Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2013)

Выпуск 12

В двух томах

Том 2. Доклады специальных секций

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2013)

Issue 12

Volume 2 of 2. Papers from special sessions

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду фундаментальных
исследований за финансовую поддержку,
грант № 13-06-06047

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19): В 2 т.

Т. 2: Доклады специальных секций — М.: Изд-во РГГУ, 2013.

Сборник включает 84 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2013», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

12-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 19-й Международной конференции «Диалог». В результате работы 54 рецензентов для сборника было отобрано 84 доклада, охватывающих различные направления исследований в области компьютерного моделирования и анализа естественного языка. В настоящем сборнике представлены:

- Лингвистическая семантика и семантический анализ;
- Формальные модели языка и их применение;
- Теоретическая и компьютерная лексикография;
- Методы оценки (evaluation) систем анализа текстов и машинного перевода;
- Корпусная лингвистика. Создание, применение, оценка корпусов;
- Новые лингвистические ресурсы;
- Интернет как лингвистический ресурс. Лингвистические технологии в Интернете;
- Онтологии. Извлечение знаний из текстов;
- Компьютерный анализ документов: реферирование, классификация, поиск;
- Автоматический анализ тональности текстов;
- Машинный перевод;
- Модели общения. Коммуникация, диалог и речевой акт;
- Анализ и синтез речи.

«Диалог» является ведущей российской конференцией по компьютерной лингвистике и, видимо, единственным в мире форумом, посвященным прежде всего проблемам компьютерного анализа русского языка. Принципиальной особенностью конференции, ее основополагающей традицией является особое внимание к технологиям автоматического анализа текста, основанным на лингвистических моделях. Именно этим объясняется и состав участников, и программа конференции, в которой соседствуют теоретические и прикладные исследования. В «Диалоге» представлены также и работы, сделанные в рамках статистических подходов, что позволяет, в частности, сравнивать полученные результаты.

«Диалог» является не только местом обмена опытом и представления новых достижений. Он является также и форумом для разработки и апробирования методик верификации и оценки как результатов лингвистических исследований, так и эффективности работы различных видов систем анализа текстов на русском языке. Целью этой работы являются единые для авторов и рецензентов принципы доказательства и оценки объективности, эффективности и научной новизны предлагаемых решений и методики проведения сравнительного тестирования, на которых могли бы основываться такие оценки.

Схожие проблемы решает в области информационного поиска семинар РОМИП: не случайно, что вот уже второй год «Диалог» и РОМИП проводят совместные дорожки тестирования, результаты участников которых докладываются на «Диалоге» и публикуются в этом сборнике.

В этом году проводилось два соревнования: по анализу тональности (продолжение тестирования 2012 года) и по оценке систем Машинного Перевода (для англо-русской языковой пары).

Особая роль русского языка обусловливает наличие в программе работ по адаптации к нему известных алгоритмов и методов, разработанных для других языков. Доказанные положительные или отрицательные результаты такого применения рассматриваются рецензентами как новые.

За год, прошедший после последней конференции, «Диалог» понес невосполнимую потерю: ушел из жизни выдающийся лингвист и один из отцов-основателей «Диалога» Александр Евгеньевич Кибрик. Трудно переоценить его роль в создании особой концепции и самой атмосферы конференции, которая сохраняется вот уже почти 40 лет, начиная с первых семинаров середины 70-х годов, из которых и вырос «Диалог». Основными чертами этой концепции всегда оставались широта взгляда, междисциплинарность, сочетание конструктивности и теоретической значимости обсуждаемых проблем. В этом году А. Е. Кибрику посвящено специальное заседание, материалы которого также вошли в сборник.

Несмотря на традиционную широту тематики докладов одного года, они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет.

Программный комитет «Диалога»
Редколлегия ежегодника «Компьютерная лингвистика
и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского Фонда Фундаментальных Исследований при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Компания Yandex
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Буате Кристиан	Гренобльский университет
Богуславский Игорь Михайлович	Политехнический университет Мадрида
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Корбетт Гревил	University of Surrey, UK
Кронгауз Максим Анисимович	Институт Лингвистики РГГУ
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Маккарти Диана	Lexical Computing Ltd., UK
Мельчук Игорь Александрович	Монреальский университет
Нивре Йоаким	Уппсальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и САПР
Раскин Виктор	Purdue University, USA
Сегалович Илья Валентинович	Компания Yandex
Селегей Владимир Павлович	Компания ABBYY
Хови Эдуард	University of Southern California, USA
Шаров Сергей Александрович	University of Leeds, UK

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В.В. Виноградова РАН
Браславский Павел Исаакович	Kontur Labs; Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitet i Tromsø
Сердюков Павел Викторович	Компания Yandex
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Александрович	University of Leeds, UK

Секретариат

Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Мытникова Татьяна Александровна, *координатор*	Компания ABBYY

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданов Алексей Владимирович
Богданова Наталья Викторовна
Богуславский Игорь Михайлович
Бонч-Осмоловская Анастасия Александровна
Браславский Павел Исаакович
Гельбух Александр Феликсович
Горностай Татьяна Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Дружкин Константин Юрьевич
Захаров Леонид Михайлович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Кэрролл Джон
Лахути Делир Гасемович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Падучева Елена Викторовна
Пазельская Анна Германовна
Подлесская Вера Исааковна
Савельев Василий Евгеньевич
Селегей Владимир Павлович
Семенова-Флюр Вера Эммануиловна
Сердюков Павел Викторович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Старостин Анатолий Сергеевич
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Урысон Елена Владимировна
Федорова Ольга Викторовна
Филиппова Екатерина Александровна
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Юдина Мария Владимировна
Янко Татьяна Евгеньевна

Contents*

Раздел II.
Анализ речи

* The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Раздел III.
Анализ тональности

Раздел IV.
Машинный перевод

Раздел II.

Анализ речи

COMBINING HMM AND UNIT SELECTION TECHNOLOGIES TO INCREASE NATURALNESS OF SYNTHESIZED SPEECH

Chistikov P. G. (chistikov@speechpro.com),
Korolkov E. A. (korolkov@speechpro.com),
Talanov A. O. (andre@speechpro.com)

Speech Technology Center ltd, St. Petersburg, Russia

We propose a text-to-speech system based on the two most popular approaches: statistical speech synthesis (based on hidden Markov models) and concatenative speech synthesis (based on Unit Selection). TTS systems based on Unit Selection generate speech that is quite natural but highly variable in quality. On the other hand, statistical parametric systems produce speech with much more consistent quality but reduced naturalness due to their vocoding nature. Combining both approaches improves the overall naturalness of synthesized speech. To reduce variability of Unit Selection results, we calculate a statistical generalization of the speaker's intonation. We created a methodology of voice model building in order to solve the task of speech parameterization. The model is a set of HMM models whose state parameters are clustered to provide good quality of synthesized speech even under conditions of insufficient training data. MFCC coefficients, pitch, energy and duration values are used as fundamental features. Objective and subjective experiments show that our method increases the naturalness of synthesized speech.

Key words: speech processing, speech synthesis, text-to-speech system, hidden Markov model, unit selection, voice model

1. Introduction

Speech synthesis (text-to-speech, TTS) is a process of transforming the character sequence of any text to a sequence of speech samples [1–3]. There are several approaches to doing this. The basic approaches are the following: rule-based speech synthesis (formant synthesis), articulatory speech synthesis, concatenative speech synthesis, and speech synthesis based on statistical models [4–8].

Currently the most popular approaches are the following: the Unit Selection algorithm (speech element selection) and statistical models (HMM TTS). The first one makes it possible to synthesize speech with maximum naturalness, given an accurately segmented voice database of a large size (10 hours and more). On the other hand, the second approach, which produces synthesized speech that is less natural, has the advantages presented below.

1. The HMM-based method provides an easy way to modify voice characteristics by using speaker adaptation/interpolation techniques. The Unit Selection algorithm generates speech with a constant style that is the same as the style of the speech in the database.
2. Speech generated by the HMM method is less natural for listeners. However, it is smoother, without detectable phone boundaries (pitch or energy leaps) which are usual for concatenative synthesis. In addition, the quality of Unit Selection TTS can be strongly reduced when some of the necessary speech elements are absent in the database. When voice models are used, absent speech elements are synthesized based on mean values which are closest to the required ones. It is possible due to tree-based context clustering, and the method provides good intelligibility when the amount of contexts is insufficient.
3. Applying the HMM-based speech synthesis method makes it possible to create a new TTS voice in much less time and to reduce the memory size required for storing the voice data.

We propose a hybrid TTS system that combines both approaches: looking for a matching sequence of speech elements in the speaker's speech database by means of the classic Unit selection algorithm, and employing a statistical intonation model which was trained on the same database. Experiments show that the naturalness of synthesized speech is increased compared to systems based only on Unit Selection or hidden Markov models.

2. System description

Structurally, the system is divided in two parts (Figure 1): the training part (the preparation stage) and the synthesis part. A speech database is created based on the speech corpus containing a set of sound files (each file contains a single recorded sentence) and a set of corresponding label files (these contain information about the speech elements in each sound file) [9–12]. Then the speech database is indexed to provide fast search for target elements by the following features: phone name, names of phones before and after the current phone, mel-frequency cepstral coefficients (MFCC) at phone boundaries, energy, pitch, and phone duration.

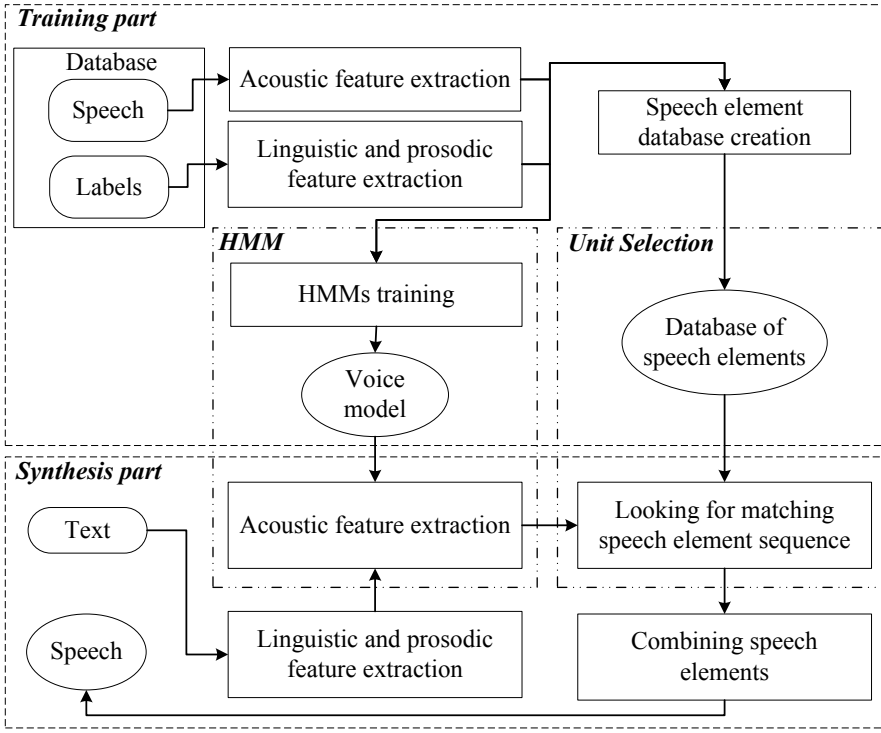


Fig. 1. Diagram illustrating the basic steps conducted by the speech synthesis engine

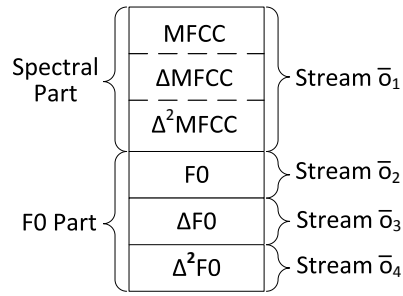


Fig. 2. Observation vector

The procedure of voice parameter modeling begins with the extraction of the feature set for all sound files [13, 14]. Each member of the set represents a short part of the signal (frame) with the length of 25 ms. The features contain the following parameters:

- Sequence $\{C_1, \ldots, C_K\}$ of MFCC vectors [15], where each vector consists of 25 coefficients and characterizes the spectrum envelope of the signal for the frame; K is the total number of frames.
- Sequence $\{F0^1, \ldots, F0^K\}$ of pitch values.

After that, linguistic and prosodic features for each allophone of all the sentences of the training database are calculated. The description of the linguistic and prosodic features is presented in Table 1.

In the next step, the HMM prototypes for each allophone are created. Each HMM corresponds to a no-skip N-state left-to-right model with N = 5. Each output observation vector $\overline{o}^i$ for the i-th frame consists of 4 streams, $\overline{o}^i = [\overline{o}_1^{iT}, \overline{o}_2^{iT}, \overline{o}_3^{iT}, \overline{o}_4^{iT}]^T$ as illustrated in Figure 2, where stream 1 is a vector composed by MFCCs, their delta and delta-delta components; stream 2 is a vector composed by F0s; stream 3 is a vector composed by F0 delta components; and stream 4 is a vector composed by F0 delta-delta components.

For each k-th HMM the durations of the N states are considered as a vector $\overline{d}^k = [\overline{d}_1^k, ..., \overline{d}_N^k]^T$, where $\overline{d}_n^k$ represents the duration of the n-th state. Furthermore, each duration vector is modelled by an N-dimensional single-mixture Gaussian distribution. The output probabilities of the state duration vectors are thus re-estimated by Baum-Welch iterations in the same way as the output probabilities of the speech parameters [16].

Table 1. Contextual features

Allophone features	
Phone before previous	Phone after next
Previous phone	Phone position from the beginning of the syllable
Current phone	Phone position from the end of the syllable
Next phone	
Syllable features	
Previous syllable	Syllable position from the end of the word
Current syllable	Syllable position from the beginning of the sentence
Next syllable	Syllable position from the end of the sentence
Number of phones in the previous syllable	Number of stressed syllables before current syllable in the sentence
Number of phones in the current syllable	Number of stressed syllables after current syllable in the sentence
Number of phones in the next syllable	Vowel name in the current syllable
Syllable position from the beginning of the word	
Word features	
Part of speech of the previous word	Number of syllables in the current word
Part of speech of the current word	Number of syllables in the next word
Part of speech of the next word	Word position from the beginning of the sentence
Number of syllables in the previous word	Word position from the end of the sentence
Sentence features	
Number of syllables in the current sentence	End punctuation type (comma, full stop, etc.)
Number of words in the current sentence	

During the voice model building, a tree-based clustering technique is applied to the HMM-states of MFCC and their delta and delta-delta components, F0 values and their delta and delta-delta components, as well as to the state duration models. In the end of the process, 4N + 1 different acoustic decision trees are generated: N trees for MFCC and their delta and delta-delta components, 3N trees for F0 features, and one tree for state duration (Figure 3). Performing this stage makes it possible to generate speech parameters for elements absent in the database, which provides intelligible output even under conditions of insufficient training data.

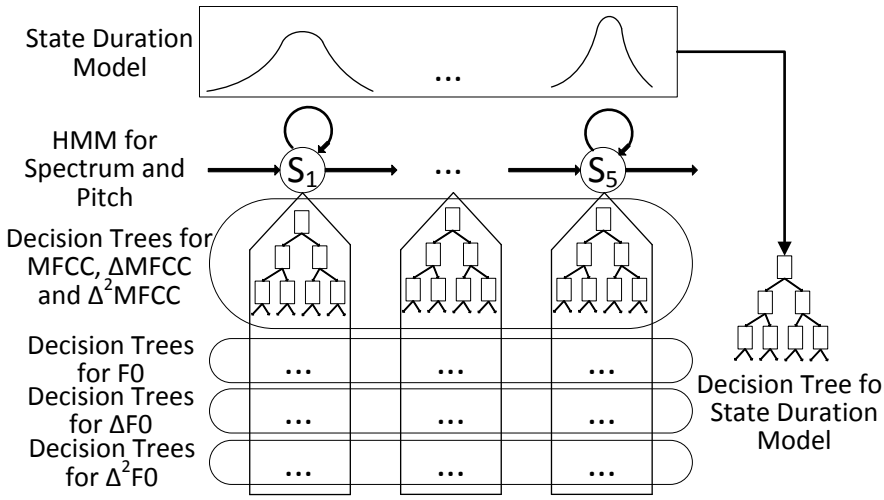


Fig. 3. Voice model

Text-to-speech system input is a raw text without any manual preprocessing. Based on the input text, the target allophone sequence is formed, and linguistic and prosodic features are calculated for each allophone. The type and structure of features are the same as those used at the stage of the speech database building. Using this information and the voice model, acoustic features are calculated for each allophone: MFCC, pitch, energy and duration. Then the most appropriate speech elements are selected from the database, based on the calculated acoustic features. Special metrics (target cost and concatenation cost) are used to estimate the suitability of each selected allophone [17].

Target cost estimation is given in equation (1):

$$C^t\left(u_i, t_i\right) = \sum_{k=1}^{p} w_k^t C_k^t\left(u_i, t_i\right) \quad (1),$$

where u_i is an element from the database; t_i is the target element; C_k^t is a distance between k-th features of elements; w_k^t is the weight of the k-th feature.

In other words, target cost is the weighted sum of differences in features between the target element and an element from the database. Any suitable linguistic and prosodic characteristics can be used as features. Usually the following information is used: pitch, duration, context, position in the syllable, position in the word, number of stressed syllables in the utterance, etc.

Selected elements should be not only close to the targets, they should also concatenate well with each other. Concatenation cost is defined as the weighted sum of differences in features between two successive selected elements:

$$C^c\left(u_{i-1},u_i\right)=\sum_{k=2}^{q} w_k^c C_k^c\left(u_{i-1},u_i\right) \quad (2),$$

where u_{i-1} is the previous element; u_i is the current element; C_k^c is the distance between k-th features of elements; w_k^c is the weight of the k-th feature.

The final cost for the whole sequence of n elements is the sum of the target cost and the concatenation cost:

$$C\left(u,t\right)=\sum_{i=1}^{n} C^t\left(u_i,t_i\right)+\sum_{i=2}^{n} C^c\left(u_{i-1},u_i\right) \quad (3).$$

The purpose of the Unit Selection algorithm is to select a sequence of elements that minimizes the final cost equation (3).

In the final step, the selected sequence of elements is concatenated to form the speech signal which is the result of TTS system work.

3. Experiments

Figures 4–6 present the results of the system's work. They are oscillograms, spectrograms, and pitch envelopes for the utterance "это очень важно!" ("eto očen' važno", Russian for "it is very important!"). A natural phrase is at the top of each figure, and its synthesized equivalent is at the bottom. It should be mentioned that this phrase had been excluded from the training data set.

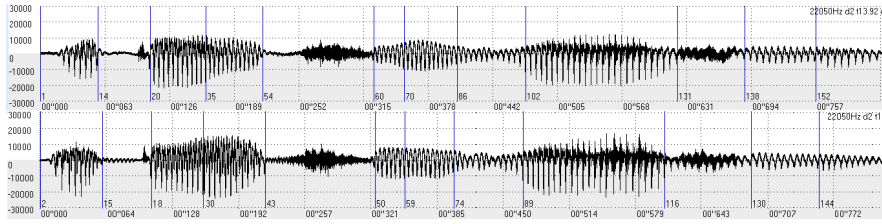


Fig. 4. Oscillograms for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

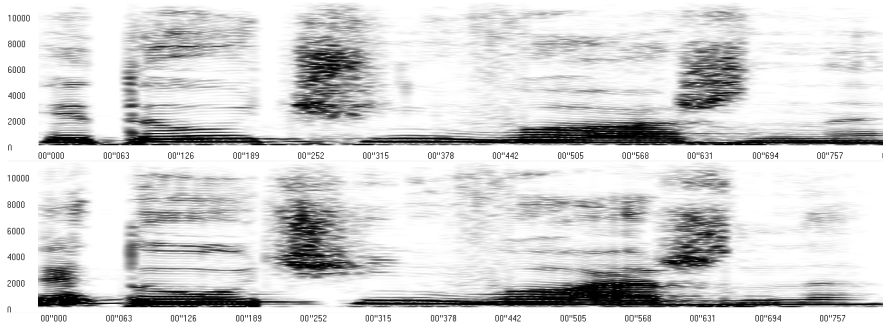


Fig. 5. Spectrograms for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

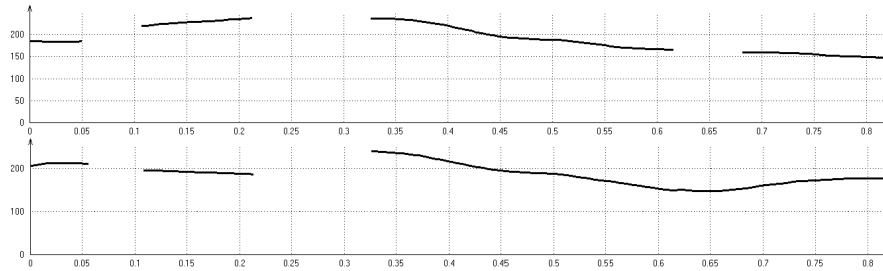


Fig. 6. Pitch envelopes for the natural sentence *"это очень важно!"* ("it is very important") (top) and its synthesized version (bottom)

From the figures above you can notice that the synthesized utterance has almost the same tempo and spectrum characteristics as the natural equivalent uttered by a real speaker. It is due to the modeling of parameters based on hidden Markov models.

We conducted a MOS (mean opinion score) evaluation to estimate the naturalness of the synthesized speech. Table 2 presents the results of the comparison for two systems: the proposed hybrid system and the system based on Unit Selection only. The comparison was performed by five experts for two voices (one male and one female); the results in the table have been averaged. The values ranged from 0 (unnatural, "mechanical" speech) to 5 (completely natural speech). The synthesized sentences were also compared to the same sentences pronounced by the speaker (they were not included in the training data set). The results show that the hybrid TTS approach increases the naturalness of synthesized speech.

Table 2. Comparison of the proposed system and the Unit Selection system

Type of TTS		Natural speech
Unit Selection	Hybrid	
4,0	4,3	4,8

4. Conclusions

This paper describes an approach for building a Russian TTS system based on the integration of hidden Markov models and Unit Selection. The TTS engine is based on a method where the speech parameters are obtained from HMMs whose observation vectors consist of MFCC, pitch and duration features; the speech signal is generated by a Unit Selection algorithm using the obtained speech parameters. We developed a voice model creation method for constructing a natural intonation contour. The experimental results confirm the improved quality of synthesized speech. It is also worth noting that the final speech quality can be improved by tuning Unit Selection weights and optimizing the training feature set.

References

1. *Dines J.* (2003), Model based trainable speech synthesis and its applications, Ph. D. Thesis, Queensland University of Technology, Brisbame, Australia.
2. *Dutoit Th.* (2002), Introduction au traitement de la parole, Faculte Polytechnique de Mons.
3. *Stilianou Y.* (1996), Harmonic plus noise models for speech, combined with statisticals methods, for speech and speaker modification, Ph.D. Thesis, Ecole Nationale Superieure des Telecommunications, Paris, France.
4. *Klatt D. H.* (1987), Review of text-to-speech conversion for English, Journal of the Acoustical Society of America, Vol. 82, pp. 737–793.
5. *Tokuda K.* (2011), HMM-based Speech Synthesis System (HTS), available at: http://hts.sp.nitech.ac.jp.
6. *Huang X., Acero A., Adcock J., Goldsmith J., Liu J., Whistler A.* (1996), Trainable Text-to-Speech System, Proc. of the International Conference on Spoken Language Processing, Philadelphia, PA, Vol. 4, pp. 2387–2390.
7. *Donovan R. E., Eide E. M.* (1998), The IBM Trainable Speech Synthsis System, Proc. ICSLP'98, Sydney, Australia.
8. *Donovan R. E., Ittycheriah A., Franz M., Ramabhadran B., Eide E., Viswanathan M., Bakis R., Hamza W.* (2001), Current Status of the IBM Trainable Speech Synthesis System, Proc. 4th ESCA Tutorial and ResearchWorkshop on Speech Synthesis, Atholl place Hotel, Scotland, UK.
9. *Prodan A., Chistikov P., Talanov A.* (2010), Voice building system for Russian TTS system "Vital Voice", Proceedings of the Dialogue-2010 International Conference, № 9 (16), pp. 394-399.
10. *Smirnova N., Chistikov P.* (2011), Software for Automated Statistical Analysis of Phonetic Units Frequency in Russian Texts and its Application for Speech Technology Tasks, Proceedings of the Dialogue-2011 International Conference, № 10 (17), pp. 632–643.
11. *Chistikov P., Khomitsevich O.* (2011), On-line automatic sentence boundary detection in a Russian ASR system, Vestnik MGTU. Priborostroenie, Special Issue "Biometric Technologies, pp. 115–123.

12. *Chistikov P., Khomitsevich O.* (2011), On-line automatic sentence boundary detection in a Russian ASR system, SPECOM 2011 International Conference, pp. 112–117.
13. *Chistikov P.* (2012), Speech parameter modeling at Russian Text-to-Speech system, Proceedings of the 1st All-Russian researcher congress, № 2, Editor-in-chief PhD, prof. V. O. Nikiforov, SPb: ITMO, pp. 227–228.
14. *Chistikov P., Korolkov E.* (2012), Data-driven Speech Parameter Generation For Russian Text-to-Speech System, Proceedings of the Dialogue-2012 International Conference, № 11 (18), pp. 103–111.
15. *Fukada T., Tokuda K., Kobayashi T., Imai S.* (1992), An adaptive algorithm for mel-cepstral analysis of speech, Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 137–140.
16. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* (2004), Hidden semi-Markov model based speech synthesis, Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 1393–1396.
17. *Black A. W., Hunt A. J.* (1996), Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database, In Proceedings of ICASSP 96, Atlanta, Georgia, Vol. 1, pp. 373–376.

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ ОПРЕДЕЛЕНИЯ МЕСТ И ДЛИТЕЛЬНОСТЕЙ ПАУЗ ПРИ АВТОМАТИЧЕСКОМ СИНТЕЗЕ РУССКОЙ РЕЧИ

Хомицевич О. Г. (khomitsevich@speechpro.com),
Чистиков П. Г. (chistikov@speechpro.com)

ООО «ЦРТ», Санкт-Петербург

Ключевые слова: синтез речи, расстановка пауз, паузирование, интонационное членение, просодический анализ, статистические методы

USING STATISTICAL METHODS FOR PROSODIC BOUNDARY DETECTION AND BREAK DURATION PREDICTION IN A RUSSIAN TTS SYSTEM

Khomitsevich O. G. (khomitsevich@speechpro.com),
Chistikov P. G. (chistikov@speechpro.com)

Speech Technology Center Ltd, St. Petersburg, Russia

The paper deals with statistical methods for predicting positions and durations of prosodic breaks in a Russian TTS system. We use CART and Random Forest classifiers to calculate probabilities for break placement and break durations, using grammatical feature tags, punctuation, word and syllable counts and other features to train the classifier. The classifiers are trained using a large high-quality speech database consisting of read speech. The experimental results for prosodic break prediction shown an improvement compared to the rule-based algorithm currently integrated in the VitalVoice TTS system; the Random Forest classifier shows the best results, although the large size of the model makes it more difficult to use in a commercial TTS system. To make the system more flexible and deal with the remaining break placement errors, we propose combining probabilities and rules in a working TTS system, which is the direction of our future research. We observe good results in experiments with predicting pause durations. A statistical model of break duration prediction has been implemented in the TTS system in order to make synthesized speech more natural.

Keywords: speech synthesis, TTS, text-to-speech, prosodic breaks, prosodic boundaries, pauses, statistical models

1. Introduction

Natural-sounding prosody is a key component for a successful Text-to-Speech (TTS) system, and correct prosodic segmentation of speech is necessary for achieving this goal. In natural speech, if an utterance is sufficiently long, it is normally divided into prosodic phrases, which are marked by intonational unity and are usually separated by pauses. Large chunks of speech pronounced without any breaks sound monotonous and are uncomfortable for the listener. In addition, accurate break placement enhances the intelligibility of speech, while pauses in the wrong positions can distort the meaning of a sentence or make it incomprehensible.

The way our natural speech is segmented prosodically depends on various factors. A major factor is syntactic structure; prosodic breaks often fall between syntactic constituents, so that syntactic structure can be seen as "mapped" onto prosodic phrases [1, 2]. However, the length of the sentence, semantics of certain words, and other features also play a role [3]. In a TTS system, these factors can be captured either by explicit rules defining which words in the synthesized sentence should be followed by a pause [4, 5], or by statistical models trained on large speech corpora and predicting probabilities of prosodic breaks. The latter method has become prevalent in the recent years (see, for example, [6–9]), and in this paper we will explore this approach as applied to a Russian TTS system.

2. Break detection using statistical methods

The principle behind automatic prosodic segmentation of speech is training a classifier on a large speech database which is labeled with word boundaries, Part-of-Speech (POS) and other grammatical tags, punctuation marks that were present in the original text (in case of read speech), and phrase breaks in the speech signal. Features like grammatical form, the place of the word in the sentence, the length of the sentence, the presence or absence of a punctuation mark, etc, are used by the classifier to predict the location and length of phrasal breaks in the synthesized speech.

This method has yielded good results for English and a number of other languages (as reported in [8] for Spanish, [9] for Arabic, etc), although some problems are bound to arise if the method is applied to Russian. Unlike English, Russian has relatively free word order, which means there is a lot of variation in possible POS sequences, and data sparseness can be an issue for model training. Russian also has rich morphology, which greatly increases the number of grammatical tags required for labeling text (and also increases variation in word form combinations). A large number of word forms in Russian are homonymous, so correct homonym resolution is essential for phrasal break detection, and errors in grammatical labeling of homonyms often lead to errors in break placement.

Despite these problems, statistical methods of phrasal break placement and break length prediction are a promising approach for Russian TTS systems, first of all because they aim to model the natural behavior of speakers, rather than rely on rigid

rules and constants. They are also easier to implement because they do not require much expert linguistic knowledge, though tuning the system for practical use may require additional linguistic constraints. In this paper we describe methods of phrasal break prediction using CART and RF classifiers, which are tested using the VitalVoice Russian TTS system [10].

3. Experimental setup

We conducted experiments using the CART classifier [11] for predicting both break placement and break duration, and the Random Forest classifier [12] for break placement only.

CART is a recursive partitioning method based on minimization of partition goodness criterion (1):

where

$$G(C_1, C_2) = \frac{D(C_1)T(C_1)+D(C_2)T(C_2)}{T(C_1)+T(C_2)} \quad (1),$$

$$D(C) = \frac{2*\left(\sum_{i=1}^{|C|}\sum_{j=1}^{|C|}\left[d(U_i,U_j)\right]\right)}{|C|^2-|C|} \quad (2),$$

$$T(C) = .5(|C|^2 - |C|) \quad (3),$$

|C| is a size of cluster C, d(U,V) is a distance between U и V vectors, stop criterion is the minimal number of items in the cluster (in our work this number is 3).

A Random Forest classifies data using a given set of features by means of a hierarchy (a "tree") of queries, based on the predictive value of each feature at each point. The classifier is capable of processing large amounts of training data. The leaves of each tree in the forest store the class distribution of all samples falling into the corresponding region of the feature space, which then serve as predictors for test samples. In our system, we use a forest containing 100 trees, and the probabilistic value is calculated by dividing the number of trees classifying the target class by the total number of trees. Each tree is built on the basis of 60% of randomized training data. This prevents the data from being dependent on noise in the training set.

For break placement prediction, we examined each sentence separately, because in the TTS system text is divided into sentences during the normalization process, and each sentence is processed separately. As for break duration, both intrasentential and intersentential breaks were taken into account.

We used the following features for the classification:

- Punctuation features. All common punctuation marks are included in the feature set. These features are calculated for the current word as well as two preceding and two following words.

- Word and syllable count features: number of words and syllables in the sentence, number of words and syllables from the previous break to the current word, from the current word to the end of the sentence, etc.
- Morphological word features. Morphological information is calculated using the VitalVoice speech synthesis engine which includes a morphogrammatical dictionary. Since using all grammatical features of Russian words would result in an enormous number of tags, which would be too large for the classifier to cope with, we decided to limit the grammatical features to part of speech and case. We also use the information on whether or not the word is a proper noun (name, geographical location, ect). Another word feature is capitalization of the first letter of the word. These features are calculated for the current word as well as two preceding and two following words. In addition, we use specific features intended for capturing grammatical agreement between words: whether or not the grammatical form of the current word matches that of the following word and the second word on the right.

Both in model training and testing, homonym resolution is necessary to minimize the number of errors due to incorrect feature calculation. We use homonym resolution provided by the VitalVoice TTS system, which labels 96% of homonyms correctly [13].

The speech database used in the experiments was originally recorded as the Unit Selection speech database for the VitalVoice TTS system and consists of read speech by nine speakers (four male and five female). The texts read by the speakers are contemporary Russian works of fiction as well as newspaper articles on the topics of politics and technology. The database comprises over 50 hours of speech, which contain over 38,000 phrasal breaks. It was divided into a training set and a test set.

4. Results and discussion

4.1. Break placement

Evaluating automatic phrase break placement is not straightforward, since the accuracy of the classification can be estimated in different ways. One way is to calculate the accuracy of the prediction (break or no break) for each pair of words (a value sometimes called “juncture”). However, significantly more junctures in speech are non-breaks than breaks, so this measure will usually yield high results even for a poorly performing system. If we only consider breaks, then two types of errors have to be taken into account: breaks added by the algorithm that were not there in the data (False Alarms, FA), and breaks incorrectly skipped by the algorithm (False Rejections, FR). Some authors [6] devise their own evaluation system which incorporates both measures; however, we prefer to use the standard precision/recall/F-score evaluation.

In Table 1 we present the results for automatic break placement (CART and Random Forest) compared to the results of the “baseline” rule-based algorithm that is currently implemented in the standard version of the VitalVoice TTS system [14]. The test set contained 47,819 junctures (word pairs inside sentences) and 6,186 phrase breaks.

Table 1. Results of automatic break detection

	Baseline TTS	CART	Random Forest
Correct junctures	43,254 (90.45%)	44,358 (92.76%)	44,723 (93.53%)
Correct breaks	5,042 (81.51%)	5,176 (83.67%)	4,624 (74.75%)
FA	3,421 (55.30%)	2,451 (39.62%)	1,534 (24.80%)
FR	1,144 (18.49%)	1,010 (16.33%)	1,562 (25.25%)
Recall	0.82	0.84	0.75
Precision	0.60	0.68	0.75
F-score	0.69	0.75	0.75

The results of both classifiers show an improvement on the baseline system: they yield a higher F-score, and the rates of FA to FR errors are more balanced. Both the CART and the RF classifiers show a maximum F-score of 75%; however, RF can be tuned so that the Precision and Recall counts are equal. CART shows a higher percentage of correct breaks due to a lower level of False Rejection errors; however, the RF classifier gives the highest percentage of correct junctures. Overall, RF can be considered the best-performing model.

The results of the classifier also compare well with those reported in the literature. For instance, for English [6] reports up to 91.1% correct junctures and the F-score of up to 71.9; [7] improves their result and attains the F-score of 74.4, which is basically equal to our result.

However, some comments on the model's performance are in order. First of all, an automatic performance test evaluates the breaks locally, without estimating the overall naturalness of the whole utterance (a discussion of this issue is given in [15]). So the fact that an utterance can usually be segmented in several correct ways remains unaccounted for. This problem is especially obvious if we consider different speakers' performance when reading the same text. Our speech database contained the same text read by several speakers, so we had a chance to find out whether they placed prosodic breaks at the same word junctures. We took a text read by three speakers (about 500 sentences) and, taking the breaks placed by one of the speakers as the model to be "tested", checked how it would fare if the other two speakers would be taken as target performances. Then we repeated the experiment with a second speaker. The results (F-scores) are given in Table 2.

Table 2. Comparison of speakers' break placement (F-score)

	Speaker 1	Speaker 2	Speaker 3
Speaker 1 as model	*1.00*	0.69	0.71
Speaker 2 as model	0.71	*1.00*	0.76

As was expected, not all breaks made by two different speakers when pronouncing the same text actually overlap; if we compare three or more speakers, the discrepancy would probably be even greater. On the other hand, treating only those breaks that coincide for several speakers as necessary and ignoring all others is clearly wrong,

because that would yield too few breaks. Interestingly, a statistical model predicts a speaker's breaks better than another human speaker; this can be explained by the fact that the model is able to generalize over multiple speakers' behavior.

Another aspect of this problem is that an automatic test that compares phrase breaks placed by the algorithm to those present in actual speech does not reflect the relative "gravity" of possible mistakes. Intuitively, some prosodic breaks in a sentence seem necessary, while others can be omitted; on the other hand, some word combination can in principle be separated by a pause, while others should be pronounced without a break. These distinctions are hard to formalize, so an automatic error detection system treats all errors as "equal". Of course, an automatic classifier should learn to avoid serious errors if the training database is sufficiently large, but in practice data sparceness is often a problem, especially for the CART classifier. In the course of subjective tests, we have identified several types of "egregious" errors that significantly worsen the impression of a model's performance, even if the overall error count is low:

- False alarm errors (inserted breaks): pauses after prepositions, conjunctions and other function words; pauses between agreeing words.
- False rejection errors (deleted breaks): lack of pauses on commas and other punctuation marks.

These errors are a particular problem for the CART classifier, though they are rare for the RF classifier. An advantage of a rule-based system is that it can easily exclude such errors by explicitly prohibiting pauses in certain word combinations and forcing them in others.

Finally, with a probability-based prosodic break model it is difficult to control rhythmic qualities of speech. The local character of decisions that the break placement algorithm takes can result in a sentence having too many pauses, while another sentence of approximately the same length and structure may have no breaks altogether. In a text that is sufficiently long, the frequency of breaks averages out and is judged by an automatic test as correct; however, specific sentences can be uncomfortable for the listener.

To sum up, even though a statistical break placement system imitates the performance of a human speaker fairly successfully, it can also make errors that should be avoided in a working TTS system. A possible solution is to "tune" the probability-based system by introducing a number of rules, which is the direction of our ongoing research.

One way to simplify the task for the break placement algorithm is to put obligatory pauses in places of punctuation marks and use the probability-based algorithm only for the text chunks without punctuation. However, in Russian punctuation is sometimes misleading in the sense that it is purely conventional and does not mark a prosodic break. So the rules need to be more elaborate than just placing a break at each punctuation mark.

In addition, breaks in certain word combinations can be prohibited. However, if we just delete breaks, the sentence may end up with too few of them. This issue is connected with the more general problem of rhythm: controlling the length of prosodic phrases and keeping the frequency of breaks constant seems to be necessary.

4.2. Break durations

The CART classifier predicts not only break positions but also the duration of each break it generates. Two versions of the model were trained. The first one predicts both break placement and break duration. The second one predicts break durations separately; that is, given a predetermined position for a prosodic break, the model predicts the break length for this position. This model can be used in combination with a rule-based break placement model or any other classifier.

In our experiments we first trained the classifier to predict the lengths of all prosodic breaks in the dataset: both those inside sentences and between sentences. After that, we decided to separate the two tasks: predicting sentence-internal vs. sentence-external breaks. It should be noted that in spontaneous speech, the notion of a sentence is controversial, and such an approach would probably fail; in that case it would be more productive to distinguish between types of breaks such as long and short breaks. However, since we were dealing with read speech, we felt that speakers were aware of sentences in the text and marked them prosodically, and we wished to imitate this effect in synthesized speech.

Break duration accuracy is much more difficult to evaluate than break placement accuracy because break lengths are not discrete and there can be no yes/no judgements. For our first model (predicting both break placement and duration), the problem is that if we evaluate the lengths of the breaks correctly predicted by the classifier, there still remain the inserted breaks (FA-type errors) whose lengths will be unaccounted for. For this reason, we decided to test the second model and to evaluate the correctness of the break length prediction that the classifier makes for each break position found in the test dataset. We considered a break as correct if its length did not deviate from the predicted length by more than a certain percentage, which we set as either 30% or 50%. The results are given in Table 3.

Table 3. Results for break length prediction

	Correct sentence — external breaks, %	**Correct sentence — internal breaks, %**
General model, 30% window	63.07%	42.74%
Specialized models, 30% window	64.12%	60.36%
General model, 50% window	81.88%	63.68%
Specialized models, 50% window	80.99%	80.16%

This table presents results for the general model (modeling all breaks in the dataset) and the specialized models (two separate models for sentence-external and sentence-internal breaks). We can see that the specialized models give a better approximation both for sentence-internal and sentence-external breaks (except for sentence-external breaks with a 50% evaluation window, where the results for the general model are slightly better). The difference is especially large for sentence-internal breaks, which are apparently not predicted accurately enough by the general model.

The baseline algorithm for break durations in the VitalVoice TTS system uses constants, so all sentence-external breaks have the same length, and there are only three types of sentence-internal breaks differing by their length. Implementing probability-based pause length prediction is promising because it makes synthesized speech sound less monotonous and more varied, which contributes to overall naturalness of speech. Subjective listening experiments with a new TTS system where constants were replaced by predicted values showed positive results.

5. Conclusions and future research

In this paper we have presented a probability-based approach to prosodic analysis of speech. The aim of our research was to evaluate different models of break placement and break length prediction for use in a Russian TTS system. The following conclusions can be drawn at the present stage of the research:

1. A break placement algorithm based on a probabilistic model gives better test results than the baseline rule-based algorithm. However, subjective evaluation shows that the presence of errors, even if they are rare, produces a bad impression on listeners, so some additional tweaking is needed in order to include the algorithm in a working TTS system. The CART model displays more errors than the RF model, and these errors are typically more "serious"; however, the RF model slows down the system due to its large size. Adapting statistical break placement methods for practical use will be the subject of future work.
2. CART-based prediction of pause lengths works well, especially if sentence-internal and sentence-external breaks are modeled separately. This model has been included in a new version of the VitalVoice TTS system to replace the old constant-based system, and has received good reviews from expert listeners.

References

1. *Bachenko, J., Fitzpatrick, E.* (1990), A computational grammar of discourse-neutral prosodic phrasing in English, Computational linguistics, Vol. 16 (3), pp. 155–170.
2. *Tepperman, J., Nava, E.* Where should pitch accents and phrase breaks go? A syntax tree transducer solution. Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 2011, pp. 1353–1356.
3. *Zellner, B.* (1994), Pauses and the temporal structure of speech, in Fundamentals of speech synthesis and speech recognition, Chichester, John Wiley, pp. 41–62.
4. *Abney, S.* (1991). Parsing by chunks. Principle-based parsing, Vol. 44, pp. 257–278.
5. *Atterer M.* Assigning Prosodic Structure for Speech Synthesis: A Rule-based Approach. Proceedings of Prosody, Aix-en-Provence, 2002, pp. 147–150.

6. *Black, A. W., Taylor, P.* (1998). Assigning phrase breaks from part-of-speech sequences. Computer Speech & Language, Vol. 12(2), pp. 99–117.
7. *Busser, B., Daelemans, W., Bosch, A. V. D.* Predicting phrase breaks with memory-based learning. 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis, 2001, pp. 29–34.
8. *Torres, H. M., Gurlekian, J. A.* Automatic determination of phrase breaks for Argentine Spanish. Proceedings of Speech Prosody, 2004, pp. 553–556.
9. *Sawalha, M., Brierley, C., Atwell, E.* Predicting Phrase Breaks in Classical and Modern Standard Arabic Text. Proceedings of LREC: Language Resources and Evaluation Conference, 2012.
10. *VitalVoice™* Russian TTS system, demo available at: http://cards.voicefabric.ru/.
11. *Loh, W.-Y.* Classification and Regression Tree Methods, in Encyclopedia of Statistics in Quality and Reliability, Wiley, 2008, pp. 315–323.
12. *Breiman, L., Cutler, A.* Random Forests, available at: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.
13. *Khomitsevich O. G., Rybin S. V., Anichkin I. M.* Linguistic analysis for text normalization and homonymy resolution in a Russian TTS system [Ispol'zovanie lingvisticheskogo analiza dlja normalizatsii teksta i snjatija omonimii v sisteme sinteza russkoj rechi.] Izvestija vuzov. Priborostroenie. Tematicheskij vypusk "Rechevye informatsionnye sistemy". [Instrument making. Thematic issue "Speech information systems"] №2, 2013 (in press).
14. *Khomitsevich O. G., Solomennik M. V.* Automatic pause placement in a Russian TTS system [Avtomaticheskaja rasstanovka pauz v sisteme sinteza russkoj rechi po tekstu]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010"]. Bekasovo, 2010. pp. 531–537.
15. *Jian, L., Bolei, H., Hairon, X., Linfang, W., Braga, D., Sheng, Z.* Expand CRF to Model Long Distance Dependencies in Prosodic Break Prediction. Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Portland, Oregon, 2012.

РАСПОЗНАВАНИЕ ДВУЯЗЫЧНОЙ РЕЧИ БЕЗ ПРЕДВАРИТЕЛЬНОЙ ИДЕНТИФИКАЦИИ ЯЗЫКА

Людовик Т. В. (tetyana.lyudovyk@gmail.com),
Пилипенко В. В. (valeriy.pylypenko@gmail.com)

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и МОН Украины, Киев, Украина

Предлагается подход к дикторонезависимому распознаванию слитной украинско-русской речи, не требующий предварительной сегментации речевого сигнала на разноязычные участки и идентификации языков. Также не требуется создание специальных речевого и текстового корпусов для обучения акустической и лингвистической моделей. Подход учитывает особенности фонетических систем русского и украинского языков. Используется разработанная ранее акустическая модель украинской речи. Двуязычная лингвистическая модель обучается на множестве украинских и русских текстов. Лексикон распознавания объединяет словоформы обоих языков, при этом фонемные транскрипции русских словоформ представлены украинскими фонемами.

Предлагаемый подход применим для распознавания двуязычной речи с межфразовым и внутрифразовым языковым переключением.

Ключевые слова: двуязычная речь, языковое переключение, украинский язык, русский язык, автоматическое распознавание речи, автоматическая идентификация языка

BILINGUAL SPEECH RECOGNITION WITHOUT PRELIMINARY LANGUAGE IDENTIFICATION

Lyudovyk T. V. (tetyana.lyudovyk@gmail.com),
Pylypenko V. V. (valeriy.pylypenko@gmail.com)

International Research/Training Center for Information
Technologies and Systems, Kyiv, Ukraine

We presents an approach to speaker-independent recognition of large-vocabulary continuous speech characterized by code-switching between

Ukrainian and Russian. The approach does not require language boundary detection or language identification. Special speech and text corpora are not needed to train acoustic and linguistic models. The approach takes into account peculiarities of phonetic systems of Russian and Ukrainian languages.

A cross-lingual speech recognition system is developed. A previously developed acoustic model of Ukrainian speech serves for both languages. A set of HMM-models representing 54 Ukrainian phonemes and several non-speech units such as breath, fillers and silence are used. Bilingual linguistic model is trained on a set of Ukrainian and Russian texts. Pronunciation lexicon combines word forms in both languages. Phonemic transcription of Russian word forms are generated using Ukrainian phonemes. Recognition post-processing can be applied to smooth recognized word sequences by using a dictionary containing Ukrainian and Russian words which sound equally but are written differently.

The proposed approach can be applied to the recognition of bilingual speech with between-phrase and within-phrase code-switching.

Developed cross-lingual speech recognition system was tested on Ukrainian, Russian, and Ukrainian-Russian speech of one bilingual speaker. Preliminary results show that the proposed approach could achieve a good performance. Accuracy of mixed speech recognition is lower only by 3–7 % as compared with monolingual speech recognition accuracy.

Key words: mixed speech, bilingual speech, bilingual code-switching, Ukrainian language, Russian speech, automatic speech recognition, automatic language identification

1. Введение

Большинство разработанных к настоящему времени систем автоматического распознавания речи ориентировано на правильную, нормативную одноязычную речь. Однако, часто приходится иметь дело с двуязычной речью — языковым переключением («code switching»), когда чередуются отрезки речи на разных языках, или смешанной речью («code mixing»), когда в речи появляются слова и обороты, образованные из элементов разных языков.

Проблема распознавания двуязычной речи актуальна в мировом масштабе, однако активно в настоящее времени она решается в Гонконге [3, 7], Сингапуре [16, 17], Тайване [11] и Индии [1, 2].

Проблему распознавания речи с языковым переключением можно рассматривать как проблему распознавания многоязычной речи [6, 8]. В такой постановке она решается, например, в Швейцарии на материале пяти европейских языков [8].

В перечисленных работах решаются задачи разной сложности: распознавание изолированных слов узкой предметной области [8], распознавание подготовленной слитной речи [1, 2], распознавание спонтанной слитной речи [16, 17].

В речи с переключением языки могут быть «равноправными» или «примешиваемыми». Наиболее часто примешивается английский язык к китайскому

[16, 17] и хинди [1, 2]. В [11] исследуется переключение между равноправными мандаринским и тайваньским диалектами китайского языка.

В Украине наблюдается как переключение между русским и украинским языками, так и смешение этих языков. Пример межфразового переключения: «*Сейчас я задам два вопроса*. ***Перше питання до вас***». Пример внутрифразового переключения: «*В наше время* ***замовник*** *становится* ***загарбником***». Смешение языков может происходить на фразовом и словарном уровнях. Это явление принято называть суржиком.

Данная работа посвящена распознаванию слитной спонтанной украинско-русской речи с языковым переключением. Чередующиеся отрезки речи могут быть произвольной длительности. Дикторами могут быть как билингвы, владеющие украинским и русским языками, так и носители только одного их этих языков. При этом системе распознавания неизвестно, на каком из языков произносятся те или иные отрезки речи. Такая ситуация типична при распознавании диалогов и интервью.

2. Существующие подходы

В типичной одноязычной системе распознавания речи используются: акустическая модель (АМ), лингвистическая модель (ЛМ), лексикон распознавания и декодер [5]. АМ — это множество HMM-моделей, статистически характеризующих акустические свойства фонем. ЛМ — это статистическая модель, задающая вероятности появления пар словоформ в речи. АМ предварительно обучается на большом корпусе речевых данных, а ЛМ, соответственно, — на большом корпусе текстовых данных. Лексикон распознавания представляет собой словарь, в котором словоформы содержатся в орфографическом виде и в виде фонемных транскрипций, что позволяет устанавливать соответствие между лексическим и акустическим уровнями. В декодере обработанный входной речевой сигнал сопоставляется с информацией, хранящейся в АМ и ЛМ, и определяется наиболее вероятная последовательность словоформ, соответствующая этому речевому сигналу.

Для распознавания двуязычной речи было предложено два подхода: многопроходный и однопроходный.

Многопроходный подход заключается в нахождении границ разноязычных фрагментов, идентификации языков и использовании соответствующих одноязычных систем для распознавания этих фрагментов [14]. Этот подход имеет существенный недостаток: он в значительной степени зависит от точности нахождения границ одноязычных фрагментов и от точности идентификации языка. В случае близкородственных языков для их идентификации проблематично использовать акустическую, фонотактическую и просодическую информацию.

При однопроходном подходе работает единая двуязычная система распознавания, для которой АМ, ЛМ и лексикон строятся как для единого смешанного языка. При этом нет необходимости идентифицировать язык.

Однопроходный подход применен в [1, 2] для распознавания хинди-английской речи. Использовалась англоязычная система распознавания речи *Sphinx* и прилагающаяся к ней АМ английских фонем [4]. ЛМ была обучена на смешанноязыковых текстах, соответствующих созданному небольшому смешанному речевому корпусу (213 предложений). Объем хинди-лексикона составлял 2071 слово. Слова хинди сначала были транслитерированы, а потом фонетически затранскрибированы с использованием английских фонем. Недостающие хинди-фонемы были аппроксимированы комбинациями английских фонем.

В [16, 17] однопроходный подход использован при распознавании спонтанной речи с большим словарем и переключением между китайским и английским языками. Обучение АМ происходило на 58,4 часа речи 139 дикторов. Алфавит фонем содержит английские и китайские фонемы, а также фонемы, принадлежащие обоим языкам (обозначаемые одинаковыми символами МФА — Международного фонетического алфавита (*IPA, International Phonetic Alphabet*). Для обучения ЛМ путем машинного перевода были созданы искусственные тексты с переключением языков. Результат распознавания контрольной выборки (10 дикторов, 2,3 часа речи) — 36,9 % MER (*mixed error rate* — смешанный показатель ошибки: пословной для английского языка и иероглифической — для китайского).

В [11] однопроходный подход применен к распознаванию речи с переключением между близкородственными мандаринским и тайваньским диалектами китайского языка, использующими одну и ту же систему письменности.

В [3] разработан гибридный подход: наряду со смешанными китайско-английскими моделями используется получаемая информация о границах разноязычных участков речи. В результате достигнута точность пословного (для английского языка) и послогового (для китайского) распознавания, равная 60 %.

3. Предлагаемый подход

Предлагаемый однопроходный подход учитывает особенности фонетических систем русского и украинского языков, позволяющие использовать в качестве АМ акустическую модель, разработанную ранее для распознавания украинской речи [12]. Двуязычная ЛМ может быть обучена на множестве украинских и русских текстов [15], при этом она может моделировать межфразовое переключение языков. Лексикон распознавания объединяет словоформы обоих языков; транскрипции русских слов представлены украинскими фонемами.

Предложенный подход не требует создания специальных речевого и текстового корпусов для обучения АМ и ЛМ.

4. Особенности украинской и русской фонетических систем

Украинский и русский языки являются близкородственными. Одной из целей данного исследования является выяснение того, положительно или отрицательно сказывается близкородственность на результатах распознавания речи.

Поскольку в работах, посвященных речевым технологиям, обычно не проводится четкое разделение между фонемами и аллофонами, в дальнейшем будем употреблять единый термин «фонема». Обычно для синтеза и распознавания русской и украинской речи пользуются расширенным алфавитом фонем [9, 10, 13].

Система фонем украинского языка включает все фонемы русского языка. При этом фонотактические различия языков незначительны. В первую очередь различия касаются частоты встречаемости звонких взрывных заднеязычных фонем (украинские [ґ] и [ґ’] встречаются в речи значительно реже, чем русские [г] и [г’]), а также частоты встречаемости ударной гласной переднего ряда среднего подъема (украинская [е], русская [э]) в позиции после мягких согласных (укр. «*суттєво*», рус. «*комитет*»).

В русской речи украинских билингвов наблюдается фонотактическая интерференция, которая наиболее сильно сказывается на замене взрывного [г] щелевым, а также на ослаблении редукции безударных гласных.

Данное исследование опирается не только на эти общие сведения, но также и на наш опыт разработки систем синтеза украинской и русской речи [16].

Для синтеза украинской речи нами использован алфавит из 59 фонем: 12 гласных (6 ударных и 6 безударных), 45 согласных (22 пары «твердая-мягкая» и непарная [й]), пауза и гортанная смычка.

Алфавит, используемый нами для синтеза русской речи, значительно меньше, поскольку отсутствуют [г], [г’] («южнорусские»), [ц’], [ч], [дз], [дз’], [дж], [дж’]. В то же время, каждая из русских фонем совпадает с одной из украинских фонем или близка к одной из них. Кроме того, на произнесении русских фонем сказывается украинский акцент, что еще более приближает русские фонемы к украинским.

Мы предполагаем, что билингвы в Украине, говоря на украинском и русском языках, пользуются одним и тем же набором фонем. Это, в свою очередь, дает основание предполагать, что при распознавании русской речи можно использовать украинскую акустическую модель.

5. Речевые и текстовые данные

Проведенное исследование базируется на данных Акустического корпуса украинской эфирной речи (АКУЕМ) [14], в котором представлена украинская и русская речь, прозвучавшая в украинском телеэфире. Количественные характеристики корпуса АКУЕМ приведены в Таблице 1. На обоих языках говорят 160 дикторов корпуса.

Таблица 1. Количественные характеристики речевого корпуса украинской и русской теле- и радиоэфирной речи

	Украинская речь	Русская речь
Длительность акустических записей	116 часов	190 часов
Общее количество словоформ	962 504	1 721 606
Количество уникальных словоформ	69 500	83 500
Количество дикторов, длительность речи которых превышает 10 сек.	1 723	2 781

Для проверки принятого подхода в качестве диктора был выбран билингв, имеющий опыт публичных выступлений. Были отобраны фрагменты его спонтанной речи на украинском и русском языках, записанной в телестудии во время трансляций двух политических ток-шоу. Все речевые фрагменты были сегментированы на моноязычные участки длительностью до 15 сек. с границами, соответствующими границам синтагм. Таким образом, исследовалась речь с межфразовым языковым переключением. Объем украинского речевого материала составил 1693 слова (856 уникальных словоформ), русского — 823 слова (503 уникальные словоформы).

Текстовый материал, использованный в исследовании, состоял из всех текстов (на украинском и русском языках) корпуса АКУЕМ. Объем текстового материала — 2,7 млн. словоформ.

6. Система распознавания украинской речи

Базовой для исследования является дикторонезависимая система распознавания подготовленной и спонтанной украинской речи [12], разработанная с помощью инструментария HTK [18]. В качестве АМ используется набор контекстно-независимых скрытых Марковских моделей, обученных на материале украинской речи корпуса АКУЕМ. Помимо акустических моделей 54 фонем украинского языка используются особые модели для паузы, звучащих пауз хезитации («э-э-э», «а-а-а», «м-м»), вдохов/выдохов, чмоканья, кашля, смеха и плача.

ЛМ представляет собой биграммную модель языка, заданную вероятностями пар словоформ. ЛМ обучена на украиноязычных текстах корпуса АКУЕМ (20 Мб) и текстах из Интернета (400 Мб).

Лексикон распознавания содержит 116 тыс. словоформ. Часто употребляемые в речи словоформы, в том числе числительные, имеют от 1 до 10 фонемных транскрипций, остальные — от 1 до 3. В среднем на одну словоформу в лексиконе распознавания приходится 1,5 фонемных транскрипций, отражающих каноническое и спонтанное произнесение. Для основной части словоформ (92%) транскрипции порождены полностью автоматически, в 7% словоформ вручную были расставлены знаки ударения (фамилии, географические названия и т. п.). Для остальных, наиболее частотных словоформ (1% от общего

объема лексикона), автоматически порожденные фонемные транскрипции были дополнены транскрипциями, написанными экспертом.

Точность пословного распознавания контрольной выборки описанной системой — 87,71 %.

7. Экспериментальная система

Наиболее трудоемким и длительным процессом при разработке системы распознавания речи является создание и обучение АМ.

Исходя из гипотезы, что все фонемы русской речи можно моделировать украинскими фонемами, в качестве АМ для распознавания речи с языковым переключением была выбрана украинская АМ.

Лексикон для распознавания речи с языковым переключением содержит 56753 украинские словоформы и 58058 русских словоформ. Количество словоформ, совпадающих по орфографическому написанию, — 4952 (например, «*новенький*», «*проводили*»); совпадающих и по написанию, и по звучанию — 257 (например, «*думаю*», «*народ*»).

Фонемные транскрипции русских словоформ были порождены транскриптором, использующимся для синтеза речи. Были изменены обозначения фонем: «и» на «i», «э» на «е», «г» на «ґ». В Таблице 2 приведены примеры фонемных транскрипций русских словоформ.

Таблица 2. Примеры словоформ и их фонемных транскрипций

Примеры словоформ	Фонемные транкрипции на базе украинского алфавита фонем
думаю	д У м а й у
государственный	ґ а с у д А р с т в’ е н и й
обстоятельства	а п с т а й А т’ е л’ с т в а

ЛМ для данного исследования была обучена на всех текстах (украиноязычных и русскоязычных) корпуса АКУЕМ. Поскольку речевые сегменты контрольной выборки также содержатся в этом корпусе, OOV = 0 (out-of-vocabulary rate — процентное отношение количества словоформ контрольной выборки, отсутствующих в ЛМ, к общему количеству словоформ контрольной выборки).

На рис. 1 представлена конфигурация экспериментальной системы распознавания слитной спонтанной украинско-русской речи с языковым переключением.

Было проведено девять экспериментов для оценки работоспособности предложенного подхода к распознаванию речи с языковым переключением. Оценивалась пословная точность распознавания. Поскольку все речевые сегменты контрольной выборки распознаются независимо друг от друга, их общее множество можно рассматривать как речь с языковым переключением.

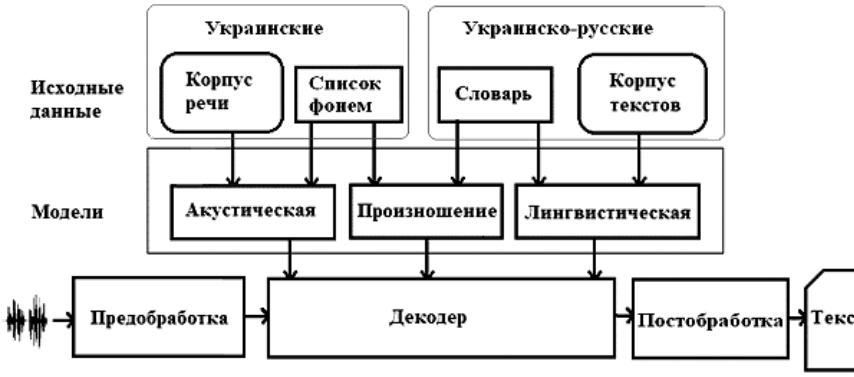


Рис. 1. Конфигурация системы распознавания украинско-русской речи

Во всех экспериментах использовалась украинская АМ. Распознавались:

1. все украинские речевые сегменты контрольной выборки;
2. все русские речевые сегменты;
3. все речевые сегменты.

Для распознавания украинской речи были разработана ЛМ, обученная на всех украинских текстах корпуса АКУЕМ. Лексикон распознавания включал только украинские словоформы из общего лексикона. Аналогично, для распознавания русской речи были разработаны ЛМ, обученная на всех русских текстах корпуса АКУЕМ и лексикон распознавания, включающий только русские словоформы из общего лексикона.

Для распознавания речи с языковым переключением использовались украинско-русская ЛМ и украинско-русский лексикон.

8. Экспериментальные результаты

Результаты экспериментов представлены в Таблице 3. Во всех экспериментах в качестве акустической модели использовалась АМ, обученная на украинской речи корпуса АКУЕМ. Обозначения лингвистических моделей:

- UKR_LM — ЛМ, обученная на всех украиноязычных текстах корпуса АКУЕМ;
- RUS_LM — ЛМ, обученная на всех русскоязычных текстах корпуса АКУЕМ;
- UKRUS_LM — ЛМ, обученная на всех текстах корпуса АКУЕМ.

ukr_lex — лексикон распознавания, включающий словоформы украинского языка;
rus_lex — лексикон распознавания, включающий словоформы русского языка;
ukrus_lex — лексикон распознавания, включающий словоформы обоих языков.

В «ukrus_lex» фонемные транскрипции русских словоформ представлены украинскими фонемами.

Таблица 3. Результаты распознавания украинской, русской и украинско-русской речи

Речевые сегменты	Количество слов	Точность распознавания, %		
		UKR_LM, ukr_lex	RUS_LM, rus_lex	UKRUS_LM, ukrus_lex
украинские	1693	85,17	10,99	83,70
русские	823	13,83	85,80	73,54
украинские + русские	2516	61,82	35,36	80,37

9. Обсуждение результатов

Результаты экспериментов свидетельствуют о том, что украинская АМ обеспечивает хорошие результаты как при распознавании моноязычной речи (украинской — 85,17%, русской — 85,80%), так и речи с языковым переключением (80,37%). Следует, однако, учесть, что в контрольной выборке представлена речь одного диктора, и этот диктор входит в число дикторов корпуса АКУЕМ. Кроме того, все слова контрольной выборки представлены в ЛМ. Однако, результаты экспериментов достаточно показательны, поскольку были проведены с целью сравнения эффективности различных конфигураций системы распознавания на одних и тех же моно- и двуязычных речевых сегментах.

Эксперименты по распознаванию русской речи системой, ЛМ и лексикон которой настроены на украинский язык (точность 13,83%) и украинской речи системой, ЛМ и лексикон которой настроены на русский язык (точность 10,99%), были проведены с целью оценки, какой должна быть точность идентификации языка в двухпроходной системе, когда на первом проходе идентифицируется язык, а на втором происходит распознавание с использованием моделей идентифицированного языка. Предварительные расчеты показывают, что для достижения 80,37% точности распознавания смешанной украинско-русской речи точность предварительной идентификации языка должна быть выше 95%.

Наиболее перспективным направлением дальнейших исследований является разработка кросс-язычной АМ, включающей модели как украинских, так и русских фонем. В связи с этим в дальнейшем предполагается проведение акустико-фонетического экспериментального исследования украинско-русской речи билингвов с целью уточнения алфавита фонем, на основе которого должна быть создана кросс-язычная АМ.

Результаты распознавания украинско-русской речи могут быть улучшены на стадии постобработки. Ответ распознавания в виде смешанного текста можно сглаживать с помощью двуязычных словарей, в которых представлены одинаково звучащие, но имеющие разную орфографию слова обоих языков. Например, «*лінія*» — «*линия*», «*ера*» — «*эра*».

10. Заключение

В статье предложен подход к распознаванию речи с языковым переключением между близкородственными украинским и русским языками.

Использование акустической модели, разработанной ранее и обученной на украинской речи, для распознавания украинско-русской речи, на первых порах позволило избежать трудоемкой и длительной процедуры разработки специальной двуязычной акустической модели. Точность распознавания двуязычной речи при этом отличается от точности распознавания моноязычной речи незначительно — на 3–7 %.

Разработанные лингвистическая модель и лексикон распознавания являются двуязычными. Обучение двуязычной лингвистической модели значительно проще, чем одноязычной модели, поскольку не требует предварительного разделения множества текстов по языковому признаку. Лексикон распознавания, обеспечивающий распознавание двуязычной речи, содержит как украинские, так и русские словоформы, причем фонемные транскрипции русских словоформ представлены украинскими фонемами.

Предложенный подход не требует предварительной идентификации языка.

Литература

1. *Bhuvanagiri, K. K., Kopparapu, S. K.* An Approach to Mixed Language Automatic Speech Recognition. Proceedings of the Oriental COCOSDA, 2010, Nepal.
2. *Bhuvanagiri, K. K., Kopparapu, S. K.* (2012) Mixed Language Speech Recognition without Explicit Identification of Language, American Journal of Signal Processing, Vol. 2 No. 5, pp. 92–97.
3. *Chan J. Y. C., Ching, P. C., Lee, T., Cao, H.* (2009) Automatic recognition of Cantonese-English code-mixing speech, Computational Linguistics and Chinese Language Processing, vol. 14, No. 3, pp. 281–304.
4. *Derbali, M., Jarrah, M., Wahid, M. T.* (2012) A Review of Speech Recognition with Sphinx Engine in Language Detection, Journal of Theoretical and Applied Information Technology, Vol. 40, No. 2, pp. 147–155.
5. *Despres, J., Fousek, P., Gauvain, J.-L., Gay, S., Josse, Y., Lamel, L., Messaoudi, A.* The LIMSI-Vecsys Research Systems for N-Best 2008. Proceedings of the N-Best: North- and South-Dutch Benchmark Evaluation of Speech recognition Technology workshop, Soesterberg, NL, 2008.
6. *Fung, P. N., Shi, B., Wu, D., Bun, L. W., Kong, W. S.* Dealing with Multilinguality in a Spoken Language Query Translator. Proceedings of the ACL/EACL'97 Workshop on Spoken Language Translation, Madrid, Spain, 7–12 July 1997, pp. 40–43.
7. *Huang, C.-L., Wu, C.-H.* Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis. IEEE Transactions on Computers, 2007, vol. 56 no. 9, pp. 1225–1233.
8. *Imseng, D., Bourlard, H., Magimai-Doss, M.* Towards mixed language speech recognition systems. Proceedings of Interspeech, Makuhari, Japan, 2010, pp. 278–281.

9. *Krivnova, O., Zinovieva, N., Zakharov, L., Strokin, G., Babkin, A.* (1997) TTS Synthesis for Russian Language, Web Journal of Formal, Computatuonal and Cognitive Linguistics, Issue 1, available at: http://fccl.ksu.ru/ar2.htm.
10. *Lamel, L., Courcinous, S., Gauvain, J.-L., Josse, Y., Le, V.-B.* Transcription of Russian conversational speech. Proceedings of the SLTU 2012 Third International Workshop on Spoken Languages Technologies for Under-resourced Languages. Cape Town, South Africa, 2012, pp. 162–167.
11. *Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., Hsu, C.-N.* Speech Recognition On Code-Switching Among The Chinese Dialects. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, May. 2006. Volume 1, pp. 1105–1108.
12. *Lyudovyk, T. V., Pilipenko, V. V., Robeiko, V. V.* Automated stenographer of Ukrainian speech (on material of acoustic corpus of Ukrainian speech) [Avtomaticheskoe raspoznavanie spontannoi ukrainskoi rechi (na materiale akusticheskogo korpusa ukrainskoi efirnoi rechi)]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2011" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011"]. Bekasovo, 2011, pp. 478–488.
13. *Lyudovyk, T., Sazhok, M.* Unit Selection Speech Synthesis Using Phonetic-Prosodic Description of Speech Databases. Proceedings of the International Conference "Speech and Computer" (SPECOM'2004). St.-Petersburg, Russia, 2004. pp. 594–599.
14. *Niesler, T., Willett, D.* Language identification and multilingual speech recognition using discriminatively trained acoustic models. Proceedings of the ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006), paper 004.
15. *Robeiko, V., Pylypenko, V., Sazhok, M., Vasylieva, N., Radoutsky, O.* Ukrainian Broadcast Speech Corpus Development. Proceedings of the 14th International Conference "Speech and Computer: SPECOM'2011". Kazan, Russia, 2011, pp. 435–440.
16. *Vu, N. T., Lyu, D. C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T., Li, H.* et al. A First Speech Recognition System for Mandarin-English Code-switch Conversational Speech. Proceedings of the ICASSP, Japan, 2012, pp. 4889–4892.
17. *Weiner, J., Vu, N. T., Telaar, D., Metze, F., Schultz, T.* et al. Integration of language identification into a recognition system for Spoken conversations containing code-switches. Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages, Cape Town; S. Africa, May 2012. MICA.
18. *Young S.* et al. The HTK Book (for HTK Version 3.4). Cambridge, UK, 2009.

EVALUATION OF NATURALNESS OF SYNTHESIZED SPEECH WITH DIFFERENT PROSODIC MODELS

Solomennik A. I. (solomennik-a@speechpro.com)
Speech Technology Ltd., Minsk, Belarus

Chistikov P. G. (chistikov@speechpro.com)
Speech Technology Center Ltd, St. Petersburg, Russia

Obtaining natural synthesized speech is the main goal of modern research in the field of speech synthesis. It strongly depends on the prosody model used in the text-to-speech (TTS) system. This paper deals with speech synthesis evaluation with respect to the prosodic model used. Our Russian VitalVoice TTS is a unit selection concatenative system. We describe two approaches to prosody prediction used in VitalVoice Russian TTS. These are a rule-based approach and a hidden Markov model (HMM) based hybrid approach. We conduct an experiment for evaluating the naturalness of synthesized speech. Four variants of synthesized speech depending on the applied approach and the speech corpus size were tested. We also included natural speech samples into the test. Subjects had to rate the samples from 0 to 5 depending on their naturalness. The experiment shows that speech synthesized using the hybrid HMM-based approach sounds more natural than other synthetic variants. We discuss the results and the ways for further investigation and improvements in the last section.

Key words: speech synthesis, unit selection, naturalness evaluation, prosodic modeling

1. Introduction

The task of speech synthesis or text-to-speech (TTS) is to convert a written text to sounds. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood, i. e. the quality of synthetic speech depends primarily on two main factors: its intelligibility and naturalness. It is possible to say that the problem of intelligibility for speech synthesis is already solved [Taylor 2009: 474]. Extensive research in the field of speech synthesis during the last few decades allowed synthetic speech to sound quite natural, and its characteristics come close to those of human speech.

At present the two main and most popular methods of natural-sounding speech synthesis are unit selection concatenative synthesis and so-called hidden Markov model (HMM) synthesis based on statistic models.

Unit selection synthesis [Black, Hunt 1996] is based on determining the best sequence of candidate units from a speech corpus. Then these candidates are concatenated

to form the resulting words and sentences. This process may be followed by modification of prosodic features of units (duration, energy and pitch) to match prescribed values.

HMM-based TTS is also called statistical parametric synthesis. A TTS system of this type models frequency spectrum, fundamental frequency (pitch) and duration of speech by HMM and then generates speech waveforms directly from HMM based on the maximum likelihood criterion [Masuko 2002; Zen et al. 2004]. Although HMM TTS provides an easy way to modify voice characteristics, speech generated without natural units usually sounds less natural than unit selection synthesis. This is the reason why we use unit selection in our TTS system.

However, naturalness of speech depends not only on segmental quality. Prosodic features including pitch, duration and energy and the way of achieving their required values are by no means less important. There are several approaches to the task [Krivnova 2000]. In the next sections we consider two ways to obtain them which are used in our VitalVoice Russian TTS system [Oparin, Talanov 2007].

2. Rule-based approach

The first approach is rule-based. It consists of two steps. During the first step we define the intonation type of the phrase (i.e. syntagma) and the word bearing the nuclear pitch accent depending on punctuation, parts of speech of words in the phrase and presence of special trigger words (question words, conjunctions, etc.). This is performed by manually constructed rules. It is worth mentioning that phrase boundaries are already defined at this stage [Khomitsevich, Solomennik 2010]. At present we have six intonation types that are reliably derived from the text: completeness, incompleteness, general and special questions and two types of exclamations. This is a reduced set of types from [Volskaya, Skrelin 2009].

At the second stage (after phonetic transcription) allophones receive tone, duration and energy values [Volskaya, Skrelin 1998]. These parameters depend on the voice used and the intonation type. For long and short phrases we use different parameters. For pitch they set declination (based on average pitch) and deviation from it depending on stress and its type. Duration and energy are also specified depending on the position in the phrase and stress as deviations from average.

The parameters are manually adjusted with respect to statistics. So, for a new voice we can immediately apply only a model from a different voice combined with the average characteristics of the new voice. But for accurate tuning we need some additional time to obtain appropriate quality.

3. Hybrid approach

Our hybrid HMM plus unit selection approach is described in detail in [Chistikov, Korolkov 2012]. It combines all the advantages of both methods. Features used for model training and then for generating the necessary physical characteristics of allophones are listed in Table 1:

Table 1. Features used in the statistic intonation model

Allophone features	
Phone before previous	Phone after next
Previous phone	Phone position from the beginning of the syllable
Current phone	Phone position from the end of the syllable
Next phone	
Syllable features	
Previous syllable	Syllable position from the end of the word
Current syllable	Syllable position from the beginning of the sentence
Next syllable	Syllable position from the end of the sentence
Number of phones in the previous syllable	Number of stressed syllables before current syllable in the sentence
Number of phones in the current syllable	Number of stressed syllables after current syllable in the sentence
Number of phones in the next syllable	Vowel type in the current syllable
Syllable position from the beginning of the word	
Word features	
Part of speech of the previous word	Number of syllables in the current word
Part of speech of the current word	Number of syllables in the next word
Part of speech of the next word	Word position from the beginning of the sentence
Number of syllables in the previous word	Word position from the end of the sentence
Sentence features	
Number of syllables in the current sentence	End punctuation type (comma, full stop, etc.)
Number of words in the current sentence	

The speech parameters are obtained from HMMs whose observation vectors consist of mel-frequency cepstral coefficients (MFCC), pitch and duration features; the speech signal is generated by a unit selection algorithm using the obtained speech parameters. The phonetic and linguistic information for the training parameters derives from the speech corpus markup [Prodan et al. 2009].

4. Experiment

In our experiment we follow the recommendations of the state standard specification GOST R 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation” [State standard specification 50840-95 1995]. This standard specification is also applied to speech synthesizers evaluation.

The new female TTS voice Julia was tested. The evaluated synthetic speech variants were the following:

1. Rule-based prosody on a small speech corpus of 20 minutes (with manually corrected labels).
2. Rule-based prosody on a speech corpus of about 2.5 hours of speech (with manually corrected labels).
3. HMM-based prosody on the same speech corpus (2.5 hours, manual correction).
4. Rule-based prosody on a large (6 hours) automatically labeled speech corpus (without manual correction).

17 listeners, 8 female and 9 male aged from 20 to 55 were subjects for the listening test. Among them 11 were trained (i. e. in one way or another closely familiar with synthetic speech) while the other 6 had little or no contact with synthetic speech before.

They were given the task to rate the naturalness of 4 synthetic and one natural speech variants of seven test utterances:

(1) *Если хочешь быть здоров, советует Татьяна Илье, чисть зубы пастой «Жемчуг»!*
(2) *Вчера на московском заводе малолитражных автомобилей состоялось собрание молодежи и комсомольцев.*
(3) *В клумбах сочинской здравницы «Пуща», сообщает нам автоинспектор, обожгли шихту.*
(4) *Тропический какаду — это крупный попугай? Ты не злословишь?*
(5) *Актеры и актрисы драматического театра часто покупают в этой аптеке антибиотики.*
(6) *Нам с вами сидеть и обсуждать эти слухи некогда!*
(7) *Так ты считаешь, что техникой мы обеспечены на весь сезон?*

Ratings could vary from 0 to 5 with a step of 0.1 with clear description of rates (from [State standard specification 50840-95 1995]):

Table 2. Rates and their meaning

Speech characteristics	**Rates**
Natural-sounding speech, some subtle distortion present. Wheeze, rattle missing. High recognizability	> 4.5
Some violation of naturalness and recognizability, a weak presence of one type of distortion (burr, twang, wheeze, rattle, etc.)	3.6–4.5
Audible violation of naturalness and recognizability, presence of several types of distortion (burr, twang, wheeze, rattle, etc.)	2.6–3.5
Constant presence of distortions (burr, twang, wheeze, rattle, etc.). A significant violation of naturalness and recognizability	1.7–2.5
Strong mechanical distortion: burr, twang, wheeze, rattle, etc., mechanical voice. A significant loss of naturalness and recognizability is observed	< 1.7

Five variants of each utterance were given in a random order with possibility of listening for each utterance several times if needed. The obtained ratings are as follows:

Table 3. Evaluation results

TTS type	Mean	Standard deviation
20 min. database	3.6	0.9
Rule-based prosody (2.5 hours)	4.1	0.7
HMM-based prosody (2.5 hours)	**4.3**	0.6
Auto-labeled database (6 hours)	3.7	0.8
Natural speech	4.9	0.1

If we exclude results for two subjects that show more than 20 % deviation from mean ratings and normalize the score to the rating of natural speech (as recommended by the standard specification) we will have 4.4 and 4.5 for rule-based and hybrid approaches respectively. All the synthetic types appeared to be in the same I class (rates from 3.6 to 4.5) of quality (according to [State standard specification 50840-95 1995]).

It should be mentioned that there was a clear connection between the rates and the subject's familiarity with synthetic speech. This may be seen in the diagram below where "a" means "naive" listener and "b" — a listener familiar with TTS (rates were averaged for all of four TTS types):

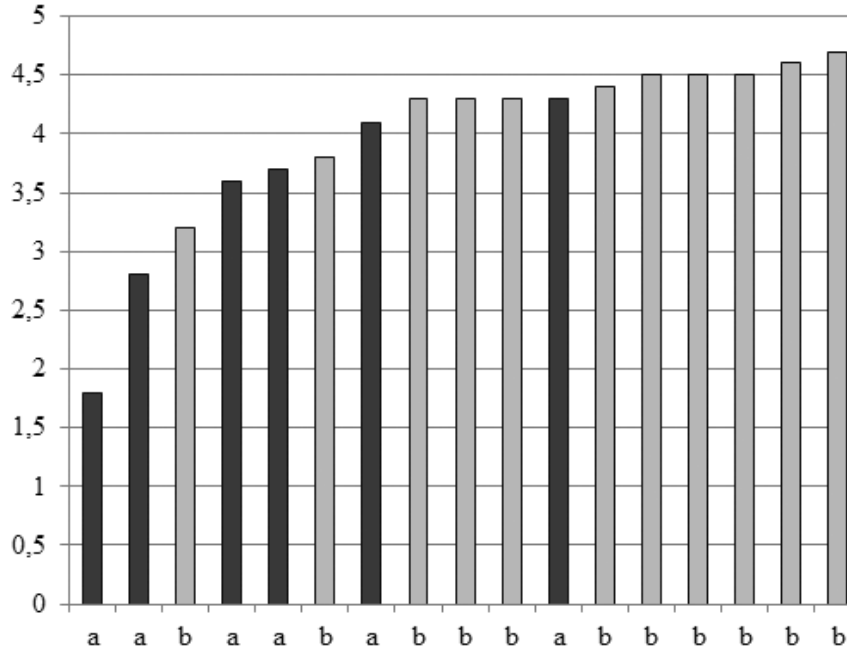


Fig. 1. Mean rates for different types of synthetic speech with respect to familiarity to TTS ("a" — "naive" listener, "b" — familiar to TTS)

We can observe that subjects unaccustomed to synthetic speech tend to give lower rates than others.

5. Conclusion

The obtained results show that by using a hybrid approach combining HMM-based and unit selection speech synthesis we have come close to natural sounding Russian synthetic speech. Also its usage permits fast adaptation of prosodic prediction for a new voice. For these reasons we plan to integrate HMM-based speech parameter generation in our voice-building system [Prodan et al. 2010]. Another important result is that even a small but phonetically balanced [Solomennik, Chistikov 2012] speech corpus can provide us with acceptable quality of synthetic speech.

However, there are still some problems to investigate and several ways of improving our system. Firstly, our evaluation of TTS using the purely automatically labeled speech corpus showed that there is room for improvement in the algorithm for detecting periods of fundamental frequency. Another way to improve prosodic quality is to include more verbal features for model training, primarily special words — potential intonation markers (specific conjunctions, particles etc.). There is also a strong need for a more powerful and at the same time generally accepted method of TTS evaluation in Russian.

References

1. *GOST R 50840-95* (1995), State standard specification 50840-95 “Speech transmission through communication channels. Methods for quality, intelligibility and recognizability evaluation” [GOST R 50840-95. Peredacha rechi po traktam svyazi. Metody ocenki kachestva, razborchivosti i uznavaemosti], Moscow.
2. *Black A. W., Hunt A. J.* (1996), Unit selection in a concatenative speech synthesis using a large speech database, Proceedings of ICASSP 96, Atlanta, Georgia, Vol. 1, pp. 373–376.
3. *Chistikov P., Korolkov E.* (2012), Data-driven speech parameter generation for Russian text-to-speech system, Proceedings of the Dialogue-2012 International Conference № 11 (18), Bekasovo, pp. 103–111.
4. *Khomitsevich O., Solomennik M.* (2010), Automatic pause placing in Russian text-to-speech system [Avtomaticheskaya rasstanovka pauz v sisteme sinteza russkoy rechi po tekstu], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2010”. [Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2010»]. Bekasovo, pp. 531–537.
5. *Krivnova O. F.* (2000), Generation of phrase tone contour in speech synthesis systems [Generaciya tonal’nogo kontura frazy v sistemah avtomaticheskogo sinteza rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2000” [Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2000»], Protvino, Vol. 2, pp. 211–220.
6. *Masuko T.* (2002), HMM-Based speech synthesis and its applications, Doctoral dissertation, Tokyo Institute of Technology, Tokyo.
7. *Oparin I., Talanov A.* (2007), Outline of a New Hybrid Russian TTS System, Proceedings of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, pp. 603–608.
8. *Prodan A. I., Korolkov E. A., Oparin I. V., Talanov A. O.* (2009), Multi-tier markup of speech corpus for hybrid Russian TTS system «VitalVoice» [Osobennosti ispol’zovaniya mnogourovnevoy rametki zvukovogo korpusa unit selection v sisteme gibridnogo sinteza «jivoy golos»], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2009» [Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2009»], Bekasovo, pp. 415–419.
9. *Prodan A. I., Talanov A. O., Chistikov P. G.* (2010), Voice building system for hybrid Russian TTS system «VitalVoice» [Sistema podgotovki novogo golosa dlya sistemy sinteza «VitalVoice»], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialog 2010» [Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii «Dialog 2010»], Bekasovo, pp. 394–399.
10. *Solomennik A., Chistikov P.* (2012), Automatic generation of text corpora for creating voice databases in a Russian text-to-speech system, Proceedings of the Dialogue-2012 International Conference, № 11 (18), Bekasovo, pp. 607–615.

11. *Taylor P. (2009), Text-to-Speech synthesis, Cambridge University Press, Cambridge.*
12. *Volskaya N. B., Skrelin P. A.* (1998), Intonation modeling for speech synthesis [Modelirovanie intonacii dlya sinteza rechi po tekstu], Ufa.
13. *Volskaya N. B., Skrelin P. A.* (2009), System of intonation models for automatic utterance intonation interpretation: functional and perceptual characteristics [Sistema intonacionnyh modeley dlya avtomaticheskoy interpretacii intonacionnogo oformleniya vyskazyvaniya: funkcional'nye i perceptivnye harakteristiki], Proceedings of the third interdisciplinary workshop "Russian spoken speech analysis" (AR3-2009) [Trudy tret'ego mejdisciplinarnogo seminara «Analiz razgovornoy russkoy rechi» (AR3-2009)], St. Petersburg, pp. 28–40.
14. *Zen H., Tokuda K., Masuko T., Kobayashi T., Kitamura T.* (2004), Hidden semi-Markov model based speech synthesis, Proceedings of the International Conference on Spoken Language Processing, Interspeech 2004, Jeju Island, Korea, pp. 1393–1396.

Раздел III.

Анализ тональности

ТЕСТИРОВАНИЕ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ НА СЕМИНАРЕ РОМИП-2012

Четверкин И. И. (ilia2010@yandex.ru),
Лукашевич Н. В. (louk_nat@mail.ru)

МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: РОМИП, извлечение отзывов из блогов, классификация отзывов, анализ тональности новостей

SENTIMENT ANALYSIS TRACK AT ROMIP 2012

Chetviorkin I. I. (ilia2010@yandex.ru),
Loukachevitch N. V. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

In 2012, Russian Information Retrieval Seminar (ROMIP) continued the investigation of sentiment analysis issues. Along with the last year's tasks on sentiment classification of user reviews we proposed two new tasks on sentiment classification of news-based opinions and query-based extraction of opinionated blog posts. For all tasks new test collections were prepared. The paper describes the characteristics of the collections, track tasks, the labeling process, and evaluation metrics. We summarize the participants' results and describe our simple approach for sentiment extraction task.

Keywords: ROMIP, sentiment classification, news-based sentiment analysis, opinion mining

1. Introduction

Recently, the sentiment analysis task received a considerable interest from the research community and industry due to the large amount of sentiment-oriented data in social media and user-generated content. The increased interest in solving the problem of sentiment analysis in social media has led to the rapid development of online reputation management systems, where political parties or companies follow the user comments to reveal the opinion trends and the trends of positive and negative comments. Other applications based on social media analytics intend to reveal new social trends in a region or a social group.

Applications dealing with sentiment analysis for social media require a combination of many different techniques for processing unstructured text data [Bing, 2011; Taboada et al., 2011], e.g. sentiment analysis (including sarcasm detection), opinion mining, information retrieval, classification, summarization, etc.

During the Russian Information Retrieval Seminar (ROMIP, http://romip.ru) cycle in 2012, the second open evaluation of sentiment analysis systems takes place. The tasks of the ROMIP 2012 were closely connected to the social media analytics and consist of:

1. Query-based extraction of opinionated blog posts,
2. Sentiment classification of news-based opinions. News-based opinions are fragments of direct or indirect speech extracted from news articles.
3. Sentiment classification of user reviews.

The first task was very similar to the TREC Blog Track 2006 [Ounis et al., 2007]. Here participants had to find all relevant opinionated posts from the blog collection according to a specific query.

The second and the third tasks had the same objective: to classify texts according to sentiment expressed in them. The main difference was in the domains of texts. Sentiment classification of the news-based opinions differs significantly from classification of user reviews and can be considered as the first step to deep sentiment analysis of news articles.

The last task concerned the sentiment classification of blog posts about different products. There were three different scales in this task:

- two-class classification task,
- three-class classification task,
- five-class classification task.

The rest of this paper is structured as follows. In Sections 2, 3 and 4 we provide a short description of each task and newly created collections used for training and evaluation. Section 5 provides an overview of runs submitted by participants. Concluding remarks can be found in Section 5.

2. Query-Based Sentiment Extraction

This task was a new one for social media analytics in Russian. The main objective was to find opinionated blog posts relevant to a specific query. Figure 1 depicts query results for the digital camera *Canon EOS 6D* with highlighted relevant posts.

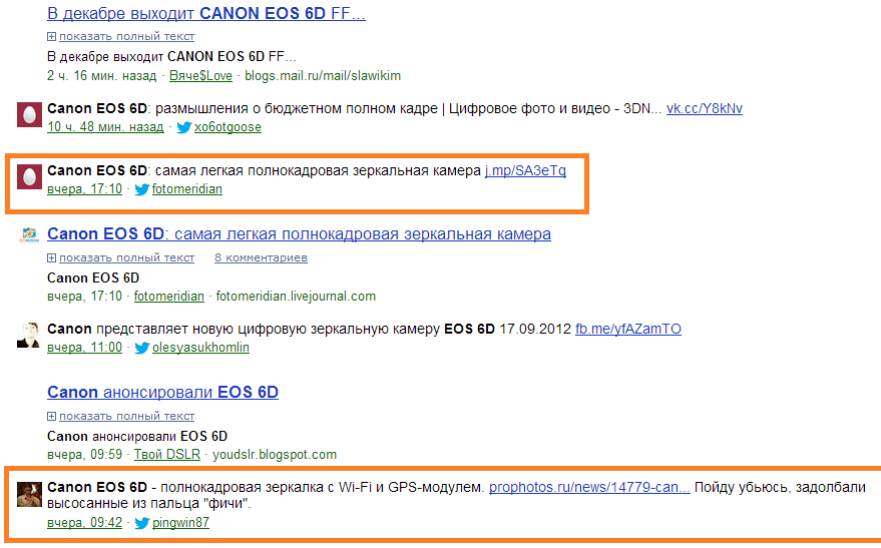


Fig. 1. Query results with highlighted opinion posts

There were three domains: books, digital cameras and movies. For the purposes of query-based extraction two new datasets were released.

The training dataset consists of 874 blog posts about various products (movies, books, digital cameras) with sentiment scores and the list of objects mentioned in this post in some opinionated context. This collection was created from the test set of sentiment classification task during the ROMIP 2011.

To evaluate the quality of sentiment classification and extraction algorithms, we needed additional collections without any authors' scores. We decided to collect blog posts about various entities in three domains (as in ROMIP 2011). For this purpose we used Yandex's Blog Search Engine (http://blog.yandex.ru).

For each domain a list of search queries was manually compiled. There were 2,713 book queries, 1,412 camera queries, and 281 movie queries. Each query was about only one entity (or related objects) from selected domains.

For each query we obtained a set of blog posts (both relevant and irrelevant). Finally results for all queries were merged. The resulting collection included 60,737 posts for entities from various domains.

From this test collection we selected a set of blog posts for human evaluation, which corresponds to randomly selected set of queries: 221 book queries, 235 movie queries and 301 queries about digital cameras.

The task for assessor was the following: for each document-query pair to decide if the document is relevant to a specific query and what sentiment is expressed about the object in the query. In situations where a blog post describes several different objects or some object which is not mentioned in the query, the assessor should mark this document as relevant to the mentioned objects.

In addition assessor was asked to put score on 2, 3 and 5 point scale for each document containing sentiment. Such document would be used in sentiment classification

task. The resulting markup for each document consists of objects mentioned in this document and sentiment scores (on three scales) associated with each object. This year we have only one assessor, but in general framework for sentiment classification is the same as in [Chetviorkin et al., 2012] where the level of annotators' agreement can be found.

The example of the evaluated blog post: *"Девушка с татуировкой дракона" — фильм крутой, вы чего. Недавно американосами был экранизирован, правда шведские книга и фильм круче..*

```
<object main="+">
  Девушка с татуировкой дракона
  <type>F</type>
  <evaluation-2>2 </evaluation-2>
  <evaluation-3>3</evaluation-3>
  <evaluation-5>5</evaluation-5>
</object>
```

3. Sentiment Classification of User Reviews

This task was similar to one from ROMIP 2011. Here the aim was to classify blog posts about different products according to sentiment expressed in documents. We consider different number of classes for classification: two, three and five.

For the sentiment classification tasks we used the same train collections as in the ROMIP 2011 sentiment analysis track [Chetviorkin et al., 2012]. There were three collections: movie and book collections with 15,718 and 24,159 reviews respectively and the digital camera review collection with 10,370 reviews. All reviews have an author's score on a ten-point scale or a five-point scale.

For testing purposes we selected all opinionated blog posts (see Section 2) from the markup which were annotated during the preparation to query-based sentiment extraction task. We obtained 408 sentiment posts about movies, 129 posts about books and 411 posts about digital cameras.

The class distribution for each task was highly skewed. For example, in the two-class task we had 96% of positive reviews for cameras, 87% of positive reviews for books and 81% of positive reviews for movies.

4. News-Based Opinions Classification

This task was new for ROMIP, and it served as the first step for sentiment analysis of whole news articles. Participants should provide sentiment classification of opinions in form of direct or indirect speech extracted from news articles. For each fragment a participant's system should classify it to one of three classes:

1. Opinion expressed in the news fragment is explicitly negative,
2. Opinion expressed in the news fragment is explicitly positive,
3. The news fragment does not contain any opinion.

We prepared a new training set for sentiment classification of direct and indirect speech from news articles, containing 4,260 text fragments. The test collection for the news-based opinion classification task has the same structure as the training set. The main difference between these collections is that test dataset was collected during the other period of time. It contains 124,647 direct and indirect speech fragments from news articles. From whole bunch of text fragments there were evaluated 5,500 quotes for testing purposes.

The example of direct speech is: *"Посредством этих структур десяткам тысяч избирателей предлагают деньги в обмен на паспортные данные и подписи за какого-либо кандидата", — сказал Черненко.*

5. Official metrics

The metrics used for the opinion classification task were *precision, recall, F1-measure, accuracy and average Euclidian distance*. For the first three measures we used traditional (separately for each category) and macro-averaged variants. In query-based sentiment extraction we used two additional measures *Precision@n*, *NDCG@n*.

To give definition to the first part of these metrics, we will use Table 1.

Table 1. Classifier output types

	actual class	
predicted class	**tp_x** (true positive) Correct result	**fp_x** (false positive) Unexpected result
	fn_x (false negative) Missing result	**tn_x** (true negative) Correct absence of result

Precision is the proportion of objects classified as X that truly belong to class X. The macro variant of this feature averages all class precision values.

$$P = \frac{tp_x}{tp_x + fp_x}$$

$$Macro\_P = \frac{1}{|S|} \cdot \sum_{x \in S} \frac{tp_x}{tp_x + fp_x}$$

Recall is the proportion of all objects of class X that is classified by the algorithm as X. The macro variant of this feature averages all class recall values.

$$R = \frac{tp_x}{tp_x + fn_x}$$

$$Macro\_R = \frac{1}{|S|} \cdot \sum_{x \in S} \frac{tp_x}{tp_x + fn_x}$$

F1-measure is the harmonic mean of Precision and Recall. Macro_F1 is the average from all F1-measures of particular classes.

$$Fmeasure = \frac{2 \cdot P \cdot R}{P + R}$$

Accuracy is proportion of correctly classified objects in all objects processed by the algorithm.

$$Accuracy = \frac{tp_x + tn_x}{tp_x + tn_x + fp_x + fn_x}$$

Average Euclidean distance is the average from the quadratic difference between the scores of the algorithm and the assessor scores (average of the assessors' scores).

$$D = \sqrt{\frac{\sum_{i=1}^{n} (q_i - p_i)^2}{n}}$$

In the query-based sentiment extraction we have the ordered list of answers for each query, and the objective was to place all relevant blog posts as close to the beginning of the answer list as possible. Because of different from sentiment classification objective function, we used the other metrics for this task.

Precision@n indicates the number of correct (relevant) objects in the first ***n*** objects in the result set. We assume that *rel(i)* is above zero (e.g. equals to one) in case of relevance of document in position *i* to the query and zero otherwise.

$$P@n = \sum_{i=1}^{n} rel(i)$$

NDCG@n measures the usefulness, or gain, of a document based on its position in the result list, where *IDCG@n* is *DCG@n* of perfect ranking algorithm.

$$NDCG@n = \frac{DCG@n}{IDCG@n} \qquad DCG@n = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_2(i)}$$

6. Results Overview

In all, sixteen groups took part in five tasks. In the review classification task there were 94 submitted runs in the two-class task, 46 runs in the three-class task, and 15 runs in the five-class task. In news-based opinion classification there were 16 runs and only two participants were in the query-based sentiment extraction with 33 runs.

For each classification task we calculated baseline values for all measures. We took as the baseline a dummy classifier that assigns all reviews to the most frequent class.

6.1. Review classification task

Primary measures for evaluating performance in review classification were macro-F1 and accuracy. Table 2–4 shows the best two runs for all tasks. Due to skewness of class distribution in the test collection in some tasks it was difficult to beat the baselines.

Table 2. Two-class classification results

Run_ID	Object	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-17	book	0.749	0.684	0.715	0.884
xxx-1	book	0.666	0.748	0.705	0.821
Baseline	book	0.434	0.500	0.465	0.868
yyy-12	camera	0.589	0.734	0.669	0.895
yyy-13	camera	0.688	0.635	0.660	0.961
Baseline	camera	0.483	0.500	0.491	0.966
zzz-19	film	0.695	0.719	0.707	0.806
zzz-23	film	0.731	0.641	0.683	0.831
zzz-12	film	0.759	0.586	0.661	0.828
Baseline	film	0.404	0.500	0.447	0.809

Table 3. Three-class classification results

Run_ID	Object	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-10	book	0.532	0.591	0.560	0.659
xxx-17	book	0.544	0.554	0.550	0.698
xxx-13	book	0.505	0.532	0.518	0.752
xxx-7	book	0.471	0.501	0.486	0.729
Baseline	book	0.258	0.333	0.291	0.775
yyy-12	camera	0.399	0.602	0.480	0.742
yyy-1	camera	0.440	0.498	0.467	0.523
Baseline	camera	0.285	0.333	0.307	0.854
zzz-11	film	0.569	0.479	0.520	0.694
zzz-6	film	0.486	0.521	0.503	0.596
zzz-1	film	0.487	0.451	0.468	0.650
Baseline	film	0.217	0.333	0.263	0.651

Table 4. Five-class classification results

Run_ID	Object	Avg_Eucl_Distance	Macro_F1	Accuracy
xxx-1	book	1.341	0.402	0.480
xxx-4	book	1.121	0.384	0.473
Baseline	book	1.180	0.131	0.488
yyy-3	camera	1.163	0.336	0.457
yyy-1	camera	1.127	0.288	0.489
yyy-4	camera	1.068	0.207	0.513
yyy-0	camera	1.005	0.245	0.494
Baseline	camera	0.992	0.134	0.504
zzz-2	film	1.388	0.377	0.407
zzz-1	film	1.387	0.323	0.385
Baseline	film	1.720	0.097	0.319

In the review classification task practically all the best results were obtained with machine learning approaches. The best results in the sentiment classification according to F1-measure were obtained by [Blinov et al., 2013] using machine learning approaches on base of SVM and MaxEnt classifiers. The features for classification were semi-automatically crafted on base of the sentiment lexicon from [Chetviorkin & Loukachevitch, 2012] and augmented by collocations with particles and adverbs. Additionally, authors took into account the weighting scheme, the fraction of positive and negative words in texts, exclamation and question marks, emoticons and obscene language. Finally, only the five class classification was conducted and then simple mapping scheme was applied to obtain two or three classes depending on the task.

In [Frolov et al., 2013] the authors use the semantics graph to complement the feature representation for machine learning and make extensive analysis of difficulties occurred during the sentiment classification of book reviews.

In paper [Panicheva, 2013] the rule-based approach using the syntactic structure and an opinion word dictionary is described. The authors obtained the best result according to F1-measure in the two-class movie review classification task. The other rule-based approach is described in [Mavljutov & Ostapuk, 2013]. The authors used the syntactic parser based on context-free grammar and text mining techniques for dictionary construction including objects, proper names, object parts and opinion expressions.

6.2. News-based opinion classification

In this task class distribution was rather balanced in comparison with the review classification task: 41% of quotes were negative, 32% of quotes were positive and 27% of quotes were neutral. Thus the majority of participants performed better than the baseline but the overall quality is still mediocre. The best results according to accuracy and F1-measure could be found in Table 5.

Table 5. News-based opinion classification results

Run_ID	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-4	0.626	0.616	0.621	0.616
xxx-11	0.606	0.579	0.592	0.571
xxx-15	0.563	0.560	0.562	0.582
Baseline	0.138	0.333	0.195	0.413

In opposite to review classification the leaders in the news-based task were knowledge-based approaches. It is due to the absence of a large training collection appropriate for this task because of the broad scope of quotation topics.

The best results in this task were obtained using the lexicon-based system described in [Kuznetsova et al. 2013]. The system has an extensive dictionary of opinion words and expressions obtained using various text mining techniques and manual refinement. Several rules taking into account intensifiers, negation and consequent opinion words were also applied.

The rule-based approach is described in [Panicheva 2013]. The authors used the same system both for sentiment review classification and news-based opinion classification. The system has an extensive rule set and manually crafted sentiment lexicon. The results of this system were second and third in news-based opinion classification.

6.3. Query-based sentiment extraction

In the query-based sentiment extraction task only one participant submitted his result before the deadline. To conduct the track we built our own very simple approach on base of TFIDF measure from [Ageev et al., 2004], which performs at the high level on the standard ad-hoc search task and the five-thousand opinion word list presented in [Chetviorkin & Loukachevitch, 2012].

This sentiment lexicon was constructed in several stages by building the supervised algorithm for sentiment lexicon extraction in the movie domain and further transfer of the model to other domains. The trained sentiment lexicon extraction model was applied to an extensive number of domains and then extracted lexicons were summed up to the single list of sentiment words. This lexicon is proved to be rather clean (P@1000 = 91.4%) to be used is various sentiment analysis tasks and is freely available on the ROMIP web site[1].

To find opinionated blog posts we build two inverted indexes with TFIDF values for all frequent lemmas using posts and headers from the full blog collection. IDF values for all words were calculated using full blog test collection. The third index was built using the aforementioned sentiment word list. For each post in the collection we calculated the fraction of opinion words in it. This fraction serves as opinion weight of each document in the third index.

[1] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

Finally, for each query we calculated weights of all documents in the collection in accordance with the following formula:

$$Weight = \alpha \cdot (\sum_{w \in q} tfidf_w + \sum_{w \in q} tfidf_w^{header}) + (1 - \alpha) \cdot SentiWeight$$

We have experimented with different values of $\alpha = \{0.2, 0.4, 0.5, 0.6, 0.8\}$. The best result was obtained with $\alpha = 0.6$. This result shows the importance of sentiment words in the task of query-based sentiment extraction. All the best results in the resulting Table 6 were obtained using aforementioned approach.

We tried to evaluate the participant results dealing with unlabeled documents as with irrelevant, but it leaded to serious underestimation of the performance. Thus we decided to use only labeled documents, excluding all other documents from the results preserving the order of the remaining documents. The main measures of the performance in this task were NDCG@10 and P@10.

Table 6. Query-based sentiment extraction results

Run_ID	Object	P@1	P@5	P@10	NDCG@10
xxx-0	book	0.3	0.32	0.286	0.305
xxx-9	book	0.3	0.31	0.323	0.304
xxx-8	book	0.25	0.31	0.332	0.298
xxx-6	book	0.25	0.31	0.327	0.302
yyy-9	camera	0.402	0.313	0.302	0.305
yyy-7	camera	0.427	0.319	0.300	0.303
yyy-1	camera	0.402	0.328	0.325	0.226
yyy-2	camera	0.440	0.325	0.311	0.303
zzz-3	film	0.494	0.449	0.438	0.338
zzz-8	film	0. 494	0.448	0.444	0.332

7. Conclusions

ROMIP 2012 is the second seminar which is dedicated to the sentiment analysis problems. In this year we continued the investigation of sentiment analysis tasks, and the list of such tasks was substantially supplemented. Several new collections were created and made available for the research purposes.

The results of this year showed that the sentiment analysis task are still very challenging and attract a lot of researchers from industrial companies and academia.

We find that sentiment classification results are consistent with the results of ROMIP 2011. In query-based sentiment extraction task we found a big role of sentiment lexicons, which is comparable to the role of underlying topic relevance task.

Acknowledgements. We are grateful to Yandex and Anton Pavlov in particular for help with collecting data for research purposes of the seminar. This work is partially supported by RFBR grant N11-07-00588-a.

References

1. *Ageev M., Dobrov B., Loukachevitch N., Sidorov A.* Experimental algorithms vs. basic line for web ad hoc, legal ad hoc, and legal categorization. In Proceedings of RIRES, 2004, (in Russian)
2. *Blinov P., Klekovkina M., Kotelnikov E, Pestov O.* Research of lexical approach and machine learning methods for sentiment analysis. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
3. *Bing L.* Sentiment Analysis Tutorial, AAAI, San Francisco, USA, 2011
4. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011 In Proceedings of Dialog, Bekasovo, 2012, pp. 1–14.
5. *Chetviorkin I.* and Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain In Proceedings of COLING 2012*,* Mumbai, India, 2012, pp. 593–610
6. *Frolov A., Polyakov P., Pleshko V.* Using semantics categories in application to book reviews sentiment analysis. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
7. *Kuznetsova E. S., Loukachevitch N. V., Chetviorkin I. I.* Testing rules for sentiment analysis system. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
8. *Mavljutov R., Ostapuk N.* Using basic syntactic relations for sentiment analysis. Computational Linguistics and Intellectual Technologies. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
9. *Ounis I., de Rijke M., Macdonald C., Mishne G., Soboroff I.* Overview of TREC-2006 Blog track. In Proceedings of TREC-2006*,* Gaithersburg, USA, 2007.
10. *Panicheva P.* Atex. A rule-based sentiment analysis system. Processing texts in various topics. Computational Linguistics and Intellectual Technologies. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
11. *Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M.* Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 2011, pp. 267–307.

RESEARCH OF LEXICAL APPROACH AND MACHINE LEARNING METHODS FOR SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com),
Klekovkina M. V. (klekovkina.mv@gmail.com),
Kotelnikov E. V. (kotelnikov.ev@gmail.com),
Pestov O. A. (oleg.pestov@gmail.com)

Vyatka State Humanities University, Kirov, Russia

Methods and approaches used by the authors to solve the problem of sentiment analyses on the seminar ROMIP-2012 are described. The lexical approach is represented with the lexicon-based method which uses emotional dictionaries manually made for each domain with the addition of the words from the training collections.

The machine learning approach is represented with two methods: the maximum entropy method and support vector machine. Text representation for the maximum entropy method includes the information about the proportion of positive and negative words and collocations, the quantity of interrogation and exclamation marks, emoticons, obscene language. For the support vector machine binary vectors with cosine normalization are built on texts.

The test results of the described methods are compared with those of the other participants of the ROMIP seminar. The task of classification of reviews for movies, books and cameras is investigated. On the whole. The lexical approach demonstrates worse results than machine learning methods, but in some cases excels it. It is impossible to single out the best method of machine learning: on some collections maximum entropy method is preferable, on others the support vector machine shows better results.

Key words: sentiment analysis, lexical approach, machine learning, maximum entropy method, support vector machine, ROMIP

1. Introduction

Text sentiment analysis is an extensively researched area of computational linguistics in last ten years. The main problem of sentiment analysis is an identification of emotional attitude to some object in a text.

Obviously there are many practical applications for sentiment analysis. For example, opinion analysis of target audience helps to reveal strengths and weaknesses of a commercial product. Automatic rating of movie or book reviews enables to make support recommendations for choice of work. Sentiment analysis systems are also used in sociological and political researches, in human-computer interfaces and in other spheres [12, 15].

A majority of researches in sentiment analysis are made for English texts. For a variety of reasons such studies on Russian text collections were not as popular. Recently, however, the situation began to change for the better: for two years a seminar ROMIP [24] has proposed the sentiment analysis tracks including classification of user reviews into 2, 3, and 5 classes. At a seminar in 2012, two new tasks appeared: the classification of news fragments into 3 classes and opinions search on requests.

The purpose of this paper is to present the results of the participation of the team of authors at sentiment analysis tracks at ROMIP-2012. Two approaches were investigated: lexical approach and machine learning approach.

The reminder of this paper is structured as follows. Section 2 gives an overview of current approaches to the problem of sentiment analysis. In section 3 the method of the lexical approach is considered. Section 4 is devoted to the machine learning methods. Section 5 presents the results of experiments at the ROMIP-2012 and their analysis. We provide concluding remarks and findings in Section 6.

2. Existing approaches

There are two main approaches to the problem of sentiment analysis: lexical approach and machine learning approach [19]. In the lexical approach the definition of sentiment is based on the analysis of individual words and/or phrases; emotional dictionaries are often used: emotional lexical items from the dictionary are searched in the text, their sentiment weights are calculated, and some aggregated weight function is applied [5, 9, 19, 20].

In the machine learning approach the task of sentiment analysis is regarded as a common problem of text classification [17] and it can be solved by training the classifier on a labeled text collection [1, 7, 14, 16].

Each approach has its advantages and disadvantages. When using the lexical approach there is no need for labeled data and the procedure of learning, and the decisions taken by the classifier can be easily explained. However, this usually requires powerful linguistic resources (e.g., emotional dictionary), which is not always available, in addition it is difficult to take the context into account.

In the machine learning approach the dictionary is not required (although it can be used), and in practice the methods demonstrate the high accuracy of classification. However, this accuracy is achieved only with a representative collection of labeled training texts and by careful selection of features. At the same time the classifier trained on the texts in one domain in most cases does not work with other domains [8].

When participating in ROMIP 2012, our team set itself the aim to research the capabilities of both approaches for the classification of user reviews.

3. Lexicon-based method

Within the lexical approach in the seminar ROMIP 2012 the lexicon-based method proposed in [22] was used. This method is based on emotional

dictionaries for each domains. The creation of dictionaries was as follows: first of all 60 most impressive emotional Russian words (*хорошо — good, превосходно — great, плохо — bad, отвратительно — disgusting,* etc.) were put in each dictionary and were assigned weights in the range [−5...+5]. Next, each domain dictionary was replenished with appraisal words of appropriate training collection that have the highest weight, calculated by the method of RF (Relevance Frequency) [10]. The weight of a word in the dictionary was also appointed manually from the range [−5...+5]. The quantity of words in the dictionaries varied from 245 to 260.

In addition the dictionaries include word-modifiers (all in all 19, for example, *очень — very, самый — most, несколько — somewhat*, etc.) and the wordnegations (*не, ни, ничего*). The word-modifier changes (increases or decreases) the weight of the following appraisal word by a certain percentage. Word-negation shifts the weight of the following appraisal word by a certain offset: for positive words to decrease, for negative — to increase. Concrete percentages for word-modifiers and the offsets for the word-negations in every dictionary were automatically selected on the basis of cross-validation for the appropriate training collection.

The procedure of the text sentiment classification was carried out as follows. First we calculated the weights of all training texts and of the classified text. The weight of text was defined as the average of the weights of emotional words from the dictionary presented in the text, taking into account the changes made by word-modifiers and word-negations. Thus, all the texts are placed into a one-dimensional emotional space. To improve the accuracy of classification the texts that were too close to the texts of another sentiment class were excluded from the consideration. The proportion of deletions was determined by the cross-validation method.

Then the average weights of training texts for each sentiment class were found. The classified text was referred to the class which was located closer in the one-dimensional emotional space.

4. Machine learning methods

For the research at the seminar ROMIP2012 we chose two machine learning methods, well proved in solving various problems of computational linguistics: Maximum Entropy method (MaxEnt) [2] and Support Vector Machines (SVM) [21].

Both methods use a vector model of the text; to obtain the vector model the only one emotional dictionary (different from the dictionaries in lexical approach) is used.

In this section first the training collections are considered, then the procedure for the building of the dictionary is given, after that the features used in the construction of the vector model of texts are listed, in the conclusion the peculiarities of machine learning methods are shown.

4.1. Training collections

The organizers of the seminar ROMIP2012 granted the following training collections: user reviews of books and movies from advisory service Imhonet[1] and user reviews of cameras from service Yandex.Market[2]. In addition in our research we used a collection of user reviews of movies from ratings "Top 250" and "100 worst" of site Kinopoisk[3] (36 000 reviews). Final training collection contained more than 83 000 reviews.

Preliminarily documents with unknown ratings were removed from the training collection. Ratings of reviews were transferred to the 5-point scale, URL addresses were removed from review contents. Then, for each review such procedures were performed: tokenization, sentence segmentation and morphological analysis; linguistic instruments FreeLing [6] and Mystem [13] were used.

4.2. Dictionary creation

For machine learning methods common emotional dictionary for the three domains: books, movies, cameras was created. Russian sentiment lexicon for product meta-domain [4] was taken as the basis. Subset of words most clearly expressing positive (969 words) and negative (1138 words) emotions were manually selected from it. Next, each word was supplemented with synonyms and antonyms, obtained from Wiktionary[4], after which the number of positive words was 1864, negative — 2215. A similar approach to the completion of the dictionary, only using WordNet, was used in [9].

In order to reflect the nearest context of words, instead of using a list of word-modifiers we included all word collocations of training collection that have the following patterns: *<particle> + <dictionary word>*, *<adverb> + <dictionary word>*, *<particle> + <adverb> + <dictionary word>*, etc. For example, an incomplete list of the resulting fragments with a verb *понравиться (like)* is: {*невероятно понравиться, понравиться, понравиться безумно, не понравиться, очень не понравиться, не очень понравиться, ...*}. As a result, the final dictionary, created by the method described above contained about 19 000 words and collocations.

For each lexical unit of the dictionary conditional probabilities were computed by means of training collection:

$$p(w|score) = \frac{|M_w|}{|N_{score}|}, \tag{1}$$

[1] URL: http://imhonet.ru.

[2] URL: http://market.yandex.ru.

[3] URL: http://www.kinopoisk.ru.

[4] URL: http://www.wiktionary.org.

where w — lexical unit of the dictionary, $score \in \{-, +\}$ — review rating (correspondence between the scales: $\{3, 4, 5\} \to +$, $\{1, 2\} \to -$); N_{score} — set of reviews with the rating *score*; $M_w \subseteq N_{score}$ — set of reviews containing lexical unit w.

4.3. Features

To obtain a vector model of a text for the Maximum Entropy method we used vectors containing 7 components:

1. a component, which takes into account the sentiment of lexical units of the text by means of likelihood ratio; it will be discussed later in details;
2. a component, reflecting the ratio of positive and negative lexical units in the text reduced to the following scale: {*much more negative*, *more negative*, *equally*, *more positive*, *much more positive*};
3. the average number of exclamation marks in the text; concrete numerical values were reduced to a scale: {*absence*, *little*, *middle*, *many*, *very many*};
4. the average number of interrogative marks in the text — it was considered in the same way as the average number of exclamation marks;
5. the ratio of positive emoticons in the text to negative emoticons; numerical value is transferred to the scale: {*less*, *equally*, *a little more*, *more*, *much more*}; emoticons are detected using regular expressions;
6. the ratio of negative emoticons in the text to positive emoticons; it was taken into account similarly to the previous component;
7. the binary feature of presence of obscene language in the text; this feature was granted with the morphological analyzer FreeLing [6].

Let's consider the algorithm of calculating the value of the first component of the feature vector for Maximum Entropy method in details. This component uses the log-likelihood ratio. For each sentence s the expression is calculated:

$$L_s = \sum_{i=1}^{m} \ln \frac{p(w_i|-)}{p(w_i|+)}, \qquad (2)^*$$

where m is the number of words and collocations included in the dictionary, which are found in the sentence s.

The resulting likelihood ratio L for review r then will be:

$$L_r = \frac{\sum_{i=1}^{m} L_i}{n}, \qquad (3)$$

where n is the number of sentences of review r.

---

* В бумажном варианте сборника опечатка: в формуле вместо *ln* (натуральный логарифм) напечатано *h*.

The values of likelihood ratios derived from the formula (3) can be considered as the values of a continuous random variable having a normal distribution $N(\mu,\sigma^2)$. As a component of the vector its values can be represented, if they are discretized. According to the three sigma rule at least 95.4% of all the values of a normal random variable fall within the range $(\mu - 2\sigma,\ \mu + 2\sigma)$. Reviews having values L_r that lie outside of this range can be attributed to the boundary values of a five-point scale of 1 and 5. Let's divide this interval into some quantity of parts. In this case the first component of the feature vector for the review will be a number of interval in which the value L_r falls.

For the Support Vectors Machines in accordance with the results of [23] the reviews were represented as binary vectors with cosine normalization. The dimension of these vectors coincides with the dimension of machine learning dictionary. In this case the i^{th} component of the vector representing the review is equal to one if the i^{th} element of the dictionary is present in a review.

4.4. Classification methods

Among the proposed formulations of the problem of classification at the ROMIP-2012 the most common is the 5-point classification task. If the solution of this task is known, the solution of 2-point and 3-point classification tasks can be automatically received by combining the reviews of different classes. For example, the conversion from 5 classes to 3 classes: $\{4,5\}\to 3$, $\{3\}\to 2$, $\{1,2\}\to 1$; from 5 classes to 2 classes: $\{3,4,5\}\to 2$, $\{1,2\}\to 1$. Guided by these considerations, it was decided to implement the classification with machine learning methods only for 5point scale. Classification decisions for 2-point and 3-point scales were obtained by combining the reviews as described above.

The Maximum Entropy method is implemented using the library SharpEntropy [18]. Conditional probability distribution $p(y \mid x)$, $y \in Y$, $x \in X$ is modeled in the method, where $Y = \{1,2,3,4,5\}$ is the set of ratings, X — the set of input vectors. Such distribution must be consistent with the training data, but also be as even as possible. Mathematical measure of the uniformity of the distribution is the *entropy* [2]:

$$H(y \mid x) = -\sum_{(x,y) \in Z} p(y,x) \log p(y \mid x), \tag{4}$$

where $Z = X \times Y$ is the Cartesian product of the sets X and Y.

From the set of all possible distributions the one that maximizes the entropy is chosen (4):

$$p(y \mid x) = \arg\max_{p(y \mid x) \in Z} H(y \mid x), \tag{5}$$

To implement the Support Vectors Machines the library LIBSVM [11] was used. The selection of the kernel and optimal parameters was conducted. As in [23] the best results a linear kernel with regulating parameter $C = 1$ produced.

5. Experimental results

Let's consider the results of the classification of user reviews at the seminar ROMIP2012. In this section our methods are identified as follows: *Dict* — the lexicon-based method, *MaxEnt* — the Maximum Entropy method; *Svm* — the Support Vectors Machines; *yyy-N* — the code of our results, *xxx-N* — the code of the results of other participants.

Tables 1-3 show the results of the classification of user reviews to 2, 3 and 5point scales (for technical reasons the lexicon-based method for a 5-point scale was not used; instead of it the variant of the Maximum Entropy method was used, which takes into account the uneven distribution of reviews in classes — in Table 3 this method is identified as *MaxEntT*). The values of *precision* (P), *recall* (R), *F1measure* (F1), computed by *macro-averaged* variant, and value of *accuracy* [3] are shown.

Table 1. Two-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
MaxEnt (yyy17)	1 from 43	book	0,749	0,684	0,715	0,884
xxx1	2 from 43	book	0,667	0,748	0,705	0,822
Dict (yyy31)	5 from 43	book	0,627	0,684	0,655	0,798
Svm (yyy7)	13 from 43	book	0,593	0,593	0,593	0,814
Svm (yyy12)	1 from 25	camera	0,589	0,774	0,669	0,895
xxx13	2 from 25	camera	0,688	0,635	0,660	0,961
Dict (yyy6)	9 from 25	camera	0,541	0,626	0,580	0,876
MaxEnt (yyy17)	10 from 25	camera	0,569	0,588	0,579	0,937
xxx19	1 from 26	movie	0,695	0,719	0,707	0,806
MaxEnt (yyy23)	2 from 26	movie	0,731	0,641	0,683	0,831
Svm (yyy13)	5 from 26	movie	0,680	0,642	0,660	0,809
Dict (yyy7)	6 from 26	movie	0,659	0,659	0,659	0,789

Table 2. Three-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
Dict (yyy10)	1 from 18	book	0,532	0,591	0,560	0,659
xxx17	2 from 18	book	0,544	0,554	0,549	0,698

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
MaxEnt (yyy13)	3 from 18	book	0,505	0,532	0,518	0,752
Svm (yyy11)	5 from 18	book	0,454	0,485	0,469	0,690
Svm (yyy12)	1 from 14	camera	0,399	0,602	0,480	0,742
xxx1	2 from 14	camera	0,440	0,498	0,467	0,523
MaxEnt (yyy2)	4 from 14	camera	0,419	0,481	0,448	0,805
Dict (yyy4)	10 from 14	camera	0,370	0,391	0,380	0,745
MaxEnt (yyy11)	1 from 14	movie	0,569	0,479	0,520	0,694
xxx6	2 from 14	movie	0,486	0,521	0,503	0,596
Dict (yyy0)	3 from 14	movie	0,505	0,477	0,491	0,627
Svm (yyy2)	7 from 14	movie	0,454	0,445	0,449	0,640

Таблица 3. Five-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
Svm (yyy1)	1 from 5	book	0,339	0,496	0,402	0,481
MaxEnt (yyy4)	2 from 5	book	0,330	0,460	0,384	0,473
MaxEntT (yyy2)	3 from 5	book	0,219	0,402	0,284	0,380
Svm (yyy3)	1 from 5	camera	0,272	0,441	0,336	0,457
MaxEntT (yyy1)	2 from 5	camera	0,258	0,326	0,288	0,489
MaxEnt (yyy2)	3 from 5	camera	0,246	0,315	0,276	0,470
MaxEnt (yyy2)	1 from 5	movie	0,401	0,352	0,375	0,407
Svm (yyy1)	2 from 5	movie	0,330	0,317	0,323	0,385
MaxEntT (yyy3)	3 from 5	movie	0,318	0,319	0,319	0,382

After analyzing the results the following theses can be concluded:

1. The lexical approach in our study showed significantly worse results than the methods of machine learning. Out of 6 tasks of 2-class and 3class classification in only one case (reviews of books, 3class) the lexicon-based method

was better than the other two methods. Perhaps this is related to the small size of the dictionary (no more than 300 words) and a lack of time to adjust the lexicon-based method.

2. We cannot make an unambiguous conclusion about the benefits of one machine learning method over the other: the Maximum Entropy method always shows the best results on a collection of movie reviews, while the Support Vectors Machines always exceed at a collection of camera reviews. In the case of book reviews for 5-class task the SVM show a slight advantage (2% by value F1), and in the other two problems — on the contrary, the results of the MaxEnt predominate (by 12% and 5%).

All in all, out of 9 tasks the Maximum Entropy method has shown results in 5 tasks higher than the Support Vectors Machines.

3. It is impossible also to summarize the effectiveness of different methods on the parameters of precision and recall: in different tasks all methods show different ratios of these important parameters — sometimes precision dominates, sometimes recall.
4. When the quantity of classes increases the results reduce, although not as dramatically as in the seminar ROMIP2011: for camera reviews the best result for binary classification is 67%, for 5-class task — 34% (43% down) while at the seminar ROMIP 2011 the decrease was 66% (from 92% to 26%). Thus we can conclude that the methods used in the seminar ROMIP2012 are more steady to the increase of the quantity of classes.

In addition to the problems of user reviews classification a new task of the classification of the fragments of direct and indirect speech from news articles was offered at the seminar. It was proposed to perform the classification for 3-point scale. The essential features of this problem should be noted: first, a small amount of the content of each fragment, secondly, the greater thematic variation. To solve this problem we used the MaxEnt and the SVM methods with emotional dictionary, created for reviews classification. Both methods showed low results because of the fact that the dictionary didn't sufficiently reflect the specific emotional terms of the news domain.

6. Conclusion

Thus this paper focuses on two main approaches to the problem of sentiment analysis — lexical approach and machine learning. In the first approach the lexicon-based method developed by the authors was used, which differs from the existing methods by the way of creating both emotional dictionaries for each domain and the algorithm which calculates the weight of texts. Machine learning approach was presented to the Maximum Entropy method and the Support Vectors Machines; it used the technique developed by the authors to create a dictionary and an algorithm for the construction of the feature vector for the Maximum Entropy method.

As a result, as in many other studies, the benefits of machine learning methods are demonstrated, but the lexical approach even with a small dictionary in some cases

shows the best results among the others methods. So, perhaps, the lexical approach should not be rejected, rather, the combination of both approaches is promising.

The participation of our team in seminar ROMIP2012 was very productive: out of 9 reviews of classification task our methods took first place in 8 cases according to the metric of F1, and in one case — the second place.

We would like to thank the organizers for their considerable efforts and express hope for further development of ROMIP which has a major positive impact on research in computational linguistics and information retrieval in Russia.

References

1. *Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R.* (2011), Sentiment analysis of twitter data, Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38.
2. *Berger A. L., Della Pietra S., Della Pietra V.* (1996), A maximum entropy approach to natural language processing, Journal Computational Linguistics, Vol. 22(1), pp. 39–71.
3. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2012) Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", No. 11(18), pp. 739–746.
4. *Chetviorkin I. I., Loukachevitch N. V.* (2012), Extraction of Russian sentiment lexicon for product meta-domain, Proceedings of COLING 2012: Technical Papers, pp. 593–610.
5. *Ding X., Liu B., Yu P. S.* (2008), A holistic lexiconbased approach to opinion mining, Proceedings of the Conference on Web Search and Web Data Mining (WSDM), pp. 231–240.
6. *FreeLing 3.0* An open source suite of language analyzer, available at: http://nlp.lsi.upc.edu/freeling/.
7. *Go A., Bhayani R., Huang L.* (2009), Twitter sentiment classification using distant supervision, Association for Computational Linguistics, pp. 30–38.
8. *He Y.* (2012), Incorporating sentiment prior knowledge for weakly supervised sentiment analysis, ACM Transactions on Asian Language Information Processing, Vol. 11(2).
9. *Hu M., Liu B.* (2004), Mining and summarizing customer reviews, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004), Seattle, pp. 168–177.
10. *Lan M., Tan C. L., Su J., Lu Y.* (2009), Supervised and traditional term weighting methods for automatic text categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(4), pp. 721–735.
11. *LIBSVM* — A library for support vector machines, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
12. *Liu B.* (2012), Sentiment analysis and opinion mining, Morgan & Claypool Publishers.
13. *Mystem,* available at: http://company.yandex.ru/technology/mystem.

14. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.
15. *Pang B., Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, No. 2.
16. *Saif H., He Y., Alani H.* (2012), Alleviating data sparsity for twitter sentiment analysis, Workshop: The 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at World Wide Web (WWW), Lyon, France.
17. *Sebastiani F.* (2002), Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34, pp. 1–47.
18. *SharpEntropy*, available at: http://www.codeproject.com/Articles/11090/Maximum-Entropy-Modeling-Using-SharpEntropy.
19. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexiconbased methods for sentiment analysis, Computational Linguistics, Vol. 37(2), pp. 267–307.
20. *Turney P.* (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424.
21. *Vapnik V.* (1998), Statistical learning theory, New York, Wiley.
22. *Klekovkina M. V., Kotelnikov E. V.* (2012), The automatic sentiment text classification method based on emotional vocabulary [Metod avtomaticheskoj klassifikatsii tekstov po tonalnosti osnovannyj na slovare èmotsionalnoj leksiki], Digital libraries: advanced methods and technologies, digital collections (RCDL-2012) [Èlektronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii], Pereslavl-Zalessky, pp. 118–123.
23. *Kotelnikov E. V., Klekovkina M. V.* (2012), Sentiment analysis of texts based on machine learning methods [Avtomaticheskij analiz tonalnosti tekstov na osnove metodov mashinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2012”], Bekasovo, pp. 753–762.
24. *Russian* Information Retrieval Evaluation Seminar (ROMIP). URL: http://romip.ru/

ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ ФИЛЬТРОВ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ О КНИГАХ

Фролов А. В. (anton_frolov@rco.ru),
Поляков П. Ю. (pavel@rco.ru),
Плешко В. В. (vp@rco.ru)

ООО «ЭР СИ О», Москва, Россия

В данной работе исследуется метод использования семантических фильтров в качестве классификационных признаков для решения задач классификации отзывов о книгах на 2 (положительный, отрицательный) и 3 (положительный, отрицательный и нейтральный) класса. Кроме того, проанализированы основные ошибки и подводные камни, которые могут встречаться в задачах подобного рода.

Ключевые слова: анализ мнений, определение тональности, автоматическая классификация, машинное обучение, извлечение классификационных признаков, метод опорных векторов, регрессия

USING SEMANTIC FILTERS IN APPLICATION TO BOOK REVIEWS SENTIMENT ANALYSIS

Frolov A. V. (anton_frolov@rco.ru),
Polyakov P. Yu. (pavel@rco.ru),
Pleshko V. V. (vp@rco.ru)

RCO LLC, Moscow, Russia

The paper studies the use of fact semantic filters in application to sentiment analysis of book reviews. The tasks were to divide book reviews into 2 classes (positive, negative) or into 3 classes (positive, negative, and neutral). The main machine learning pitfalls concerning sentiment analysis were classified and analyzed.

Key words: opinion mining, sentiment analysis, document categorization, machine learning, classification feature extraction, support vector machine, regression, two-class classifier, multi-class classifier

Introduction

The classification problem of goods reviews is very important today. This fact is supported by increased popularity of commercial resources offering services for monitoring social networks and blogs (i.e. [7]). However, until recently there were no public collections in Russian language that could be used to test research methods. New ROMIP tracks devoted to classification of books, films and digital cameras reviews, are to fill this gap.

This paper studies methods for solving the book reviews classification problem, involving 2 classes (positive, negative) and 3 classes (positive, negative, neutral), within the framework of ROMIP 2012 [3].

Problem specification

The participants were offered a training collection, composed of blog users reviews of books of different genres (24,160 reviews in total). Each review was graded on a decimal scale. It was decided to participate in the following tracks: classification of book reviews into 2 classes (positive, negative) and 3 classes (positive, negative, neutral). In the former case the task was to divide reviews into positive and negative, in the latter — into positive, neutral (the review mentions both positive and negative features) and negative.

Generalizing facts with semantic filters

It was decided to improve the linguistic approach based on fact extraction which was presented in [6] and demonstrated good results on the last year track. Therefore, we analyzed last year results and tested a hypothesis that the training collection was too small to ensure that individual facts have high enough frequencies to be used as good classification features. A possible solution for this problem is an application of semantic filters that allow combining several facts into one class.

Recall, that fact extraction is performed by the means of semantic templates. Semantic template is a directed graph with certain restrictions applied to its vertices. The restrictions can be applied to part of speech, name, semantic type, syntactic connections, etc. (see Fig. 1). Fact extraction is performed by finding a subgraph of a sentence syntax tree which is isomorphic to the template (with all restrictions applied).

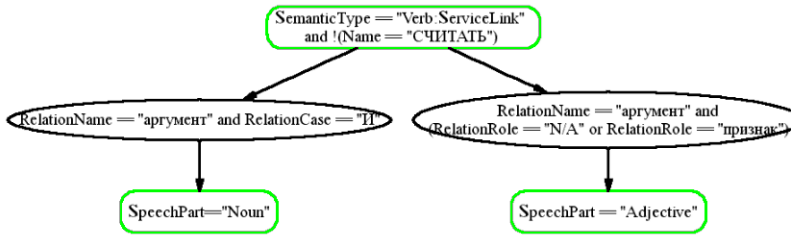


Fig. 1. Semantic template for detecting book review tonality

Moreover, facts can be generalized by the use of special dictionaries (so-called filters), containing synonyms for positive, negative and neutral appraisals. The main flaw of this approach is the necessity of manual selection of terms for the filters. It makes filter generation a labor-intensive task that requires help of a linguistic expert. On the other hand, the expert may form only the most basic vocabulary and all the additional terms can be added by an automated system. It was decided to rely on this method.

The idea was implemented as follows: we took filters used in last year track and expanded them by terms that the system was able to find independently. For more detailed explanation of how fact extraction and filters application works see [6].

New vocabulary was constructed as follows:

The training collection was processed by a system tuned to fact extraction. Then, the collection was classified by using the obtained facts as the only classification features. It was decided to use Naïve Bayes classifier with Poisson function as the PDF for words [5]. The system considered the profiles for each class individually and used filled frames slots to form the word lists for the filters vocabulary. Then, the lists were filtered against a frequency threshold and merged with the existing ones. Also, for better quality we used the vocabulary published by ROMIP organizers [2]. If ROMIP vocabulary contained a fact slot, the slot's weight was multiplied by 10.

Table 1. Filter example

Subject	Quality Verb	Quality Emotion	Quality Adjective
КОНЕЦ КНИГИ	УБИТЬ	ЖДАТЬ	УМОПОМРАЧИТЕЛЬНЫЙ
КОНЦОВКА	ИДТИ	УБИТЬ	ОПТИМИСТИЧНЫЙ
ФИНАЛ	РАСТЯГИВАТЬСЯ	НЕ ЖДАТЬ	ДУРАЦКИЙ
РАЗВЯЗКА	СДЕЛАТЬ	ИДТИ	ЗАКРЫТЫЙ
ХЭППИЕНД	НЕ ПОНРАВИТЬСЯ	РАСТЯГИВАТЬСЯ	УТОМИТЕЛЬНЕЙШИЙ
ХЭППИ ЭНД		НЕ ПОНРАВИТЬСЯ	СКУЧНЫЙ
ХЭППИ		НАЗВАТЬ	ПЕЧАЛЬНЫЙ
ХЭППИ-ЭНД			

Table 1 shows an example of an automatically filled filter, which combines several facts into one class: "negative review concerning book ending". In this case, facts with four slots (subject, quality verb, quality emotion, quality adjective) will be merged if the contents of their corresponding slots belong to the same filter.

Despite rare errors (in example — term "оптимистичный" has been included in a filter for the negative class) most of the vocabulary is adequate. Furthermore, the quality of selected terms can be improved by increasing representativeness and size of the training sample.

Thus, we obtained fact classes for identifying tonality of reviews concerning characters, language, storyline, and author evaluations.

Classification methods

To obtain a good training set two of our experts independently evaluated the collection and marked reviews as being mostly positive, mostly negative or having both positive and negative features. Every expert evaluated about 4,000 reviews with most of them been marked as positive. The experts agreement equals $r \sim 0.8$, where r is Pearson's correlation coefficient. Two approaches were used for the experiment.

In the former approach the classifier was trained using blog users' evaluations and these evaluations were used to build a linear regression model (SVM-Light implementation, see [4]). Then, this model was used to compute weights of documents from the training collection and to determine thresholds for relating documents to corresponding classes so that the difference between the system's partitioning and the experts' partitioning is minimized (F-measure was used as an utility function).

In the latter approach the classifier was based on the training set, formed by the experts. Following classification methods were used:

- Linear classifier with the learning stage been conducted for each class independently (SVM-Light implementation, see [4]). In the case when the same document is classified as being a member of several classes, we select the class where the document has the greatest weight.
- Linear classifier, with the learning stage been conducted independently for 2 classes (positive and negative), that was used to classify documents into 3 classes (SVM-Light implementation, see [4]). In this case, a document is marked as being a member of the neutral class if the classifier considers it as being a member of both negative and positive classes.

Results

This paper studies the results of 4 runs devoted to classification into 2 classes and 4 runs devoted to classification into 3 classes. The runs are parameterized with the classifier's type:

- SVM: support vector machines method with “one against all” partitioning
- Regression: linear regression model

and classification features sets:

- Base: classification features are lemmas (single words) and themes (word-combinations)
- Hybrid: fact classes are used in addition to Base features

We used F1-measure as a primary evaluation metric [1]. Additionally, for convenience, recall, precision and accuracy are also present in the tables.

Table 2. Runs for 2 classes

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.676425	0.620273	0.647133	0.86046
Hybrid SVM	0.577041	0.552521	0.564515	0.82945
Base Regression	0.627363	0.627363	0.627363	0.82945
Hybrid Regression	0.605004	0.634454	0.619379	0.79845

The data given in Table 2 indicates that the Base SVM classifier demonstrates the best result. The explanation for this fact is given in the next section of this paper. Also, it is evident that in the case of the hybrid model, the regression based classifier shows better result than SVM.

Table 3. Runs for 2 classes (detailed)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.898305	0.946428	0.921739	0.454545	0.294117	0.357143
Hybrid SVM	0.881355	0.928571	0.904348	0.272727	0.176471	0.214286
Base Regression	0.901785	0.901786	0.901786	0.352941	0.352941	0.352941
Hybrid Regression	0.905660	0.857143	0.880734	0.304348	0.411765	0.350000

Table 3 indicates that correct identification of negative reviews was the most difficult task for the classifier. The complexity of this task can be explained by the following factors:

1. Most of reviews in both test and training collection were positive: 112 (positive) vs 17 (negative).
2. The size of the test sample was small: as little as 129 documents. This, in addition to factor 1, leads to the result being statistically biased.
3. A significant part (8 out of 17) of negative reviews did not contain explicit negative opinions. Such reviews were correctly identified as neutral under classification into 3 classes.

It is worth mentioning, that the test collection was evaluated out by only one expert which results into increased bias in the final result.

The classification into 3 classes demonstrates completely different picture: SVM performs better than the regression model.

Table 4. Runs for 3 classes

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.544343	0.554074	0.549165	0.697674
Hybrid SVM	0.450879	0.467037	0.458816	0.666666
Base Regression	0.354825	0.333703	0.343940	0.542636
Hybrid Regression	0.354826	0.333704	0.343941	0.542636

Table 5. Runs for 3 classes (neutral class)

	P-neu	R-neu	F-neu
Base SVM	0.891566	0.74	0.808743
Hybrid SVM	0.870588	0.74	0.800000
Base Regression	0.857142	0.72	0.782608
Hybrid Regression	0.864864	0.64	0.735632

Table 6. Runs for 3 classes (negative, positive)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.341463	0.70	0.459016	0.400000	0.222222	0.2857140
Hybrid SVM	0.282051	0.55	0.372881	0.200000	0.111111	0.1428571
Base Regression	0.147058	0.25	0.185185	0.090909	0.111111	0.1000000
Hybrid Regression	0.116279	0.25	0.158730	0.083333	0.111111	0.0952380

Results analysis

The result was strongly affected by several properties of the test collection. Namely: collection's small size (twice as small as the last year collection) and strong odds towards neutral reviews (positive, in case of binary classification). Additionally, negative reviews are biased: about half of them are devoted to the same book, namely, "Angels and demons" by Dan Brown.

The agreement between our expert and ROMIP expert equals $r = 0.78$

As it is possible to see from the tables, negative reviews posed the main problem for the classifier. We analyzed and classified errors, made by the system. They can be divided into following categories:

The author mostly retells the storyline. In this case, the text may contain enough noise terms for the classifier to make an error.

Despite the author speaks about book's positive features, the final evaluation is negative, e.g: "Сюжет есть. И интрига присутствует. А вот то, как разворачиваются действия — не вдохновляет ни коим образом." As a result, positive terms overweight negative terms only due to their number. The methods employing fact extraction are particularly vulnerable to errors of this kind. The reason is that it is much more difficult to gather enough statistics for facts than for lemmas.

Although the author mentions positive reviews by other people, his/her own evaluation is negative, e. g. "С сожалением сообщаю: не для моих мозгов. Говорят, книга очень хорошая. Промолчу."

The presented system used a semantic filter rather than a regular stop-words list, i. e. all numerals and auxiliary words were filtered out. This method demonstrated good results in classification of reviews from Imho-net. However, current track contains blog posts rather than ordinary reviews. Blog posts vocabulary contains significantly more noise terms that cannot be filtered out by semantic filters solely.

It is worth mentioning, that we used "one against all" partitioning and chose class where the document had the greatest weight. As a result, many incorrectly classified documents had negative weight for both classes. In classification into 3 classes the system correctly identified such reviews as being neutral.

Table 7. Comparison of last year and this year results

	Expert 1 F-macro	Expert 2 F-macro
New hybrid SVM	0.503129181	0.500560892
Old hybrid SVM	0.467705308	0.484938518
Base regression	0.490300000	0.499800000

It is evident, that the classifier that employs fact extraction demonstrates worse results than the basic one. We suspected that the reason is that the bias of the collection. To prove it we conducted experiments with the last year collection. It turned out that the new classifier demonstrated improvement in classification into 3 classes in comparison to hybrid system and even regression method [6] that was the leader among all the systems participated in the last year track. It follows that the new classifier performs better than the old ones, provided the collection is not biased.

Possible improvements

The above mentioned problems can be solved by changing the set of classification features. First of all, it is important to be able to distinguish the summarizing assessment. Indeed, such reviews mostly contain retelling of a storyline or an irrelevant discussion. The same time the statements that truly characterize the review are contained in a few sentences in the beginning or the end of the text.

Secondly, it is desirable to be able to identify the object being reviewed. The point is the same review can discuss several books simultaneously, e. g.: "Сегодня

я читал X и мне не понравилось. Гораздо хуже замечательной книги Y, которую я читал вчера". If the object is not specified the system should be able to identify it itself.

Thirdly, in classification into three classes the author's opinion should be distinguished from outer sources opinions ("говорят книга хорошая, но мне не очень понравилась"). In this case, the author's opinion obviously has a greater weight. In classification into three classes this factor is not so critical and different sources can be assigned with similar weights.

Conclusion

We tested several methods of classification into 2 and 3 classes and improved the linguistic approach, based on application of evaluative vocabulary, by application of automated filters generation. Additionally, main errors made by the classifier were analyzed and categorized. Finally, the direction for future work has been set.

References

1. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2011), Sentiment Analysis Track at ROMIP 2011, available at: www.dialog-21.ru/digests/dialog2012/materials/pdf/83.pdf.
2. *Chetviorkin I., Loukachevitch N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012.
3. *Chetviorkin I., Loukachevitch N.* (2012), Sentiment analysis track at ROMIP'12.
4. *Joachims T.* (1998), Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Machines, MIT Press: Cambridge, MA.
5. *Pleshko V. V., Polyakov P. Yu., Ermakov A. E.* (2009), RCO at RIRES 2009 [RCO na ROMIP 2009]. Trudy ROMIP 2009 [Proc. ROMIP 2009]. Petrozavodsk, Saint Petersburg, pp. 122–134.
6. *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.* (2012), Research of applicability of thematic classification to the problem of book review classification. Dialog '12. Naro-Fominsk.
7. *Sentiment 140* (2012), Available at: www.sentiment140.com.

ТЕСТИРОВАНИЕ ПРАВИЛ ДЛЯ СИСТЕМЫ АНАЛИЗА ТОНАЛЬНОСТИ

Кузнецова Е. С. (knnika@yandex.ru)
ГК «Геострим», Москва, Россия

Лукашевич Н. В. (louk_nat@mail.ru),
Четверкин И. И. (ilia2010@yandex.ru)
МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: анализ тональности, общественно-политическая область, РОМИП, правила

TESTING RULES FOR A SENTIMENT ANALYSIS SYSTEM

Kuznetsova E. S. (knnika@yandex.ru)
GK «Geostream», Moscow, Russia

Loukachevitch N. V. (louk_nat@mail.ru),
Chetviorkin I. I. (ilia2010@yandex.ru)
Lomonosov Moscow State University, Moscow, Russia

The paper is devoted to testing rules useful for sentiment analysis of Russian. First, we describe the working principles of the POLYARNIK sentiment analysis system, which has an extensive sentiment dictionary but a minimal set of rules to combine sentiment scores of opinion words and expressions. Then we present the results achieved by this system in ROMIP-2012 evaluation where it was applied in the sentiment analysis task of news quotes. The analysis of detected problems became a basis for implementation of several new rules, which were then tested on the ROMIP-2012 data.

Key words: sentiment analysis, socio-political domain, ROMIP, rules

Introduction

In recent years sentiment analysis is one of the most rapidly developing branches of computational linguistics. A lot of studies in this area have been conducted for English language, also there already exist working systems (for example, TwitterSentiment), different resources (WordNet, SentiWordNet and others) and natural language processing tools, which simplify the task (Liu, 2010; Pang, Lee, 2008).

For Russian language the sentiment analysis task is complicated by the lack of analogues of the above mentioned resources and tools in Russia and a significantly smaller number of related research papers. Recently, there is some growth of interest in the sentiment analysis task for Russian, both from organizations and researchers. In connection with this interest, the evaluation of sentiment analysis systems for Russian language was organized in 2011 and 2012 as a part of the Russian Information Retrieval Seminar — ROMIP (Chetviorkin et al., 2012; Chetviorkin, Loukachevitch, 2013).

There are two main approaches to the sentiment analysis task (Liu, 2010; Pang, Lee, 2008):

- Machine learning methods, when a system is trained using a labeled text collection,
- Dictionary-based methods, which are based on usage of opinion lexicons, linguistic rules and taking into account contexts of the words (Taboada et al., 2011; Pazelskaya, Solovyev, 2011).

In this paper we investigate the influence of linguistic rules on the quality of sentiment analysis in Russian on the example of the POLYARNIK system. At first, we describe the basic principles of the system, which has an extensive dictionary of opinion words and expressions, but a minimal set of rules to combine them (section 2). Then we provide the results of the POLYARNIK at ROMIP-2012 news-based sentiment analysis task and analyze revealed problems (section 3). Finally, we add a set of rules taking into account multiword sentiment expressions, multiword operators, irrealis markers and evaluate the impact of these added rules to the performance of the system (section 4).

1. Linguistic rules in sentiment analysis systems

The most wide-spread linguistic rules used in the sentiment classification are as follows:

- use of operator words, which increases the base score of a sentiment word (*очень, значительно*) or revert the score to the opposite (*не, нет*…) (Taboada et al., 2011; Pazelskaya, Solovyev, 2011; Chetviorkin, Loukachevitch, 2010);
- aggregation of the scores of sentiment words (Taboada et al., 2011; Liu, 2010)

The rule-based system for sentiment analysis of texts in English is described and tested in the detailed study (Taboada et al., 2011). Additionally, the study describes several rules that take into account the appearance of irrealis markers (words indicating that a certain situation or action is not known to have happened) — the score of sentiment words that occur in the same fragment with them is nullified. The list of these markers includes modals, conditional markers (*if*), some negative sentiment

words like *any*, *anything*, certain verbs (for example, *expect* and *doubt*), questions, and words enclosed in quotes.

The analysis of rules for sentiment classification systems from various Russian and English-language studies on the basis of ROMIP-2012 training collection of news quotations is conducted in the study (Kuznetsova, 2012).

2. POLYARNIK system for sentiment classification of socio-political texts

Sentiment classification of socio-political texts is distinguished from the other domains by the fact that this domain includes a wide variety of topics. This fact hampers the creation of a training collection for machine learning algorithms within realistic time limits. For this reason it is necessary to use engineering technologies for creating sentiment dictionaries, searching dictionary items in texts and combining them with linguistic rules.

The four main dictionaries for socio-political sentiment classification are presented in the POLYARNIK system:

- **Positive words and expressions dictionary** (*новаторский, нравственный*, *огромный потенциал, очень убедительный* — about 7 thousand words and expressions*)*. The +1 value was assigned to the most part of entries in this dictionary;
- **Negative words and expression dictionary** (*осквернить (desecrate), отсутствие диалога (absence of the dialog)* — about 15 thousand words and expressions*)*. The −1 value was assigned to the most part of these entries;
- **Dictionary of operators**, which can revert or intensify the value of the sentiment expressions. Operators can intensify the base value of expressions (*очень, значительно (very much, considerably),* etc.) or revert the value to the opposite (*не, отменить* (*not, abolish*) etc.). There are about 140 operators in the dictionary;
- **Dictionary of stop-expressions** — the list of multiword expressions containing sentiment words, but not expressing any overall sentiment, for example, *фонд* ***эффективной*** *политики (foundation of effective politics) etc.* This dictionary contains about 250 items.

To create these dictionaries the following procedure was used:

At the first stage the list of sentiment word candidates was made on the basis of news text collection *Coll_news* (2 million documents). For this purpose the sentiment words and expressions extracted for the movie domain (see Chetviorkin, Loukachevich, 2010) were taken. Documents that contain more than 3 different sentiment words from this list were chosen from the news collection *Coll_news*. It was supposed that if there were at least three sentiment expressions in a text, then this text was likely to contain more sentiment words. In this way the sub-collection *Coll_Sent* with presumably high share of sentiment words was constructed. Lemmas (words in a dictionary form) that appeared at least in 100 documents were extracted from this *Coll_Sent* sub-collection. As a result, thirty thousand lemmas were obtained.

So-called *weirdness* formula (Ahmad et al, 1999) was applied to the sub-collection lemmas:

$$Weirdness = \frac{P_s(w)}{P_g(w)}$$

where $P_s(w)$ — probability of the word appearance in documents of *Coll_sent* sub-collection, $P_g(w)$ — probability of the word appearance in documents of *Coll_news* collection.

The first ten thousand lemmas ordered by *weirdness* were manually refined and it appeared that these lemmas contained more than 30% of sentiment words. These sentiment words formed the first version of the sentiment dictionary.

At the second stage derivational variants (adverbs formed from adjectives; participles formed from verbs, etc.) of the described sentiment words were added to the obtained dictionary.

At the third stage POLYARNIK system was used in the sentiment analysis of news articles, and during the analysis of the results the dictionaries were specified and supplemented. For this purpose we selected big articles with a large share of sentiment words, and thus a single article could become a source of various additional sentiment expressions.

The algorithm of assigning the sentiment value to a document or a text fragment in POLYARNIK system was as follows:

- The words in the processed text are matched with the dictionaries. If an ambiguous match between the text fragment and dictionary entries is found, the longest entry is chosen;
- When an operator word is found at a distance of 5 words (this parameter of the algorithm can be changed), the system is looking to the right of it for a sentiment word, to which this operator can be applied. The search is performed until a punctuation mark is found.

3. The POLYARNIK system in ROMIP-2012 news-based opinion classification task

One of the ROMIP-2012 evaluation tasks was the task of sentiment classification of news-based opinions (direct and indirect speech, further *quotations*) extracted from news articles. The task was to classify quotations as neutral, positive or negative speaker comment about the topic of the quotation. The example of a negative quote (opinionated expressions are underlined): *По мнению эксперта, глава белорусского государства больше всего* ***боится (afraid of),*** *что страну все-таки* ***лишат права*** *(**deprive the right**) провести чемпионат мира по хоккею в 2014 году.*

The results of POLYARNIK system in the news-based sentiment classification are shown in Table 1.

Table 1. The best results in ROMIP-2012 news-based sentiment classification task

Run_ID	Macro_P, %	Macro_R,%	Macro_F1,%	Accuracy,%
POLYARNIK	**62.6**	**61.6**	**62.1**	**61.6**
xxx-11	60.6	57.9	59.2	57.1
xxx-15	56.3	56.0	56.2	58.2

So, POLYARNIK sentiment analysis system obtained the best results in the news-based sentiment classification task, and to our opinion this fact can be explained from the above-described technique of dictionary creation, which allowed us to extract an actual sentiment lexicon for the socio-political domain.

To evaluate additional rule types, which could improve our existing sentiment classification system, we analyzed the reasons of incorrect classification on the basis of 140 news quotations from the training collection. Altogether we have analyzed 40 quotations, which were incorrectly classified by the base version of POLYARNIK system.

Most errors in the news-quotation sentiment classification occurred due to the lack of sentiment words and expressions in the system dictionary — 16 quotations (40 %), including the lack of sentiment expressions or stop-expressions — 13 quotations (32.5%). Examples of these expressions are shown in Table 2.

At the same time, the dictionary can also contain wrong (or not always relevant) data about sentiment of a word or expression. Thus, in the system dictionary it was indicated that expression *не думать (not to think*) has negative sentiment, while this expression did not have this sentiment in one of the analyzed examples: *Сам игрок заявил, что пока не думает о переходе в другой клуб (The player said that he does not think of moving to another club…).*

Table 2. The examples of expressions that are useful to include in a sentiment dictionary

Expression	Automatically assigned sentiment	Real sentiment
индустрия гостеприимства (hospitality industry)	Positive	neutral
падение личности (fall of personality)	Neutral	negative
заработать удаление (remove from the field)	Neutral	negative
чувствовать себя как дома (feel at home)	Neutral	positive
не видеть в этом смысл (do not see the point)	Neutral	negative

Table 3. Exploring the use cases of the linguistic rules

Quotations	Automatically assigned /real sentiment	Rule needed to improve analysis
Он заявил, что речь идет о **прискорбном недоразумении**, ведь он всегда считал, что литература и **искусство** должны служить **морали**	0/–	Irrealis factor: the sentiment score of the fragment going after expression like «*думал, что*» (*thought that*) should be reduced
Секретарь президиума генсовета «Единой России», зампредседателя Госдумы Сергей Неверов в субботу заявил, что партия **не боится раскола в** связи с появлением в ней разных идеологических платформ	0/+	The negation operator should be applied to a group of sentiment words rather than to a single word (*do not afraid of a split*)
Алла Джиоева заявила, что не пойдет на выборы и в качестве избирателя, потому что «не видит в этом смысла и **не верит** в их **объективность**»	0/–	The negation operator should be applied to a group of sentiment words rather that to a single word (*do not believe in their objectivity*)
Один из руководящих сотрудников Hermitage Capital, жизни которого уже **не раз угрожали** из России, утверждает, что явное сотрудничество между британской полицией и российским МВД подвергает его семью **опасности**	+/–	The negation operator is a part of the other operator «*не раз*» (*not a single time*), which should be applied as an intensifier

Additionally, authors of the paper did not agree with assessor classification of 5 quotations (12.5%). Therefore, the misclassification of 25 quotations from 40 examples came from either the dictionary or from the complexity of short utterance sentiment classification.

At the same time it was found that the classification quality of 4 quotations could be improved by developing existing rules and implementing new rules. Table 3 shows the examples of such quotations and indicates rules that could be applied here. The sentiment words and expressions found by system are underlined. The rule definitions are partially based on rules described in the work (Kuznetsova, 2012).

Therefore, as we can see from the analysis, using the additional rules can improve the performance for 4 quotations, what equals to more than 3% of classification accuracy growth, and quotations appear to be the appropriate material for testing different types of rules proposed in the literature.

4. Testing various rules for news-based sentiment classification

To evaluate the impact of linguistic rules to the sentiment classification quality, the following scheme was used. We implement a certain set of rules and test its performance on the ROMIP-2012 quotation training set. After all rules are tested, we evaluate the sentiment analysis system quality on the separate ROMIP-2012 test set.

The set of linguistic rules, which was tested on available quotation sets, can be divided in two groups. The first group is the consideration of various combinations of sentiment words and operators. The second is a group of rules considering the irrealis factor of a text fragment and reducing the sentiment score in such fragment. Further, a fragment (=clause) is a part of a sentence between two punctuation marks.

The first group contains the following set of rules (referred to below as *algo*)

1.1. If an operator word is a part of a longer stop-word or sentiment expression, it does not act as an operator;
1.2. If a group of operators appears together, their scores are multiplied;
1.3. If there is unknown hyphenated word appeared in a text fragment, it is divided in two words and their scores are considered separately;
1.4. If there is a sentiment word sequence, and a negative word appears among them then the score of the whole sequence becomes negative, otherwise positive;
1.5. An operator is applied to the resulting score of a group of sentiment words.

The second group contains the following set of rules (referred below as *rules*). The rules were modified from (Kuznetsova, 2012):

2.1. If there is a question mark in a sentence, and the sentence does not begin with the words *почему/зачем (why, for what),* its sentiment score should be reduced;
2.2. If there is *если (if)* in a clause, the sentiment scores of the words in this fragment that go after *если* should be reduced;
2.3. If there is *ли* particle in a clause, and there is no such words as *чуть/то/вряд/видишь/видите/мало/едва/что* just before *ли,* the sentiment score of the clause should be reduced;
2.4. If there is *бы* particle in a clause then the sentiment score of the words in this clause, which go after *бы,* should be reduced.

Note that it was supposed in (Kuznetsova, 2012) that all the above mentioned rules of the second group, result in nullifying the corresponding fragment sentiment score, but our experiments demonstrated that the reduction of the sentiment score is more efficient. The sentiment score of a fragment containing irrealis is reduced by current algorithms with a certain specified coefficient (in this version 0.4).

Table 4. The results of both groups of rules on the ROMIP-2012 training collection

	Macro_P, %	Macro_R, %	Macro_F1, %	Accuracy, %
Baseline	60.9	61.0	60.9	60.5
Baseline + rules	61.1	61.3	61.2	60.9
Baseline + algo	61.4	61.5	61.5	61.4
Full composition	**61.5**	**61.6**	**61.6**	**61.5**

Table 4 shows the results of the aforementioned rule group implementation in POLYARNIK system. The evaluation metrics used in ROMIP-2012 (Chetviorkin, Loukachevitch, 2013) are applied here. Table 5 shows how the number of correctly and incorrectly classified quotations changes depending on the rule set.

There were 3893 quotations in the training collection, and the scores of 333 of them changed in case of the full set of rules. Therefore we can see that we managed to improve the system performance without any changes in the sentiment dictionaries.

Table 5. The quality of quotation sentiment classification with various rule sets

	Number of quotations changed to the correct class	Number of quotations changed to the incorrect class	Growth of correctly classified quotations compared to the baseline
Baseline	—	—	—
Baseline + rules	20	7	13
Baseline + algo	53	21	32
Full composition	60	22	38

The new version of POLYARNIK system was applied to the test collection of ROMIP-2012 news sentiment classification task for the final evaluation. Table 6 shows the quality metrics of the system with various groups of rules on the test collection. The resulting quality of the full rule set is less than on the training set, but in general we can see performance improvements for all groups of rules.

Table 6. The results of both groups of rules on the ROMIP-2012 test collection

	Macro_P, %	Macro_R, %	Macro_F1, %	Accuracy, %
Baseline	62.6	61.6	62.1	61.60
Baseline + rules	62.8	61.9	62.3	61.90
Baseline + algo	63.0	62.2	62.6	62.25
Full composition	62.9	62.2	62.6	**62.32**

Conclusion

In this paper POLYARNIK sentiment analysis system was presented. The system performance yielded the best results in the ROMIP-2012 news-based sentiment classification task, what in our opinion is due to the extensive system dictionaries, which were created beforehand.

Then without any changes to the sentiment lexicon we implemented the set of rules to take into account groups of opinion words and operators and irrealis markers. Using these new rules, the system performed better both on the train and test collections. In prospect we suppose to continue incorporation of different kinds of rules into POLYARNIK system and testing them on the available quotation collections. Furthermore, we plan to examine rules performance in sentiment analysis in specific domains such as movies, books, etc.

Acknowledgements

This work is partially supported by RFFI grant N11-07-00588-a

References

1. *Ahmad K., Gillam L., Tostevin L.* University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval. In the Proceedings of Eigth Text Retrieval Conference (Trec-8), 1999.
2. *Chetviorkin I., Loukachevich N.* Automatic extraction of domain specific opinion words. Computational Linguistics and Intellectual Technologies. Proc. of International Conference Dialog, 2010, pp. 565–571.
3. *Liu B.* Sentiment analysis and Subjectivity. Handbook of Natural Language Processing, CRC Press, Taylor and Francis Group, Boca Raton, 2010, pp. 1–38.
4. *Chetviorkin I., Braslavski P., Loukachevitch N.* Sentiment analysis track at ROMIP 2011. Computational Linguistics and Intellectual Technologies. Proc. of International Conference Dialog, 2012, pp. 739–746.
5. *Chetviorkin I., Loukachevitch N.* Sentiment analysis track at ROMIP 2012. Computational Linguistics and Intellectual Technologies. Proc. of International Conference Dialog, 2013.
6. *Kuznetsova E. S.* Linguistic support for sentiment analysis of opinionated quotes in Russian. Dipoma thesis of Lomonosov Moscow State University, 2012, Moscow.
7. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2008, v. 2, n. 1–2, pp. 1–135.
8. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment classification using machine learning techniques. In Proceeding of the conference on empirical methods in natural language processing, 2002, Philadelphia, PA, USA, pp. 79–86.
9. *Pazelskaya A., Solovyev A.* method of sentiment analysis in Russian texts. Computational Linguistics and Intellectual Technologies. Proc. of International Conference Dialog, 2010, pp. 510–522.
10. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-based methods for Sentiment Analysis. Computational linguistics, 37(2), 2011.

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ПАРАМЕТРОВ ПРОДУКТОВ ИЗ ТЕКСТОВ ОТЗЫВОВ ПРИ ПОМОЩИ ИНТЕРНЕТ-СТАТИСТИК

Марчук А. А. (aamarchuk@gmail.com)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Уланов А. В. (alexander.ulanov@hp.com)
Научно-исследовательская лаборатория Хьюлетт-Паккард в России

Макеев И. В. (ilya.makeev@gmail.com)
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, Санкт-Петербург, Россия

Чугреев А. А. (artemij.chugreev@gmail.com)
Санкт-Петербургский государственный политехнический университет, Санкт-Петербург, Россия

Ключевые слова: анализ мнений, извлечение информации, параметры продуктов, классификация

EXTRACTING PRODUCT FEATURES FROM REVIEWS WITH THE USE OF INTERNET STATISTICS

Marchuk A. A. (aamarchuk@gmail.com)
St. Petersburg State University, St. Petersburg, Russia

Ulanov A. V. (alexander.ulanov@hp.com)
Hewlett-Packard Labs Russia

Makeev I. V. (ilya.makeev@gmail.com)
St. Petersburg State University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia

Chugreev A. A. (artemij.chugreev@gmail.com)
St. Petersburg Polytechnical University, St. Petersburg, Russia

The paper studies the task of extracting product features from reviews. We consider this task as a classification problem and propose a number of classification features. These features are computed using different statistics returned by queries to Yandex search engine, the Internet library and the Russian National Corpus. To justify our approach, we create and manually label a product features dataset, compute the proposed classification features and conduct classification experiments. The results produced by various classifiers applied to different subsets of the data show the feasibility of our approach. We also look at the usefulness of the proposed classification features.

Keywords: opinion mining, sentiment analysis, information extraction, product features, classification

1. Introduction

A lot of useful information is stored in user-generated content, especially when it contains opinions. These days, users are able to express their opinions and write reviews about almost everything on the Web. Opinion mining or sentiment analysis area of study analyzes such kind of content. Its ultimate goal is to detect opinionated texts and extract who and when expressed which degree of positivity towards which entity or its attribute [12]. Then such tuples can be analyzed computationally. In this work, we are focusing on the problem of entity extraction, or, more specifically, mining product features from reviews.

The task of mining product features can be considered as information extraction task [15], or in particular, relationship extraction problem, when one mines relationships for a given product. Many methods from those fields were adapted to the problem of mining product features. One of the first works [9] dealing with this problem suggests that most frequent nouns and noun phrases in reviews are product features. Infrequent features are extracted by relationships with the same opinion words that accompany frequent features. Paper [2] proposes several useful features to detect noun phrases as product features. Pointwise mutual information (PMI) is computed between a candidate phrase and a product with a relationship discriminator. An example of the latter is “scanner comes with”, where “scanner” is a product, and “comes with” is a discriminator. PMI is also computed between a product and a candidate noun phrase. Statistics for PMI is gathered from a Web search. Other features used in [2] are WordNet’s component/parts relationships. The authors of [7] deal both with explicit and implicit product feature extraction. They perform classification into feature groups as well. Dependency parsing is employed in [17]. Hidden Markov models (HMM) are used and part of speech information is employed in [11]. Conditional random field (CRF) classifier is used in [10]. Token, part of speech, dependency path, word distance and opinion class of a sentence are used as classification attributes there. Another line of work is concerned with the use of topic modeling. Multi-grained topic model is proposed in [16], however, opinion words and product features are not distinguished into separate groups. The authors of [3] construct a localized Latent Dirichlet Allocation (LDA) model that allows them to perform clustering of product aspects and to infer sentiment orientation of them.

Most works on sentiment analysis for Russian are devoted to either sentiment classification or opinion word mining, for example [5], [14] and [4]. The latter approaches the task of opinion lexicon generation as a classification task.

This paper is motivated by development of a service for automatic sentiment analysis [6] and addresses the problem of product features extraction in Russian. We consider this problem as a classification task. We build a labeled dataset of product features extracted from product reviews, propose a number of attributes and use them to perform supervised classification. We use attributes proposed in several other works [2], [1], as well as those motivated by common sense. We also study the usefulness of features and the performance of various classifiers.

2. Classification features

We consider the problem of product feature extraction as a classification task. Candidates are extracted from text and classified into two classes: feature and not feature. Candidates can be words or phrases. The following classification features are employed: frequency, opinion word proximity, weirdness, TF-IDF, PMI. Below are their definitions. Section 3 contains a description of their implementations.

Frequency

Frequency is computed with the following formula:

$$freq_{corpus}\ (c) = \frac{N(c)}{N},$$

where *N(c)* is the number of occurrences of the candidate *c* in the corpus of size *N*.*N(c)* and *N* may be words, phrases or documents. Further, we will compute word frequency and document frequency.

Opinion word proximity

An opinion word lexicon is needed to compute this feature. The trivia is that if there is an opinion word near the candidate, then it is probable that the opinion is expressed about it and it may be a product feature. We compute the number of documents in which the opinion word *ow* is in proximity of *p* words within the candidate *c*.

Weirdness

Weirdness represents the difference in distribution of lexical items in a specialized corpus and in a general one [1]. We need such general corpus, where the product features are weird. Weirdness is computed as follows:

$$weirdness(c) = \frac{freq_{special}\ (c)}{freq_{general}\ (c)},$$

where *special* means a specialized corpus and *general* is a general corpus. In our case, a specialized corpus is a collection of reviews.

TF-IDF

TF-IDF stands for the Term Frequency Inverse Document Frequency. It is a well-known feature that can be computed in a number of ways. In this work, we use the following formulae:

$$TFIDF(c,d) = TF(c,d)IDF(c),$$
$$TF(c,d) = freq_d(c),$$
$$IDF(c) = log\left(\frac{N(d)}{N(d_c)}\right),$$

where d is a document, d_c is a document with the candidate c.

It is important to note that TF-IDF depends nonlinearly on the size of the corpus, unlike the previously mentioned features.

PMI

The Pointwise Mutual Information (PMI) between two lexical items is a measure of the degree of statistical dependence between them and is defined as follows:

$$PMI(c,l) = log\left(\frac{freq(c,l)}{freq(c)freq(l)}\right),$$

where c is a candidate, l is some lexical item (word or phrase). $freq(c,l)$ is a frequency of them occurring together. It may mean, for example, occurrence one by one, in one sentence, in one document etc. We use different types of occurrences in this work.

3. Experiments

3.1. Dataset

We create and label a dataset with product features. We make an assumption that product features are single nouns and they explicitly appear in the text. This means that we consider only a part of the product feature name if it is a multi-word noun phrase. The side-effect is that representation of product features in a single noun may become ambiguous and hard to understand without context. However, the type of the product is known in advance and provides the context for disambiguation. We don't consider implicit product features [12] due their complex nature; however, they occur rarely because people usually use explicit descriptions to mention a product feature.

We extract all nouns from the reviews dataset described in [6]. It consists of 810 laptop reviews crawled from on-line shopping site Citilink[1]. The nouns were extracted and normalized using Mystem[2] part of speech tagger. It resulted in 1,994 unique nouns.

[1] http://www.citilink.ru/

[2] http://company.yandex.ru/technologies/mystem/

Then these nouns are manually labeled by 3 persons with 3 classes: a product feature (PF), not a product feature (NF) and a possible product feature (PPF). We agree to assume that a product feature is a product part, property or an attribute. All related entities and their parts are considered as well. For example, “keyboard”, “thickness” and “soft” are labeled as laptop features; “air”, “consumer”, “moment” are labeled as non-features; “resource”, “brain”, “glue” are labeled as possible product features.

The PPF class is hard to work with because it is very uncertain. It can be interpreted both as PF and NF. Depending on this, there will be different classification results and correlation agreement between the assessors. We will consider 3 solutions to this problem: remove all PPF, use them as PF and use them as NF.

The difficulty of product feature classification emerges already during the manual labeling process. Uncertainty and the lack of a formal feature definition result in low agreement between the assessors. The values of pair wise correlations between the assessors are 39%, 42% and 61% respectively. Considering PPF as PF produces even worse results: 32%, 32% and 45%. Considering PPF as NF gives correlations similar to the initial ones. If PPF is removed, then correlations are 61%, 62% and 97%. Such dispersion in agreements once again proves the difficulty of the work with product features.

Nine datasets for classification experiments are created from the mentioned labeled dataset. Three different approaches to treat the assessors’ agreement are used: intersection of labels, voting and the author’s labels. PPF label is assigned if there are 3 different votes. The three mentioned ways are applied to treat PPF label. Additional datasets are constructed for extra experiments.

Data imbalance is dealt both with oversampling the minority class and undersampling the majority class. Oversampling is performed in two rounds with a synthetic minority over-sampling technique (SMOTE) [13]. Each round doubles the minority class data. Then the order of instances is randomized. Undersampling is performed by means of removing instances of the majority class in order to make it the same size as the minority class.

3.2. Computation of classification features

We relied on the Yandex[3] search index as on the corpus for computing statistics, because it is supposed to be the biggest and all-embracing and, thus, the most precise from the freely available. The service YandexXML[4] provides a query API to the search engine. It has a limitation of the number of queries per day. The query result contains various fields, out of which we are interested in “found-docs”. It means an approximate number of documents relevant to the query. A simple software for making such queries has been written.

The mentioned approach has a number of restrictions. One cannot accurately argue, what is considered as a document, what percentage of document text is indexed,

[3] http://www.yandex.ru/

[4] http://xml.yandex.ru/

how the relevancy is computed, how precise the approximate number of documents is, etc. The search index is constantly changing and this puts certain restrictions on repeatability of our experiments. Another important issue is that it is impossible to compute pure statistics because the size of the index is unknown. As we mentioned earlier, there are some features that depend linearly (or on a constant) on the size of a corpus. In this case, we can deal with the unknown size of the corpus by means of normalization. However, TF-IDF depends non-linearly and we have to compute pseudo TD-IDF.

Let us consider the practical aspects of feature computation.

Frequency

We decide to use two different frequencies. The first is computed by means of Yandex Market[5] and represents a review corpus. The second is computed by means of Yandex and represents the whole Internet. We use the number of relevant documents returned for the queries "candidate host:market.yandex.ru" and "candidate". As mentioned earlier, there is no need to know the size of the Yandex Market and Internet corpora to compute frequencies.

Opinion word proximity

Yandex has quite a few query parameters that allow creating rather complex queries. One can search for the keywords occurrence in the same sentence and for the keywords occurring together not farther than a given number of words. We use two opinion words, "bad" and "good". Opinion word proximity is computed as the number of documents returned by the query "candidate /3 (good | bad)". This means that "good" or "bad" must be no farther than 3 words from the phrase "candidate". We will refer to it as to "OpinionNEAR3".

Weirdness

We employ two general purpose corpora: the Internet library lib.rus.ec[6] (LIB) and the Russian National Corpus[7] (RNC). LIB contains predominantly fiction and its size is 257,000 books. From RNC, the newspaper corpus is used that contains 332,720 documents (173,521,766 words). These corpora have been chosen because they are able to provide reasonable weirdness for the laptop product features. Weirdness-LIB is computed using the number of documents returned by the "candidate host:market.yandex.ru" and "candidate" queries. The software for querying RNC has been written. It returns the number of keyword occurrences and the number of documents with a keyword. Weirdness-RNC is computed using the mentioned numbers and the number of documents returned by the query "candidate host:market.yandex.ru". Frequencies from both general corpora are included as classification features as well. Interestingly, RNC provides a number of different sub-corpora and returns precise statistics. This is an area for further investigation.

---

[5] http://market.yandex.ru/

[6] http://lib.rus.ec/

[7] http://www.ruscorpora.ru/en/index.html

TF-IDF

The TF part of TF-IDF is computed as the number of documents returned by the query "candidate host:market.yandex.ru". The IDF part is computed using general corpora, as proposed in [4]. IDF-LIB cannot be computed precisely because the total number of documents is unknown. We use the number of books instead of it. IDF-RNC can be computed precisely because all the needed statistics is returned by RNC. The number of documents returned by RNC is used as a separate feature. We will refer to TF-IDF computed with LIB as to "TF-IDF-LIB" and to the one computed with RNC as to "TF-IDF-RNC".

PMI

We compute PMI with respect to the word "laptop" and a candidate. We try two different approaches to estimate . We use the number of documents returned by "candidate && laptop", that means search for both keywords in the same sentences. The second approach is to use the number of documents returned by "candidate & laptop", that means search for both keywords in the same documents. We perform search in Yandex and Yandex Market. Eventually, we have four versions of and 4 PMI consequently: "PMI-snt", "PMI-doc", "PMI-YM-snt", and "PMI-YM-doc".

We add the value 0.5 to the document count if it is used in logarithm or as a denominator. Finally, we have 23 different features including the assessors' labels.

3.3. Product feature classification

We use Weka[8] data mining tool [8] to conduct classification experiments. We chose 3 different classifiers: logistic regression, a decision tree and support vector machines (SVM). "J48" implementation of C 4.5 decision tree and "SMO" implementation of SVM is used. Logistic regression and the decision tree are run with default parameters. SVM is used with "data standardization", "build logistic models" and parameters. Two kernel types are set: radial basis (RBF) and normalized polynomial.

As we mentioned earlier, we prepared nine datasets. Table 2 reports classification results for 3 out of 9 prepared datasets and 1 additional one. These are the datasets created with the use of voting and with 3 different approaches to treat the possible product feature (PPF) class. "Vote-strong" dataset does not contain any converted PPF instances. All PPF labels were converted to non-product features (NF) in the "Vote-negative" and to product features (PF) in the "Vote-positive". "Vote-negativeO" is an oversampled "Vote-negative". The properties of these datasets are listed in Table 1. We conduct experiments with all remaining 6 datasets as well. They behave similarly to the "Vote-strong" classification and thus we didn't put them into the resulting table. The experiments were performed with 10-fold cross validation. The results in the table are the averages. Confidence interval for the F1-measure is similar for all experiments and is no more than 0.02 (alpha is 0.01). SVM column contains the best result of two kernels.

[8] http://www.cs.waikato.ac.nz/ml/weka/

Table 1. Properties of selected datasets

	PF	NF	total
Vote-strong	367	837	1204
Vote-negative	367	1627	1994
Vote-positive	1157	837	1994
Vote-negative^{O}	1468	1627	3095

Table 2. Aspect classification results

	Decision Tree			SVM			Logistic Regression		
Dataset	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Vote-strong	0.757	**0.711**	**0.733**	**0.801**	0.624	0.700	0.785	0.619	0.692
Vote-negative	0.509	**0.316**	0.390	**0.679**	0.294	**0.411**	0.609	0.259	0.363
Vote-positive	**0.790**	0.728	0.758	0.702	0.828	**0.760**	0.688	**0.831**	0.753
Vote-negative^{O}	**0.819**	**0.841**	**0.830**	0.816	0.766	0.790	0.785	0.727	0.755

Classification performance is quite good for the first dataset because it doesn't contain possible product features about which the assessors were not sure. The second dataset is imbalanced and the results are unsurprisingly mediocre. Interestingly, "Vote-positive" shows good performance despite the low agreement between the assessors. One of the reasons for this is that the real amount of single noun product features in our dataset may be comparable to the real amount of "neutral" or non-product feature nouns. The reason why the assessors did not agree on this was ambiguity of the nouns. Classification of the oversampled "Vote-negative" dataset provides the best results. We also conduct experiments with the undersampled "Vote-negative" and it performs very similarly to the first one, which is reasonable.

Different classifiers perform more or less as expected. SVM wins on the hardest imbalanced data, however due to some parameter tuning. The decision tree performs well on everything except the mentioned imbalanced data. In general, the classification results show applicability of the proposed approach to the product feature extraction. They also show that the possible product feature class can be considered both as a feature and as a non-feature. It may depend on the user's requirement: show more uncertain features or only precise ones.

Interestingly, our results are comparable to the results reported in papers on product features extraction for English [2], [9], [10], and [12]. They report an average F1-measure ranging from 0.76 to 0.86.

We are also interested to find out, which classification features are the most useful. We conduct experiments with each feature separately, but some of them produced zeros. We decide to combine at least two features instead. PMI is chosen as a default feature because it was used as a base feature in a similar work for English [2]. We have 2 modifications of PMI: "PMI-snt" and "PMI-doc". Experiments with a pairwise combination of different features with them are performed. SVM classifier is used with the same settings as mentioned previously and RBF kernel. The results are represented in Table 3.

Table 3. Aspect classification results with different features

	PMI-snt			PMI-doc		
Feature	**P**	**R**	**F1**	**P**	**R**	**F1**
TF-IDF-LIB	0.643	0.172	0.271	0.610	**0.226**	**0.330**
Weirdness-LIB	0.778	0.038	0.073	**0.789**	0.041	0.078
Weirdness-RNC	0.321	0.025	0.046	0.529	0.025	0.047
TF-IDF-RNC	0.344	0.030	0.055	0.481	0.071	0.124
PMI-YM-docs	0.383	0.049	0.087	0.154	0.005	0.011
PMI-YM	0.242	0.022	0.040	0.278	0.014	0.026
OpinionNEAR3	0.231	0.016	0.031	0.512	0.060	0.107

One can see that the “TF-IDF-LIB” and “Weirdness-LIB” are the most useful features in combination with PMI. Interestingly, TF-IDF and Weirdness computed with a different general corpus provide worse results. It is accounted for by the use of the newspaper corpus from RNC, while the corpus in LIB is mostly fiction. Newspapers are more probable to have product features, rather than fiction. Another interesting observation is that PMI computed using “the same document” (“PMI-doc”) query perform slightly better than the one computed with “the same sentence” query (“PMI-snt”).

4. Conclusion

We performed the task of product features extraction from Russian reviews. It was addressed as a classification problem. A product feature dataset was created and labeled. A number of different classification features were used and several classification algorithms applied. The experiments demonstrated efficiency of our approach.

Our further work is to use additional linguistic and statistical attributes for classification. Spelling corrector will be employed to correct the spelling of candidates. We plan to apply sequence labeling classifiers as well. We will do product features clustering to group them into meaningful groups. This may help us to filter features as well.

References

1. *Ahmad, K.; Gillam, L.; and Tostevin, L.* 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In The Eighth Text Retrieval Conference (TREC-8).
2. *Ana M. Popescu, Oren Etzioni.* Extracting Product Features and Opinions from Reviews. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (2005)
3. *Brody, S. and S. Elhadad.* An Unsupervised Aspect-Sentiment Model for Online Reviews. In Proceedings of The 2010 Annual Conference of the North American Chapter of the ACL, 2010.

4. *Chatviorkin Ilya, Lukashevich Natalia*. Automatic Extraction of Domain-Specific Opinion Words. Proceedings of the International Conference Dialog, 2010.
5. *Chetviorkin I. I., Braslavski P. I.* Sentiment analysis track at ROMIP 2011. Dialog 2011.
6. *Chugreev A., Marchuk A., Makeev I., Mokaev T., Skudarnov Y., Bat'kovich D., Ulanov A.* GoodsReview — A Service for Automatic Sentiment Analysis [GoodsReview — servis avtomaticheskogo analiza portrebitel'skogo mneniya]. Proceedings of KESW, 2012.
7. *Ghani, R., K. Probst, Y. Liu, M. Krema, and A. Fano.* Text mining for product attribute extraction. ACM SIGKDD Explorations Newsletter, 2006, 8(1): p. 41-48.
8. *Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H.* (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
9. *Hu, M. and B. Liu*. Mining and summarizing customer reviews. In Proceedings of ACM SIGKD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 2004.
10. *Jakob, N., & Gurevych, I.* (2010, October). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 1035–1045). Association for Computational Linguistics.
11. *Jin, W. and H. Ho.* A novel lexicalized HMM-based learning framework for web opinion mining. In Proceedings of International Conference on Machine Learning (ICML-2009), 2009.
12. *Liu, Bing.* Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. 2nd ed. 2011, XX, 622 p.
13. *Nitesh V. Chawla* et. al. (2002). Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16:321–357.
14. *Pak A., Paroubek P.* Language independent approach to sentiment analysis (LIMSI Participation in ROMIP'11), Dialog 2011.
15. *Riloff E*. Automatically constructing a dictionary for information extraction tasks. Proceedings of the National Conference on Artificial Intelligence, 811–811, 1993
16. *Titov, I. and R. McDonald.* Modeling online reviews with multi-grain topic models. In Proceedings of International Conference on World Wide Web (WWW-2008), 2008.
17. *Wu, Y., Q. Zhang, X. Huang, and L. Wu*. Phrase dependency parsing for opinion mining. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2009), 2009.
18. *Zhang Z., Iria J., Brewster C., Ciravegna F. A.* Comparative Evaluation of Term Recognition Algorithms. In the sixth international conference on Language Resources and Evaluation, (LREC 2008).

USING BASIC SYNTACTIC RELATIONS FOR SENTIMENT ANALYSIS

Mavljutov R. R. (m-ceros@yandex.ru),
Ostapuk N. A. (nataxane@yandex.ru)

Yandex, Moscow, Russia

The paper describes a rule-based approach to sentiment analysis. The developed algorithm aims at classifying texts into two classes: positive or negative. We distinguish two types of sentiments: abstract sentiments, which are relevant to the whole text, and sentiments referring to some particular object in the text. As opposed to many other rule-based systems, we do not regard the text as a bag of words. We strongly believe that such classical method of text processing as syntactic analysis can considerably enhance sentiment analysis performance. Accordingly, we first parse the text and then take into account only the phrases that are syntactically connected to relevant objects. We use the dictionary to determine whether such a phrase is positive or negative and assign it a weight according to the importance of the object it is connected with. Than we calculate all these weights and some other factors and decide whether the whole text is positive or negative. The algorithm showed competitive results at ROMIP track 2012.

Keywords: sentiment analysis, opinion mining, syntactic relations, context-free grammar, thesaurus

1. Introduction

Automatic sentiment analysis is a comparatively new field in computational linguistics. With developing of Web and particularly blogosphere every Internet user got the opportunity to leave a review, expressing his or her opinion about some product or service. Such information is useful both for other users and for market departments of service providers. The problem is this information is large, so it cannot be processed manualy. As an illustration, the website TripAdviser.com publishes about 40 reviews every minute, and booking.com has almost 18 million reviews overall. The methods of natural language processing may be helpful to takle the issue with big amount of data. On the basis of these methods systems of sentiment analysis are being developed. The goals of the SA systems vary from text tone assessment to extraction and assessment of specific parameters, which are discussed in the text.

Automatic sentiment analysis task encounters a lot of problem, such as implicit expression of emotional component in the text, too informal language of reviews and until recently lack of annotated corpus for Russian to measure the quality. To settle the last problem, ROMIP offers sentiment ananlysis track, which aims at classifying blog posts about books, films and cameras according to the sentiment they express into 2, 3 or 5 groups.

The current version of our system classifies reviews into two groups. The algorithm is based on rules, which take into account syntactic relations in the text. The main goal of our participation in ROMIP 2012 was to measure the quality of work of our system and to compare it with others in order to understand if we are on the right way and at what else we have to work.

2. Related works

All existing approaches to sentiment analysis can be divided into two large categories: rule-based and machine learning based.

Sentiment analysis based on machine learning in general is similar to classical task of text classification, where sentiment words act as features. The commonly used method here is support vector machines trained on large annotated corpora [3], [5], [8].

Rule-based methods make use of sentiment lexicon of the text. Such methods vary from simple lists of positive and negative words to more sophisticated methods, using sentiment patterns and syntactic relations between words in the text. Approaches which involve syntactic relations are mostly developed for English language [11], [14]. For the Russian language the task of constructing syntactic tree is much more complicated, taking into account rich morphology and free word order.

In [7] the syntactic approach to sentiment analysis for Russian was implemented. This system aimed at determining news texts tone. It extracts the object of evaluation as well as syntactic groups with opinion words and according to some set of rule combines them.

3. Method description

In our work we implemented the following algorithm: first, we gathered object thesaurus, including terms to which opinion phrase could refer. Then we detected phrases syntactically connected to objects from the thesaurus, as well as negations relevant to these phrases — such syntactic groups became potential entries of our sentiment dictionary. It's worth noticing, that we considered the whole syntactic group including the object as a sentiment; not just opinion phrase: this issue will be considered in details in Section 3.3. After that we compiled a sentiment dictionary using mined syntactic groups and some additional resources and finally we searched for sentiments in the text and weighed them to determine text tone.

3.1. Objects thesaurus

For each class of objects (films, books, digital cameras) we have gathered a thesaurus, that has three categories of terms:

1. Common nouns that denote objects of the class. For digital cameras such terms are *“камера”, “фототехника”, “аппарат”, “фотоаппарат”,* etc.

2. Proper names of objects of the class. The names of the camera models and movies and books titles.
3. Common nouns that denote parameters, properties, and parts of objects of the class. As an illustration, for digital cameras the parameters are *"формфактор", "качество фото", "разрешение", "матрица", "объетив", "вспышка"* etc. For films and books they are *"автор", "режиссер", "игра актеров", "атмосфера", "дубляж"* etc.

Each element of the thesaurus had its unique id, class id and type id.

The distribution of terms quantity by object class and type was the following:

	books	**films**	**digital cameras**
common nouns	69	74	475
proper names	2,713	208	1,412
parameters	161	252	512

We have filtered ambiguous proper names (e.g., *"камень"*), to be sure that we wouldn›t mix up class objects with other entities in texts. For the digital cameras we have also made a vocabulary of contracted proper names that consists of company names and parts of the full names of the models. This vocabulary is helpful since the camera names are usually complex, so writers (especially in blogs and comments) prefer to use simplified versions. For example, instead of *"BenQ DC C1450"* they may write *"BenQ DC"*, *"BenQ", "benq", "benq dc", "c1450"*, and so on.

Gathering data for thesaurus

At the beginning of ROMIP competition we were given a vocabulary of proper names for each class as a source data. We used this vocabulary to mine common nouns and parameters. To perform this task we executed the following algorithm:

1. Gather text snippets where proper names from the thesaurus are mentioned. Each text got a class id according to the class of the proper name that was found in it. We used a part of Russian Web as a source, and we restricted the search area with texts enclosed by the paragraph tag <p>.
2. Extract all noun phrases (which do not coincide with the matched proper name), and sequences of noun phrases connected by genitive case. Let's call them potential thesaurus terms.
3. Calculate Pointwise Mutual Information between a potential term and text class, where it was found:

$$PMI\,(potential\ term,\ text\ class) = \log_2\left(\frac{p(term,\ text\ class)}{p(term) \times p(text\ class)}\right)$$

where p is probability.

The idea was that common nouns and parameters that denoted objects of a certain class would have the value of PMI for this class much bigger, than for the other two classes. So, we could choose the closest class for each potential term and calculate its affinity to the class:

$$affinity\ (term\_i,\ class\_j) = MIN\ (PMI(term\_i,\ class\_j) - PMI(term\_i,\ class\_k))$$

where $k\ != j$ is probability.

Now for each class we have a set of potential terms, and for each potential term we have the value of its affinity to the class.

4. Sort potential terms for each class by the value of their affinity and filter manually the part of them with highest values.

In our case on the first stage we have gathered 2 billion of text snippets in which proper names from the thesaurus were mentioned. On the second stage we got 60 thousand of potential thesaurus terms. We cut off a part of them with low value of affinity, and only 17 thousand were left. After correction of misprints 8 thousands were left. Then we have filtered those that left manually, and only 1.5 thousand terms became a part of the thesaurus.

3.2. Syntactic relations used for opinion extraction

Unlike to other approaches that use syntax, we didn't make full text parsing. According to our experience, there is a set of the specific syntactic relations are generally used to express subjectivity.

Previously we conducted a research which aimed at determining how subjective evaluation of an object could be expressed in the text. The training set consisting of 10 thousand hotel reviews was annotated manually. According to this markup the following distribution was received:

1. 80% of subjective evaluations are grammatical modifiers expressed by adjectives, e.g. *"громкая музыка", "плохое обслуживание"*.
2. 7% — predicates expressed in different ways: *"бармен кричал", "обслуживание было плохим", "обслуживание оставляло желать лучшего", "отель чудовищен", "отношение к клиентам просто ужас"*.
3. 4% — adverbials expressed by adverbs and prepositional phrases connected to predicate, grammatical modifier or directly to an object: *"кран работал плохо", "плохо работающий кран", "связь на троечку"*.
4. 9% — other ways. This ways include expression of subjectivity with interjections (*"брр", "фууу",* etc.), objects comparison (*"А лучше Б", "А понравился меньше, чем Б"*), reference to self (*"мне стало плохо", "я замучался его смотреть"*), and expressions, where object and opinion phrase are not connected syntactically (*"Вчера посмотрел этот фильм. До сих пор противно"*).

We didn't make a detailed study for classes proposed by the ROMIP task and texts related to blogs; however, we made an assumption that the trend would remain the same. In current research we concentrated on the first three ways of subjectivity expression. We also considered independently cases where opinions were expressed by reference to self.

We have used the Tomita-parser[12] for extracting syntactic relations between object and other parts of a sentence. The Tomita-parser is an instrument for extracting structured data (facts) from texts in natural language by means of context-free grammars. To extract a fact, we should write a set of rules, describing the structure of this fact in the text. For example, to extract an adjective agreed with a noun, we should write the next rule:

$$S \rightarrow Adj\text{<}gnc\text{-}agr[1]\text{>}\ Noun\text{<}gnc\text{-}agr[1],\ rt\text{>};$$

For our task we have written set of rules for each of three syntactic structures. In sum we got about 50 rules. The main difficulty was to describe predicates and adverbials, expressed by collocation (*оставляло желать лучшего, на троечку* etc). We searched for such collocations in the text and tried to generalize them and to describe their structure. Of cause, we could not find all of them, and that's why the grammar did not cover all desired syntactic structures — empirically, we managed to detect about 80–90% of them.

Text chunks, which were find by the grammar, were converted into facts. In Tomita, fact is a structured entity, which consists of fields. To convert text chunk into fact means to point out, with which part of the chunk we should fill every fact field. In our case facts consisted of four fields:

1. *an object from the thesaurus*
2. *type of syntactic relation between the object and the other part of the sentence*
3. *related part of the sentence*
4. *negation*

For example, the initial phrase is *"Неделю назад я купил водонепроницаемую камеру от Nikon."* In this sentence the object is *"камера"*. From all syntactic connections of the object, only one may potentially express subjectivity (the grammatical modifier), so one fact will be extracted:

1. *object: "камера"*
2. *relation: grammatical modifier*
3. *related part: "водонепроницаемый"*
4. *negation: false*

Negation extraction

Determining negations is an important part of sentiments extraction. We define negation as a part of text structure that inverts the sign of a sentiment.

In Russian negation is expressed in different ways for different parts of speech. So for each type of syntactic relations in facts we wrote a different set of rules for extraction of negations.

Examples:

(1) *'нет'* | *'без'* | *'отсутствие'* | *'лишенный'* | *'лишивший'* | *'мало'* | *'никакой'* | *'ни'* + *noun in genitive case*

(2) *‘не’ | ‘мало’ + verb in a finite form*

(3) *‘нельзя› | ‹невозможно› + verb in an infinite form*

(4) *‹не› | ‹мало› | ‹ничего› + adjective*

(5) *‘не’ + adverb, preposition phrase*

The presence of “не” (particle of negation) doesn’t necessarily express negation. For example, the expression *“не только мерзкий”* doesn’t change the sign of *“мерзкий”*. Therefore, we have also described the class of expressions, where negation words didn’t express negation.

3.3. Sentiment dictionary

As opposed to usual practice, we don’t consider opinion words apart from their context. An entry in our sentiment dictionary is a fact, not a separate word. This approach is justified by the fact that a sentiment sign depends not only on an opinion word, but also on the object, and type of the syntactic relation that characterize their connection.

Compare two facts with the same opinion word, but different objects:

1) object: “официант”	1) object: “скорость обработки сигнала”
2) relation: grammatical modifier	2) relation: grammatical modifier
3) related part: “бешеный”	3) related part: “бешеный”
4) negation: false	4) presence of negation: false

In the first case the fact describes a negative sentiment; in the second — a positive sentiment; however, the opinion word “бешеный” stays the same.

Also, some sentiments don’t base on opinions words. For example, let’s consider phrase “Брюс уже не тот”. The fields of the fact are:

1. *object: “Брюс”*
2. *relation: predicate*
3. *related part: “мом”*
4. *negation: true*

This fact denotes a sentiment; but, the word “тот” cannot be classified as an opinion word.

The task of compiling the sentiment dictionary was to collect facts, that express a subjective evaluation.

In addition to facts with all fields filled, we also considered their modifications, where values of some fields were empty. It could be a fact with empty “object” or “related part of sentence” field.

A fact with empty “object” field denotes context-free sentiment (the sign of which doesn’t depend on object). For example, the phrase *“что-то было ужасным”* represents a negative attitude regardless of the object.

A fact with empty "related part of sentence" field denote object, which convey a subjective evaluation by itself. For example, the parameters of digital cameras, like *"блики экрана", "поломка", "царапина", "битый пиксель"*, convey a negative attitude.

Compiling the sentiment dictionary

We used several sources to compile our dictionary:

1. Object-independent sentiments, which we gathered at the previous stage of our research.
2. Filtered manually and translated to our format vocabulary of sentiments given for the competition. Again, we used only object-independent sentiments.
3. The training set. The algorithm was very similar to that we used for thesaurus mining. In this case, the classes were negative and positive reviews. For each fact we have calculated its PMI with each of two classes. Then for each class we made a list of facts with the highest values of affinity to it. These facts formed the sentiment dictionary.

The size of the final vocabulary was 43 thousands of facts. Among them 5.5 thousands of facts were with empty field "object" (object-independent sentiments).

3.4. Two class classification of blog texts

After the Tomita-parser extracted facts from a text, we searched for these facts in the sentiment dictionary. Those sentiments which were found became features for the review classification.

The class of the texts was defined by the sign of the weighed sum:

$$predicted\ class = SUM\big(object\_i\_weight \times relations\_in\_sentiment\_i\_weight \times sentiment\_i\_class\big) - TRESHOLD,\ sum\ of\ all\ found\ sentiments$$

We have made the following assumptions:

1. the weight of the object expressed by a proper name or by a common noun is 1. The weight of the object parameter is 0.5
2. if the text has more than two mentions of different proper names, we consider this text as not a review, and refuse to classify it.

Thereby, the weighed sum has 4 variables to define: 3 weights for different types of relations in sentiments (modifier, predicate and adverbial) and the TRESHOLD parameter.

We used the training set to find optimal values for the parameters. As an algorithm for learning we chose SVM with cross-validation. The best results on the training set were precision 0.94, recall 0.89 for the positive class.

4. Results and further work

Here are official results from ROMIP 2012 for 2-class sentiment classification track. Our results are highlighted with blue color:

System_ID	Precision_P	Recall_P	F_Measure_P	Precision_N	Recall_N	F_Measure_N	Accuracy
Object — book							
xxx-17	0.914530	0.955357	0.934498	0.583333	0.411765	0.482759	0.883721
xxx-8	0.868217	1.000000	0.929461	0.000000	0.000000	0.000000	0.868217
xxx-27	0.873016	0.982143	0.924370	0.333333	0.058824	0.100000	0.860465
xxx-10	0.898305	0.946429	0.921739	0.454545	0.294118	0.357143	0.860465
xxx-41	0.872000	0.973214	0.919831	0.250000	0.058824	0.095238	0.852713
xxx-39	0.866142	0.982143	0.920502	0.000000	0.000000	0.000000	0.852713
xxx-3	0.910714	0.910714	0.910713	0.411765	0.411765	0.411765	0.844961
xxx-25	0.901786	0.901786	0.901786	0.352941	0.352941	0.352941	0.829457
Object — film							
xxx-23	0.857534	0.948485	0.900719	0.604651	0.333333	0.429752	0.830882
xxx-12	0.836788	0.978788	0.902235	0.681818	0.192308	0.300000	0.828431
xxx-18	0.823980	0.978788	0.894737	0.562500	0.115385	0.191489	0.813725
xxx-15	0.854749	0.927273	0.889535	0.520000	0.333333	0.406250	0.813725
xxx-14	0.817043	0.987879	0.894376	0.555556	0.064103	0.114943	0.811275
xxx-17	0.808824	1.000000	0.894309	0.000000	0.000000	0.000000	0.808824
xxx-13	0.860000	0.912121	0.885294	0.500000	0.371795	0.426471	0.808824
xxx-19	0.895899	0.860606	0.877898	0.494505	0.576923	0.532544	0.806373
Object — camera							
xxx-5	0.965937	1.000000	0.982673	0.000000	0.000000	0.000000	0.965937
xxx-13	0.975062	0.984887	0.979950	0.400000	0.285714	0.333333	0.961071
xxx-15	0.970297	0.987406	0.978777	0.285714	0.142857	0.190476	0.958637
xxx-14	0.965602	0.989924	0.977612	0.000000	0.000000	0.000000	0.956204
xxx-20	0.972431	0.977330	0.974874	0.250000	0.214286	0.230769	0.951338
xxx-2	0.977099	0.967254	0.972152	0.277778	0.357143	0.312500	0.946472
xxx-10	0.977041	0.964736	0.970849	0.263158	0.357143	0.303030	0.944039
xxx-17	0.972010	0.962217	0.967089	0.166667	0.214286	0.187500	0.936740

Precision, recall and F-measure were counted separately for positive and negative texts. Accuracy is proportion of correctly classified objects in all objects processed by the algorithm it is calculated according the following formula:

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

where tp is correct results, fp — unexpected results, fn — missing results and tn — correct absence of results. [2]

Our classifier has the second result (among 26 participants) in film classification and the third — in book classification (among 40 participants). A little bit worse we performed at camera classification — we are the sixth of 25. It can be explained by the fact that reviews about books and films are very much alike both in sentiment lexicon and parameters which are evaluated. Camera reviews have more specific lexicon and it was more complicated to extract sentiment facts from them. In such cases training process should be more domain-specific with less "object-independent" sentiments.

From complete result table one can see that regardless to object class precision and recall of classification of negative reviews is considerably lower than positive ones. The explanation is that negative reviews form only 10% of the flow. This correlation is true both for training set and for the Web in general. Prevalence of one class impacts on machine learning. Moreover, it complicates the process of gathering sentiment dictionary for negative class.

Despite pretty bad performance in negative reviews classification, total accuracy is still high enough. It means that test set also contained less negative reviews.

On the basis of existing system we are going to implement 3 or 5 groups classificator. Moreover, at the previous stage of our research we tried to evaluate not the whole text, but separate parameters of it, such as service, beach, rooms for hotel reviews or service, interior, food for restaurant reviews. We believe, that for such objects as hotels and restaurants, as well as cameras, cars and so on, such parametric evaluation is much useful, and that's why we are going to continue our investigation in this area.

References

1. *Chetviorkin I. I.* (2012), Testing the sentiment classification approach in various domains — ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 747–755.
2. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012) Sentiment analysis track at ROMIP 2011 Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 739–746.
3. *Kotelnikov E. V., Klekovkina M. V.* (2012) Sentiment analysis of texts based on machine learning methods [avtomaticheskij analiz tonal'nosti tekstov na osnove metodov machinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 756–763.
4. *Nakagawa T., Inui K., and Kurohashi S.* (2010), Dependency tree-based sentiment classification using crfs with hidden variables, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Morristown, NJ, USA, pp. 786–794

5. *Pak A., Paroubek P.* (2012) Language independent approach to sentiment analysis (LIMSI participation in ROMIP '11), Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 764–771.
6. *Pang B. & Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, v.2 n.1–2, pp.1–135.
7. *Pazel'skaja A. G., Solov'jev A. N.* (2011) A method of sentiment analysis in Russian texts [metod opredelenija emocij v russkih tekstah], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2011"], Bekasovo, pp. 510–522
8. *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.* (2012), Research on applicability of thematic classification methods to the problem of book review classification [issledovanie primenimosti metodov tematicheskoj klassifikacii v zadache klassifikacii otzyvov o knigah], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 772–779.
9. *Poroshin V.* (2012), Proof of concept statistical sentiment classification at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 780–788.
10. *Prabowo R. and Thelwall M.* (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2) pp. 143–157.
11. *Shilpa Arora, Elijah Mayfield, Carolyn Penstein-Rose and Eric Nyberg* (2010), Sentiment classification using automatically extracted subgraph features. NAACL workshop on Computational approaches to analysis and generation of emotion in text
12. *Tomita-parser:* http://api.yandex.ru/tomita/
13. *Vasilyev V. G., Khudyakova M. B., Davydov S.* (2012), Sentiment classification by fragment rules [klassifikacija otzyvov pol'zovatelej s ispol'zovaniem fragmentnyh pravil], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 789–796.
14. *Yi J., Nasukawa T., Niblack W. & Bunescu R.* (2003), Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques, In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003), Florida, USA, November 19–22, pp. 427–434.
15. *Zirn Cacilia, Niepert Mathias, Stuckenschmidt Heiner, Strube Michael.* (2011), Fine-Grained Sentiment Analysis with Structural Features. In Proceedings of 5th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand

СИСТЕМА СЕНТИМЕНТНОГО АНАЛИЗА ATEX, ОСНОВАННАЯ НА ПРАВИЛАХ, ПРИ ОБРАБОТКЕ ТЕКСТОВ РАЗЛИЧНЫХ ТЕМАТИК

Паничева П. В. (ppolin86@gmail.com)

EPAM Systems, Санкт-Петербург, Россия

Ключевые слова: сентиментный анализ, анализ тональности, РОМИП

ATEX: A RULE-BASED SENTIMENT ANALYSIS SYSTEM PROCESSING TEXTS IN VARIOUS TOPICS

Panicheva P. V. (ppolin86@gmail.com)

EPAM Systems, Saint-Petersburg, Russia

ATEX is a rule-based sentiment analysis system for texts in the Russian language. It includes full morpho-syntactic analysis of Russian text, and highly elaborated linguistic rules, yielding fine-grained sentiment scores. ATEX is participating in a variety of sentiment analysis tracks at ROMIP 2012. The system was tuned to process news texts in politics and economy. The performance of the system is evaluated in different topics: blogs on movies, books and cameras; news. No additional training is performed: ATEX is tested as a universal 'ready-to-use' system for sentiment analysis of texts in different topics and different classification settings. The system is compared to a number of sentiment analysis algorithms, including statistical ones trained with datasets in respective topics. Overall system performance is very high, which indicates high usability of the system to different topics with no actual training. According to expectations, the results are especially good in the 'native' political and economic news topic, and in the movie blog topic, proving both to share common ways of expressing sentiment. With regard to blog texts, the system demonstrated the best performance in two-class classification tasks, which is a result of the specific algorithm design paying more attention to sentiment polarity than to sentiment/neutral classes. Along these lines areas of future work are suggested, including incorporation of a statistical training algorithm.

Keywords: rule-based sentiment analysis, sentiment classification, Russian language processing, ROMIP

1. Введение

Сентиментный анализ, или анализ тональности — молодой, но быстро развивающийся раздел автоматической обработки текстов. В середине 1990-х гг. исследователи начали проявлять интерес к выражению субъективного отношения автора в тексте [Wiebe], включая в это понятие мнения, настроения, отношение автора, выраженные каким-то образом в тексте[Pang].

С развитием интернета сентиментный анализ привлекает внимание исследователей как один из разделов анализа субъективности, задачей которого является определение значения «тональности» текста, а именно, классификация текста как отражающего позитивное, негативное или нейтральное отношение автора к объектам, явлениям, персонам, упомянутым в тексте.

Важно отметить, что до сих пор не сформулированы четкие теоретические критерии, по которым тот или иной отрезок текста может быть отнесен к позитивному, негативному или нейтральному классам, несмотря на успешные попытки некоторых исследователей теоретически обосновать сентиментный анализ (к примеру, [Balahur]). Таким образом, оценка значения тональности устанавливается опытным путем, с помощью разметки асессорами, которая затем используется в качестве «золотого стандарта» для обучения и оценки результатов сентиментного анализа. Наличие данных, размеченных таким образом, является критическим для развития этой области, в том числе потому, что бóльшая часть исследований сосредоточена на обучаемых методах классификации.

В России сентиментный анализ стал привлекать внимание исследователей в конце 2000-х гг., что отразилось в появлении в 2011 г. в программе семинара РОМИП дорожек по оценке сентиментного анализа на русском языке. Особенность отечественных работ в данной области заключается в большей производственной и коммерческой направленности описываемых систем. В результате оказываются решающими не только численные показатели результатов работы алгоритмов, обученных и проверенных на определенных текстовых выборках, но и более детальная настройка алгоритмов, прозрачная схема определения значения тональности, основанная на явных и четких лингвистических показателях, а также доступность поддержки системы и ее развития для обработки текстов новых жанров/тематик. С этой точки зрения особенно удобными в применении оказываются системы, основанные на правилах ([Kan, Vasilyev]).

Целью данного исследования является тестирование работы системы ATEX, основанной на правилах, настроенной на новостных текстах различного происхождения, без предварительного обучения. Тестирование призвано показать применимость системы к сентиментному анализу текстов различных тематик в сравнении с другими системами сентиментного анализа, в том числе основанных на машинном обучении. Для этого система ATEX была представлена на семинаре РОМИП в наборе дорожек по сентиментному анализу; при этом не проводилось никакого обучения или дополнительной настройки системы.

2. Алгоритм сентиментного анализа на основе лингвистических правил

Система, которую мы представляем на семинаре РОМИП, автоматически реализует сентиментный анализ на основе правил для русскоязычных данных. Правила содержат богатую линвистическую информацию и применяются к структуре текста, полученной в результате работы морфо-синтаксического модуля системы.

2.1. Морфо-синтаксический анализ

Во-первых, на основе морфологического словаря, содержащего для редактирования в текстовом виде парадигмы слов, происходит определение «нормальной формы» каждой из словоформ в тексте и его грамматических атрибутов. Изначально словарь порожден автоматически по базе данных Грамматического словаря А. А. Зализняка [Zaliznyak].

Для словоформ, которые не были найдены в морфологическом словаре, происходит поиск возможной грамматической информации на основе суффикса и окончания, отбрасывания приставки, а также неточный поиск для потенциальных форм с ошибками и опечатками.

Затем грамматическая информация используется для работы синтаксических правил. Синтаксическая обработка представляет собой формальную грамматику, состоящую из нескольких сотен правил, которые разрешают омонимию, объединяют слова в группы, группы, в свою очередь, в более крупные группы, доходя до размера клаузы и сложного предложения, включая предложения с прямой речью. В полученной многоуровневой структуре происходит простановка синтаксических связей ко всем значимым словам.

2.2. Сентиментный анализ

Сентиментный анализ производится на основе ключевых слов, а также сентиментных правил. И в том, и в другом случае результатом сентиментного анализа является значение тональности (+1, −1, 0 или никакое) для одного или нескольких слов[1].

[1] Текущая версия алгоритма не учитывает силу сентимента, т.е. все слова по умолчанию имеют одинаковый вес при вычислении сентимента предложения. Это упрощение оказывается адекватным и не препятствует достижению высоких результатов подготовке системы на основе наших данных, см. «Подготовка системы к сентиментной классификации»

2.2.1. Ключевые слова

В качестве ключевых сентиментных слов выступают слова, которые несут сентиментную окраску в любом контексте, или в подавляющем большинстве контекстов. Ключевые слова хранятся в виде списков нормальных форм в текстовых файлах и содержат, к примеру, такие слова как «хороший», «плохой», «неприятный», «трус», «успех», «провал», «угроза», «позитив», «оперативно», «своевременно», «слишком», и т.п.; всего 1590 негативных и 510 позитивных слов с указанием части речи, что необходимо для правильной обработки омонимичных форм. Если слово из этого списка с соответствующей частью речи встречается в тексте, ему приписывается соответствующее значение тональности.

2.2.2. Сентиментные правила

Сентиментные правила используются для более точного определения сентимента слов и работают на основе более полной морфо-синтаксической информации. Сентиментные правила реализованы на предметно-ориентированном языке программирования и на входе обрабатывают синтаксическую структуру предложения, состоящую из слов и связей между ними, или ее часть, присваивая значение атрибутам отрицания или сентимента определенным словам на выходе.

2.2.2.1. Сентимент словосочетания

В некоторых случаях отдельные слова не несут в себе сентиментного значения, но сентиментом нагружено определенное сочетание некоторых слов или форм слов. Правило, приписывающее сентимент, основано на синтаксической связи определенных слов или словоформ в предложении.

Например, с помощью этих правил обрабатываются такие сочетания, как «пойти навстречу, на лапу, душа компании, по фазе, так себе, промыть мозг, поставить крест, с ума, из ума, ниже плинтуса», и многие другие.

2.2.2.2. Инверсия сентимента

При отрицании в сочетании со словом, которое содержит значение сентимента, его сентимент инвертируется: если слово имеет позитивную тональность, то отрицание модифицирует его на негатив; при отрицании негативной тональности слово в общем случае получает нулевую тональность сентимента.

Важно отметить, что отрицание может выражаться несколькими способами. Основной способ выражения отрицания — частица «не», предикатив «нет». Они приписывают отрицание словам, с которыми они связаны определенными синтаксическими связями. Также при определенных синтаксических связях отрицание ставится за счет группы слов, отличающихся семантикой отрицания, таких как «отсутствие, удаление, лишение, отрицание, устранение, отсутствовать, удалять, лишать, отрицать, устранять», и предлог «без».

При работе с отрицанием также были выделены группы слов — имен существительных, прилагательных, глаголов, — которые получают или меняют значение тональности специфическим образом в сочетании с отрицанием. Эти слова, во-первых, могут не входить в список ключевых сентиментных слов,

но получают значение сентимента при отрицании; во-вторых, могут содержаться в списке ключевых слов с негативным сентиментом, но при сочетании с отрицанием получают, в отличие от общего правила, позитивный сентимент. В первом случае примером могут служить такие слова, как «будущее, дело, желание, мозг, надежда, объяснение, ответ, смысл, ум»; во втором — «вопрос, дефект, конфликт, нарекание, перебой, препятствие, проблема»[2]. Всего в системе порядка 120 таких слов; они хранятся в виде списков в текстовых файлах, обозначенные как «слова, позитивные с отрицанием» и «слова, негативные с отрицанием».

2.2.2.3. Синтаксически связанные слова, входящие в значимые семантические списки

Категория правил, которая заслуживает особого внимания, — правила, приписывающие сентимент на основе синтаксической связи между словами. При этом каждое из слов по отдельности не входит в ключевые сентиментные слова, а связь двух слов не является устойчивым словосочетанием.

К примеру, такие слова как «деньги, доход, зарплата, качество, оборот, отдача, оценка, потенциал, рейтинг, уровень» не содержат позитива сами по себе. С другой стороны, экспериментально подтверждается, что когда явления, обозначенные этими словами, велики, высоки, максимальны, это добавляет положительную тональность, и наоборот — когда они низки, добавляет отрицательную. Ср. «наш рейтинг»/«высокий рейтинг»/«повышение рейтинга»/«низкий рейтинг»/«понижение рейтинга».

Наоборот, такие слова как «издержка, очередь, потеря, расход, риск, урон, ущерб» будут получать позитивный сентимент, когда такие явления минимальны, и негативный, когда максимальны.

Таким образом, если слова из данных списков синтаксически связаны со словами, обозначающими увеличение или уменьшение степени, количества явления или предмета, то главному слову в данной синтаксической связи приписывается соответствующее значение сентимента. Данные правила также включают в себя списки слов, относящиеся к семантике проблем, ситуаций и решения; нехватки; порядка, правил и их гибкости, жесткости, и т. п.

3. Постановка задачи

3.1. Дорожки РОМИП по сентиментному анализу

На семинаре РОМИП были предоставлены 2 вида дорожек по сентиментному анализу: отрывки с цитатами прямой и косвенной речи из новостей, а также тексты блогов, причем последние включали 3 тематики: отзывы

[2] Пример такой инверсии сентимента из данных РОМИП (id отрывков 1049, 1188) см. в разделе «Результаты системы ATEX и их анализ».

о фильмах, книгах и фотокамерах. Тестирование системы проводилось в следующих дорожках:

Новостные фрагменты:

- дорожка по классификации прямой и косвенной речи из новостных лент — 3 класса: положительные, отрицательные, нейтральные (не содержащие оценки).

Отзывы о товарах:

- дорожка по классификации отзывов пользователей на 2 класса: положительные и отрицательные;
- дорожка по классификации отзывов пользователей на 3 класса: положительные, отрицательные и содержащие достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

3.2. Подготовка системы к сентиментной классификации

Для каждого вида данных на семинаре были предоставлены размеченные выборки для обучения системы. Следует подчеркнуть, что в цели участия в семинаре входило тестирование системы ATEX с исходными настройками, в том числе для новых неисследованных тематик. Поэтому тренировочные выборки не использовались для обучения системы и подготовки к тестовому этапу. Система была настроена заблаговременно в ходе работы над текстами с русскоязычных новостных сайтов русского и казахского доменов на тему политики и экономики — в частности, сентиментно размеченного корпуса, состоящего из 3 тыс. предложений.

Значение сентимента предложения в системе вычисляется как знак среднего арифметического значений сентиментов входящих в него слов. Позитивное или негативное значение сентимента предложения, как и слова, обозначается соответственно как «+1» или «−1». При этом если количество положительно и отрицательно окрашенных слов в предложении одинаково, в том числе и равно нулю, то общий сентимент предложения получается нейтральным. Таким образом, в системе не проводится различие между «нейтральным» сентиментным классом и классом, содержащим достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

Так как большинство данных содержало отрывки, состоящие из более чем одного предложения, система тестировалась в двух режимах:

1. «С предложениями»: вычислялся сентимент каждого предложения в отрывке. Общий сентимент отрывка вычислялся как среднее арифметическое между сентиментами предложений.
2. «Без предложений»: сентимент всего отрывка вычислялся как среднее арифметическое между сентиментами всех входящих в него слов, без учета границ предложений.

3.3. Данные тестирования РОМИП

В Таблице 1 представлена статистика размеченных тестовых данных РОМИП по сентиментным дорожкам. Новостные отрывки были размечены на 3 класса; отрывки из блогов были размечены и оценивались двумя способами: на 2 и 3 класса. Учитывая настройку системы на новостных текстах политической и экономической тематик, именно в ней ожидается получить наиболее высокие результаты сентиментного анализа.

Таблица 1. Статистика тестовых данных РОМИП

Тематика	Всего отрывков	Положительных	Отрицательных	Нейтральных/ содержащих + и −	Процент наибольшего класса во всей выборке, %
Новости	4573	1448	1234	1890	41
Фильмы	408	330	78	—	80
		266	63	79	65
Книги	129	112	17	—	87
		100	9	20	78
Камеры	411	397	14	—	97
			7	53	85

4. Результаты системы ATEX и их анализ

Оценка результатов работы систем сентиментного анализа проводилась на основе четырех показателей: Аккуратность (Accuracy), Полнота (Recall), Точность (Precision), Мера F1 (F-measure) ([Chetviorkin]). В Таблице 1 видно, что тестовые данные не являются сбалансированными относительно итоговых сентиментных классов; поэтому для представления относительных результатов работы систем они были упорядочены по значению F-measure, которое является более подходящей оценкой, чем Accuracy, для несбалансированных данных [van Rijsbergen]. В таблицах ниже приведены результаты работы различных систем; результаты представленной в данном докладе системы выделены жирным. Выведены наилучшие четыре результата, упорядоченные по F-measure.

Таблица 2. Результаты сентиментной классификации отрывков прямой и косвенной речи из новостей

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	xxx-4	news	3	0,626	0,616	0,621	0,616	
2	**ATEX**	**news**	**3**	**0,606**	**0,579**	**0,592**	**0,571**	**без предложений**
3	**ATEX**	**news**	**3**	**0,606**	**0,576**	**0.590**	**0,569**	**с предложениями**
4	xxx-5	news	3	0,579	0,568	0,574	0,575	

Согласно ожиданиям, результаты по тематике «новости» оказались высокими и абсолютно, и относительно среди других систем. Это говорит о том, что настройка системы на новостных текстах оказалась полезной, несмотря на различные источники и время появления новостных текстов, используемых для настройки и для тестирования системы.

Таблица 3. Результаты сентиментной классификации на 2 класса блогов по тематике «Фильмы»

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	**ATEX**	**film**	**2**	**0,695**	**0,719**	**0,707**	**0,806**	**с предложениями**
2	xxx-23	film	2	0,731	0,641	0,683	0,831	
3	xxx-2	film	2	0,667	0,687	0,677	0,787	
4	xxx-12	film	2	0,759	0,586	0,661	0,828	

Несмотря на различие в тематиках, система показала наилучший результат в дорожке по классификации отзывов о фильмах на два класса. Следует отметить, что в классификации на 2 класса отзывов о книгах и о фотокамерах система занимает третью и пятую строки соответственно относительно других систем. Предположительно, язык выражения сентимента в описании фильмов оказывается наиболее близким к языку выражения сентимента в политике и экономике, и по-видимому, наиболее общим, не обладающим большим количеством специфических сентиментных слов и выражений. В действительности, для сравнения, описания фотокамер содержат большое количество подробностей о функциональных качествах, свойствах самих камер, которые не могут быть освоены без знакомства с самой тематикой и создания специфических правил; что делает такие тексты специфическими и близкими к техническим описаниям.

Таблица 4. Результаты сентиментной классификации на 3 класса блогов по тематике «Фильмы»

Number	System ID	Object	Classes	Precision Macro	Recall Macro	F_Measure Macro	Accuracy	
1	xxx-11	film	3	0,569	0,479	0,520	0,694	
2	**ATEX**	**film**	**3**	**0,486**	**0,521**	**0,503**	**0,596**	**с предложениями**
3	xxx-0	film	3	0,505	0,477	0,491	0,627	
4	xxx-7	film	3	0,566	0,429	0,488	0,360	

Предположение о более общих сентиментных моделях в тематиках фильмов, политики и экономики подтверждается также в результатах классификации на 3 класса: система занимает вторую строку в тематике фильмов, и пятую и шестую строки соответственно для тематик камер и книг.

Важно отметить, что система, основанная на правилах в сравнительном анализе значений F-measure и Accuracy результатов, получает высокий показатель F-measure при относительно более низком значении Accuracy. Это говорит о более равномерном механизме классификации такой системы относительно других систем, особенно при условиях несбалансированной обучающей выборки вероятностных систем, которые при тестировании, предположительно, демонстрируют «перекос» результатов в сторону наиболее частотного класса, что проявляется в их высоком значении Accuracy, но низком значении F-measure относительно системы, основанной на правилах. С другой стороны, это характеризует относительно более высокую воспроизводимость результатов последней на различных данных.

Для иллюстрации работы сентиментных правил приводятся примеры работы системы на цитатах из новостных лент. Подчеркиванием выделены слова, получившие соответствующую тональность в результате работы всей последовательности сентиментных правил и повлиявшие на правильный конечный результат.

Таблица 5. Примеры работы системы для цитат из новостных лент

Id отрывка	Текст	Общий сентимент
1049	«На данный момент не вижу перспективы(−1) никаких военных действий за исключением мер по защите дипломатических представителей, а также справедливого наказания ответственных за эту ужасную акцию(−1)», — сказал Терци.	−1
1068	«Льоренте все еще принадлежит Атлетику и, похоже, готов играть. Впрочем, мы все равно потеряли(−1) одного отличного(0) футболиста(0) и хорошего(0) человека(0)», — сказал Бьелса, намекая на уход Хави Мартинеса в мюнхенскую «Баварию».	−1
1108	«В период после нашей предыдущей встречи мировая экономика по-прежнему испытывала немалые трудности и продолжает подвергаться рискам падения; финансовые рынки остаются(−1) нестабильными, тогда как высокий уровень(−1) дефицита госсектора и государственной задолженности в некоторых развитых экономиках в значительной мере сдерживает процесс восстановления экономики», — отмечается в документе.	−1
1151	«Еще одна трата(+1) на проведение саммита — обеспечение безопасности. Но деньги пошли на обеспечение спецслужб, оборудование не будет выброшено(0), но будет использовано для проведения Универсиады в Казани, Олимпиады в Сочи, на форумы «восьмерки» и «двадцатки». Ничего не пропадает(0). Все траты в целом абсолютно обоснованы», — подчеркнул Путин.	+1
1188	Конкуренция — не проблема(+1) для меня<…>.	+1
7943	«Принять данный документ позволило повышение(+1) возможностей медицинских учреждений по диагностике и лечению заболеваний», — констатируют в оборонном ведомстве.	+1

5. Выводы и дальнейшая работа

Система, основанная на правилах, настроенная на новостных текстах без дополнительной настройки и обучения, в сентиментной классификации на 2 и 3 класса для различных тематик демонстрирует хорошие результаты, сравнимые с результатами систем, в том числе обученных на текстах соответствующих тематик. Особенно высокие показатели полноты, точности и F-measure система демонстрирует, как и ожидалось, в тематике новостей, а также в тематике отзывов о фильмах, что характеризует особенности выражения сентимента в последней.

Числовые показатели оценки говорят о высокой воспроизводимости результатов системы на различных текстах в различных тематиках, при отсутствии тренировочной размеченной выборки и связанных с ней ограничений.

При более детальном исследовании результатов в дальнейших работах наиболее полезной представляется информация о сработавших в ходе сентиментного анализа правилах. Это позволило бы, во-первых, сформировать статистику наиболее частотных моделей выражения сентимента; во-вторых, охарактеризовать различные тематики исследования с точки зрения специфических присущих им моделей, правил и лексики; наконец, это создало бы основу для автоматического создания и лексического наполнения недостающих правил.

В дальнейшем будет полезно, учитывая значительный объем текстов в некоторых тематиках и важность понятия «нейтрального» сентиментного класса, настраивать систему с помощью машинного обучения. В качестве параметров следует использовать количество и, возможно, качество положительных, отрицательных и нейтральных слов в тексте, обработанном системой. В результате следует с помощью алгоритма обучения настраивать общий сентиментный класс, соответствующий всему тексту. Такое дополнение позволило бы, во-первых, более четко определять границу между сентиментным и нейтральным текстом; во-вторых, разграничивать действительно нейтральные тексты как не содержащие сентимент от текстов, в которых указываются достаточно значимые положительные и отрицательные стороны оцениваемой сущности.

Литература

1. *Balahur A., Montoyo A.* Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification, Proc. AISB Convention Comm., Interaction and Social Intelligence. 2008.
2. *Chetviorkin I., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2012. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2013" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013"]. Bekasovo, 2013.
3. *Kan D.* Rule-based approach to sentiment analysis at ROMIP 2011. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]. Bekasovo, 2012.
4. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, no. 2, pp. 1–135. 2008
5. *van Rijsbergen C. J.* Information Retrieval, Butterworths, London, (1979)
6. *Sokolova M., Lapalme G.* A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4, pp. 427–437. Jul. 2009
7. *Vasilyev V. G., Khudyakova M. B., Davydov S.* Sentiment classification by fragment rules. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2012" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012"]. Bekasovo, 2012.
8. *Wiebe J. M.* Tracking point of view in narrative. Computational Linguistics 20 (2), pp. 233–287. 1994.
9. *Zaliznyak A.* Grammaticheskij slovar' russkogo jazyka. Moskva, 1977, (further editions are 1980, 1987, 2003).

Раздел IV.

Машинный перевод

AUTOMATIC EVALUATION OF MACHINE TRANSLATION QUALITY[1]

Màrquez L. (lluism@lsi.upc.edu)

TALP Research Center Software Department
Technical University of Catalonia

This paper contains an extended abstract of the invited talk presented, with the same title, at Dialogue 2013, the 19th International Computational Linguistics Conference. The presentation will cover several works carried out at the Technical University of Catalonia (UPC) in collaboration with several researchers and colleagues. I would like to especially mention the following: Jesús Giménez, Meritxell Gonzàlez, Lluís Formiga and Laura Mascarell. Sincere thanks to all of them.

1. Introduction

Automatic evaluation of machine translation (MT) quality deals with computing the similarity between a system's output and one or several reference translations for a given source text. Automatic machine translation evaluation metrics are not only useful to provide a quality assessment for machine translation results, but also constitute important guidance for machine translation development and tuning. One real challenge in MT evaluation is that the similarity measure should be able to discriminate whether translation and reference texts convey the same meaning, so the comparison is at the level of semantic equivalence. Additionally, MT evaluation is an open task, that is, there is not a unique *good translation* for a given source text. Instead, a large number of variants can be usually considered correct translations or acceptable to some extent. Indeed, translation correctness is not black and white, but a matter of degree. Finally, it is well known that translation quality aspects are diverse and they complement each other (fluency, adequacy, grammaticality, etc.).

The first approaches to automatic MT evaluation were all based on lexical similarity. Lexical measures (also called *n*-gram-based or string-based measures) work by rewarding lexical matches between automatic translations and a set of manually-produced reference translations. BLEU (Papineni et al., 2002) is the most popular representative in this family, and has been widely accepted as a *de facto* standard for years. However, it has been shown that lexical similarity is neither a sufficient nor a necessary condition for two sentences to convey the same meaning (Culy and Riehemann, 2003; Coughlin, 2003; CallisonBurch et al., 2006). Actually, the reliability of lexical metrics depends very strongly on the heterogeneity and representativity of reference translations. It has been shown that currently used lexical metrics have

trouble distinguishing raw, inadequate machine translation output from fully fluent and adequate translation obtained from them through professional postediting (Denkowski and Lavie, 2012). Note that lexical-based metrics are not able to capture the syntax or semantic structure of sentences; therefore, they are not directly sensitive to the improvement of machine translation systems on these aspects. Moreover, they tend to favor statistical MT systems when compared to rule-based MT or other paradigms in a particular data set.

2. Linguistically-motivated evaluation measures

In order to cope with the above mentioned issues, a number of authors have suggested exploiting linguistic information beyond the lexical level to increase robustness. Some have used additional linguistic knowledge to extend the reference lexicon. For instance, Rouge, Meteor and TER allow for morphological variations via stemming. TER and Meteor may perform an additional dictionary-based lookup for synonyms and paraphrases (Snover et al., 2010; Denkowski and Lavie, 2010). Russo-Lassner et al. (2005), Zhou et al. (2006), Kauchak and Barzilay (2006), and Owczarzak et al. (2006) have also studied the use of automatically-generated paraphrases to find potential phrase matchings.

In a complementary direction, Dreyer and Marcu (2012) introduced HyTER, an edit-distance based metric designed to avoid the problem of comparing to a very reduced set of references. In the HyTER approach, human annotations are used to construct networks encoding a exponentially large number of meaning-equivalent reference translations, and the similarity is computed over the whole set of translation equivalents.

Other authors have suggested modeling language variability by directly comparing the syntactic and semantic structure of candidate and reference translations. For instance, Reeder et al. (2001) defined a similarity measure based on named entity overlap. Liu and Gildea (2005) introduced several syntactic measures based on comparing head-word dependency chains and constituent subtrees. Popović and Ney (2007) proposed a series of measures based on edit distance over parts of speech. Owczarzak et al. (2007) presented a measure based on comparing dependency structures from a probabilistic lexical-functional grammar parser. Mehay and Brew (2007) defined a measure based on combinatorial categorial grammar parsing which differs from others in that it does not require the parse of the possibly illformed automatic candidate translations, but only the parse of the reference translations. Kahn et al. (2009) used a probabilistic context-free grammar parser and deterministic head-finding rules. Chan and Ng (2008) presented MaxSim, a general framework which allows for using arbitrary similarity functions between items, and to incorporate different information in the comparison (dependency relations, lemmas, parts of speech and synonymy lookup). Padó et al. (2009) suggested measuring the quality of MT output through its semantic equivalence to the reference translation, based on a set of textual entailment features. Finally, Lo and Wu (2011) introduced MEANT, a semi-automated metric, which assesses translation utility by matching semantic role fillers.

Over the last years, at UPC we have worked on the definition of generic MT evaluation measures which include information at different linguistic levels, ranging from lexical to syntactic and semantic (Giménez and Màrquez, 2010b). This approach is based on the assumption that measures at different levels capture different aspects of translation quality. So, rather than looking for the single best evaluation metric, we aim at combining several partial measures to provide a richer and broader assessment of translation quality. A toolkit for MT evaluation, called Asiya has been developed to integrate all previous measures (Giménez and Màrquez, 2010a; Giménez and Màrquez, 2010b).[2]

Although linguistically-enriched evaluation measures have shown good properties and higher correlation with human assessments at several MT evaluation campaigns, they are not still widely adopted by developers and researchers when doing real machine translation evaluations and comparisons. One of the actual problems of such methods, compared to lexical metrics, is robustness. They can be unreliable in certain situations, because they depend strongly on parsers or machine learning algorithms which are trained on specific corpora and because these parsers may fail when applied to the generally noisy text output by a translation system. From the point of view of system tuning, another issue with linguistically-rich based metrics is their high computational cost, which prevented them from being introduced in costly optimization and tuning procedures.

3. Intelligent MT output and error analysis

In MT system development, a qualitative analysis of translation quality is a fundamental step in order to spot the limitations of a system, compare the linguistic abilities of different systems or tune the parameters during system refinement, among others. The need for analyzing and comparing automatic translations with respect to evaluation metrics is also paramount for developers of translation quality measures, who need elements of analysis to better understand the behavior of their evaluation measures.

Existing measures for MT quality evaluation, and especially those working at higher linguistic levels, can be very useful for assisting a manual exploration of MT output and its error analysis. At UPC, we have been working on this direction by providing a web-based version of the Asiya evaluation toolkit, called Asiya Online Interface, which provides a graphical visualization and an interactive access to the evaluation results (Gonzàlez et al., 2012).

The benefits of the online interface are multiple. First, it facilitates the use of the Asiya toolkit for rapid evaluation of test beds. Second, it aids the analysis of the errors produced by MT systems under comparison by creating meaningful visualizations of the information related to the evaluation metrics. The intermediate structures generated by the parsers used to compute the metric scores are priceless for MT developers, who can use them to compare the structures of several translations and see how they affect the internal performance of the metrics, providing more understanding

[2] Find the ASIYA toolkit available at the following URL: www.lsi.upc.edu/~nlp/Asiya/

in order to interpret the actual performance of the automatic translation systems. Finally, search capabilities have been also included into the Asiya On-Line Interface for an intelligent analysis of MT output and system comparison. The search module, *t*Search, is build on top of Asiya and connected to the Asiya On-line Interface. It provides a flexible query language, which allows to retrieve and export from the test bed all the translation examples satisfying virtually any criterion related to the evaluation measures (including a large number of alternative metrics, their numerical scores, and any internal syntactic and semantic structure of their intermediate analyses) and the MT systems under comparison.

Currently, there are no freely available automatic tools for aiding MT evaluation tasks. For this reason, we believe that *t*Search can be a very useful tool for MT system and evaluation metric developers. So far, other related works in the field addressed (semi)-automatic error analysis from different perspectives. A framework for error analysis and classification was proposed in (Vilar et al., 2006), which has inspired more recent works in the area, such as (Fishel et al., 2011). They propose a method for automatic identification of various error types. The methodology proposed is language independent and tackles lexical information. Nonetheless, it can also take into account language-dependent information if linguistic analyzers are available. The user interface presented in (Berka et al., 2012) provides also automatic error detection and classification. It is the result of merging the Hjerson tool (Popović , 2011) and Addicter (Zeman et al., 2011). This web application shows alignments and different types of errors colored. In contrast, the Asiya interface and the *t*Search tool together facilitate the qualitative analysis of the evaluation results yet providing a framework to obtain multiple evaluation metrics and linguistic analysis of the translations. They also provide the mechanism to search and find relevant translation examples using a flexible query language and export the results.

4. Quality Estimation

The term Quality Estimation (QE) refers to the task of estimating translation quality in the absence of human reference translations (Specia et al., 2010; Callison-Burch et al., 2012). That is, the only information available is that of the source and translated texts and, possibly, some information on the translation system itself. This problem was already introduced ten years ago (Blatz et al., 2003), with the term *Confidence Estimation*, but it has not been until more recently that it concentrated a broader attention from the community, with the creation of specific shared tasks for evaluating QE systems and approaches under the umbrella of the WMT workshops on Statistical Machine Translation (Callison-Burch et al., 2012).[3]

QE measures have a wide range of applications in practical MT system development, analysis and usage. For instance, they can be useful for: system parameter tuning, informing MT end-users about estimated translation quality, quality-oriented

[3] The 2013 edition is also under development. Find more information at: http://www.statmt.org/wmt13/ quality-estimation-task.html

filtering of translation cases (e.g., to identify translations requiring manual post-edition, or to identify casual users' post-editions that are useful for enriching the MT system), selecting the best translation among a set of alternatives (e.g., in a system combination scenario), etc.

Quality Estimation is usually addressed as a scoring task (Specia et al., 2009; Specia et al., 2010), where some regression function predicts the absolute quality of the automatic translation of a source text. QE has recently evolved towards two separate subtasks consisting in scoring itself and ranking, where different MT outputs for a given source sentence have to be ranked according to their comparative quality. Results obtained so far on QE have been more satisfactory for the ranking approach (Specia et al., 2010; Avramidis, 2012; Callison-Burch et al., 2012).

System ranking based on human quality annotations has been established as a common practice for MT evaluation in shared tasks (Callison-Burch et al., 2012). Therefore, training corpora are available for researchers to train ranking functions with supervised machine learning methods to perform automatic ranking mimicking human annotations. Learned models can be reusable, provided they are system independent and based on a generic analysis (i.e., no system dependent features can be used for training), and applicable to other sets containing any input and multiple outputs. The applications of *QE-for-ranking* are diverse: from hybrid MT system combination to their internal optimization and evaluation. The most popular practical scenario of QE models (both rankers and regressors) consists of ranking alternative MT systems' outputs to predict the best translation at segment level.

It is worth noting that the research conducted in QE for training ranking models from human annotations has always been done in *controlled environments*, consisting of well-formed text with little presence of noise (such as News or EU Parliament acts). However, MT in real life has to deal with a more complex scenario, including non-standard usage of text (e.g., social media, blogs, reviews, etc.), which is totally open domain and prone to contain ungrammaticalities and errors (misspellings, slang, abbreviations, etc.). An example of noisy environment is found in the publicly available FAUST corpus[4] (Pighin et al., 2012b), collected from the 24/7 Reverso.net MT service. This corpus is composed of 1,882 weblog source sentences translated with 5 independent MT systems. The systems were ranked according to human assessments of adequacy by several users using a graph-based methodology, obtaining considerably high agreement and quality indicators (Pighin et al., 2012a).

At UPC we have studied the supervised training of QE prediction models from the aforementioned FAUST corpus to rank alternative system translations. Our study focused on different aspects, such as: *i*) the typology of the problem (regression vs. binary classification), *ii*) suitability of the learning algorithm, and *iii*) best combination of features to learn. Results showed that is possible to build reliable QE models from an annotated *real life MT* corpus. Concretely, correlation results are comparable to those described in the literature for standard text. Furthermore, we also observed that comparative (ranked–based) QE models fit better to the system selection task (i.e. predict always the best translation) compared to absolute (regression-based) QE models.

---

[4] http://www.faust-fp7.eu/faust/Main/DataReleases

References

1. *Avramidis, Eleftherios*. 2012. Comparative quality estimation: Automatic sentence-level ranking of multiple machine translation outputs. In Proceedings of 24th International Conference on Computational Linguistics (COLING), pages 115–132, Mumbai, India.
2. *Berka, Jan, Ondrej Bojar, Mark Fishel, Maja Popović , and Daniel Zeman.* 2012. Automatic MT error analysis: Hjerson helping Addicter. In Proceedings of the 8th International Conference on Language Resources and Evaluation, pages 2158–2163, Istanbul, Turkey.
3. *Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing.* 2003. Confidence estimation for machine translation. Final Report of Johns Hopkins 2003 Summer Workshop on Speech and Language Engineering. Technical report, Johns Hopkins University.
4. *Callison-Burch, Chris, Miles Osborne, and Philipp Koehn.* 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
5. *Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia*. 2012. Findings of the 2012 workshop on statistical machine translation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 10–51, Montréal, Canada, June.
6. *Chan, Yee Seng and Hwee Tou Ng.* 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In Proceedings of ACL-08: HLT, pages 55–62.
7. *Coughlin, Deborah.* 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In Proceedings of Machine Translation Summit IX, pages 23–27.
8. *Culy, Christopher and Susanne Z. Riehemann.* 2003. The Limits of N-gram Translation Evaluation Metrics. In Proceedings of MT-SUMMIT IX, pages 1–8.
9. *Denkowski, Michael and Alon Lavie.* 2010. Meteornext and the meteor paraphrase tables: Improved evaluation support for five target languages. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 339–342, Uppsala, Sweden, July. Association for Computational Linguistics.
10. *Denkowski, Michael and Alon Lavie*. 2012. Challenges in predicting machine translation utility for human post-editors. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA'2012), San Diego, CA, USA, October. AMTA.
11. *Dreyer, Markus and Daniel Marcu.* 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 162–171, Montréal, Canada, June. Association for Computational Linguistics.
12. *Fishel, Mark, Ondˇrej Bojar, Daniel Zeman, and Jan Berka.* 2011. Automatic Translation Error Analysis. In Proceedings of the 14th Text, Speech and Dialogue (TSD).

13. *Giménez, Jesús and Lluís Màrquez.* 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. The Prague Bulletin of Mathematical Linguistics, 94: 77–86.
14. *Giménez, Jesús and Lluís Màrquez.* 2010b. Linguistic features for automatic MT evaluation. Machine Translation, 24(3–4): 209–240.
15. *Gonzàlez, Meritxell, Jesús Giménez, and Lluís Màrquez.* 2012. A graphical interface for MT evaluation and error analysis. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). System Demonstrations, pages 139–144, Jeju, South Korea, July.
16. *Kahn, Jeremy G., Matthew Snover, and Mari Ostendorf.* 2009. Expected Dependency Pair Match: Predicting translation quality with expected syntactic structure. Machine Translation.
17. *Kauchak, David and Regina Barzilay.* 2006. Paraphrasing for Automatic Evaluation. In Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLTNAACL), pages 455–462.
18. *Liu, Ding and Daniel Gildea.* 2005. Syntactic Features for Evaluation of Machine Translation. In Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, pages 25–32.
19. *Lo, Chi-kiu and Dekai Wu.* 2011. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 220–229, Portland, Oregon, USA, June. Association for Computational Linguistics.
20. *Mehay, Dennis and Chris Brew.* 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI).
21. *Owczarzak, Karolina, Declan Groves, Josef Van Genabith, and Andy Way.* 2006. Contextual BitextDerived Paraphrases in Automatic MT Evaluation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pages 148–155.
22. *Owczarzak, Karolina, Josef van Genabith, and Andy Way.* 2007. Dependency-Based Automatic Evaluation for Machine Translation. In Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation, pages 80–87.
23. *Padó, Sebastian, Michel Galley, Daniel Jurafsky, and Christopher D. Manning.* 2009. Machine translation evaluation with textual entailment features. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 37–41, Athens, Greece, March. Association for Computational Linguistics.
24. *Papineni, Kishore, Salim Roukos, Todd Ward, and WeiJing Zhu.* 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).
25. *Pighin, Daniele, Lluís Formiga, and Lluís Màrquez.* 2012a. A graph-based strategy to streamline translation quality assessments. In Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA'2012), San Diego, USA, October.

26. *Pighin, Daniele, Lluís Màrquez, and Jonathan May.* 2012b. An Analysis (and an Annotated Corpus) of User Responses to Machine Translation Output. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.
27. *Popović, Maja and Hermann Ney.* 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 48–55, Prague, Czech Republic, June. Association for Computational Linguistics.
28. *Popović, Maja.* 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. The Prague Bulletin of Mathematical Linguistics, 96: 59–68.
29. *Reeder, Florence, Keith Miller, Jennifer Doyon, and John White.* 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. In Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at Machine Translation Summit VIII, pages 55–59.
30. *Russo-Lassner, Grazia, Jimmy Lin, and Philip Resnik.* 2005. A Paraphrase-Based Approach to Machine Translation Evaluation (LAMP-TR-125/CSTR-4754/UMIACS-TR-2005-57). Technical report, University of Maryland, College Park.
31. *Snover, Matthew G., Nitin Madnani, Bonnie Dorr, and Richard Schwartz.* 2010. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. Machine Translation, 23(2–3): 209–240.
32. *Specia, Lucia, Marco Turchi, Nicola Cancedda, Mark Dymetman, and Nello Cristianini.* 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT-2009), pages 28–35, Barcelona, Spain.
33. *Specia, Lucia, Dhwaj Raj, and Marco Turchi.* 2010. Machine Translation Evaluation Versus Quality Estimation. Machine Translation, 24:39–50, March.
34. *Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney.* 2006. Error Analysis of Machine Translation Output. In Proc. 5th Intl. Conference on Language Resources and Evaluation (LREC), pages 697–702, Genoa, Italy.
35. *Zeman, Daniel, Mark Fishel, Jan Berka, and Ondrej Bojar.* 2011. Addicter: What Is Wrong with My Translations? The Prague Bulletin of Mathematical Linguistics, 96: 79–88.
36. *Zhou, Liang, Chin-Yew Lin, and Eduard Hovy.* 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 77–84.

ДОРОЖКА ПО ОЦЕНКЕ МАШИННОГО ПЕРЕВОДА ROMIP MTEVAL 2013: ОТЧЕТ ОРГАНИЗАТОРОВ

Браславский П. (pbras@yandex.ru)
Kontur Labs; Уральский федеральный университет,
Екатеринбург, Россия

Белобородов А. (xander-beloborodov@yandex.ru)
Уральский федеральный университет,
Екатеринбург, Россия

Шаров С. (s.sharoff@leeds.ac.uk)
University of Leeds, Лидс, Великобритания

Халилов М. (maxim@tauslabs.com)
TAUS Labs, Амстердам, Нидерланды

Ключевые слова: машинный перевод, оценка, англо-русский перевод

ROMIP MT EVALUATION TRACK 2013: ORGANIZERS' REPORT

Braslavski P. (pbras@yandex.ru)
Kontur labs; Ural Federal University, Russia

Beloborodov A. (xander-beloborodov@yandex.ru)
Ural Federal University, Russia

Sharoff S. (s.sharoff@leeds.ac.uk)
University of Leeds, Leeds, UK

Khalilov M. (maxim@tauslabs.com)
TAUS Labs, Amsterdam, Netherlands

The paper presents the settings and the results of the ROMIP 2013 machine translation evaluation campaign for the English-to-Russian language pair. The quality of generated translations was assessed using automatic metrics and human evaluation. We also demonstrate the usefulness of a dynamic mechanism for human evaluation based on pairwise segment comparison.

Keywords: machine translation, evaluation, English-to-Russian translation

1. Введение

Русский и английский были одной из первых языковых пар на заре исследований в этой области машинного перевода (МП) в 1950-х годах [Hutchins2000]. С тех пор парадигмы МП поменялись много раз, многие системы для этой языковой пары появлялись и исчезали, но, насколько нам известно, до сих пор не проводилась систематическая сравнительная оценки систем МП, аналогичная DARPA'94 [White et al., 1994] и более поздним мероприятиям. Семинар по статистическому машинному переводу (Workshop on Statistical Machine Translation, WMT) в 2013 году впервые включил русско-английскую пару в свою программу.[1] На данный момент эта оценка еще не проведена, к тому же в семинаре примут участие системы, обученные на данных, предоставленных организаторами. За рамками оценки останутся существующие системы, в частности — системы на основе правил и гибридные системы.

Кампании по оценке играют важную роль в развитии технологий МП. В последнее время был проведен ряд открытых кампаний для различных комбинаций европейских, азиатских и семитских языков, см. [Callison-Burch et al., 2011; Callison-Burch et al., 2012; Federico et al., 2012]. В этой статье мы описываем кампанию по оценке англо-русского машинного перевода в рамках РОМИП.

РОМИП (Российский семинар по Оценке Методов Информационного Поиска)[2] — это российский аналог TREC и других инициатив по оценке задач информационного поиска. Первый цикл оценки был организован в 2002 году. В течение этих десяти лет РОМИП организовал серию дорожек по оценке, включая классическую задачу поиска по запросу, задачи тематической классификации документов, вопросно-ответного поиска, формирования сниппетов, анализа тональности текста, поиска изображений и т. д. В рамках этой деятельности было подготовлено несколько свободно распространяемых наборов данных, содержащих документы и оценки релевантности, сделанные асессорами. Российские сообщества, занимающиеся информационным поиском и машинным переводом, имеют давние связи, их представители тесно общаются. Поэтому было естественным организовать кампанию по оценке МП в рамках РОМИП, используя накопленный опыт семинара. Кроме того, важной целью мероприятия была консолидация групп, разрабатывающих как статистические системы МП (SMT), так и системы, основанные на правилах (RBMT).

Одна из проблем для систем МП, работающих с русским языком, и для их оценки — это необходимость иметь дело с относительно свободным порядком слов в предложении и развитой морфологией. За счет развитой морфологии у русских лемм много словоформ (в среднем 8,2 формы для существительных, 34,6 — для глаголов [Sharoff et al., 2013]), что осложняет выравнивание на уровне слов при статистическом подходе. Дистантные зависимости создают дополнительные проблемы, особенно для SMT-систем.

---

[1] http://www.statmt.org/wmt13/

[2] http://romip.ru

Для оценки было выбрано одно направление перевода (английский → русский). Во-первых, для этого направления нам намного проще было найти асессоров, для которых целевой язык является родным. Во-вторых, системы-участницы в основном используются именно в этом направлении (перевод английских текстов для русскоязычных пользователей).

2. Данные

При формировании тестового корпуса текстов мы руководствовались двумя соображениями. Во-первых, известно, что предметная область и жанр текста влияют на качество перевода [Langlais, 2002; Babych et al., 2007]. Таким образом, мы хотели обеспечить хотя бы минимальное жанровое разнообразие текстов, входящих в корпус. Во-вторых, мы хотели использовать источники, допускающие дальнейшее распространение текстов по лицензии Creative Commons. В итоге корпус был сформирован из двух источников, соответственно — из текстов двух жанров. Новостные тексты было собраны с английского раздела Wikinews[3]. Формальные тексты (регламенты, инструкции, положения, официальные документы) был собраны из Веба с использованием жанрового классификатора [Sharoff, 2010]. После применения автоматической классификации был проведен ручной отбор текстов.

Начальный корпус состоял из 8356 оригинальных документов общим объемом 148 864 английских предложений. В корпусе были представлены оригинальные документы целиком, т. к. некоторые системы могут использовать для перевода контекст предложения. Источник 100 889 предложений в корпусе — Wikinews; 47 975 предложений относятся к формальным текстам. Первые 1002 предложения были опубликованы заранее, чтобы участники могли адаптировать свои системы к используемому формату. Так как корпус был подготовлен полностью автоматически, он не лишен дефектов (например, часто встречается некорректная разбивка на предложения, остатки HTML-разметки и т. п.). Участники должны были прислать организаторам русские переводы 147 862 предложений в течение недели после публикации исходного тестового корпуса.

Примеры предложений тестового корпуса:

90237	*Ambassadors from the United States of America, Australia and Britain have all met with Fijian military officers to seek insurances that there wasn't going to be a coup.*
102835	*If you are given a discount for booking more than one person onto the same date and you later wish to transfer some of the delegates to another event, the fees will be recalculated and you will be asked to pay additional fees due as well as any administrative charge.*

[3] http://en.wikinews.org/

Тексты в исходном корпусе не были до этого переведены на русский язык, т. е. системы-участники не могли заранее использовать переводы для обучения. Для оценки мы выбрали 947 «чистых» предложений (т. е. с корректными границами, без паразитной HTML разметки и т. п.), из них 759 — новостных и 188 — из формальных текстов.

Эти предложения примерно равными порциями были назначены для перевода трем переводчикам (переводчик 1: предложения 1–316; переводчик 2: 317–632; переводчик 3: 633–947). Переводчики 1 и 2 сообщили, что они потратили от 20 до 30 часов на перевод всего задания. Оба переводчика сообщили, что время, потраченное на перевод отдельного предложения значительно различалось. В отличие от перевода связного текста, дополнительная сложность возникает из-за необходимости переключаться между темами и понимать контекст предложения (переводчикам иногда приходилось обращаться к набору данных, содержащему оригинальные документы). Переводчик 3 не смог выполнить перевод в срок, поэтому ему принадлежат только 152 перевода в третьей порции. Остальные предложения переведены двумя членами одной из групп-участниц. Все 947 переводов использовались для автоматической оценки качества переводов, 330 предложений из 947 были выбраны для ручной оценки (190 новостных и 140 формальных текстов).

Дополнительно мы сделали объявление в списке рассылки конкурса, нескольких онлайн форумах переводчиков и в группах Facebook с просьбой принять участие в коллективном переводе тестовых предложений на сайте TranslatedBy.[4] Сравнение профессионального и коллективного перевода — тема отдельного исследования.

Дополнительно организаторы предоставили участникам доступ к следующим ресурсам:

- 1M предложений англо-русского параллельного корпуса, распространяемого Яндексом (этот корпус используется в WMT13)[5];
- 119K предложений англо-русского параллельного корпуса из репозитория TAUS.

Эти наборы данных не связаны с корпусом, который был подготовлен в рамках кампании по оценке; цель этих дополнительных данных — снизить порог участия для групп, которые не имеют собственных данных достаточного объема для этого направления перевода.

3. Ручная и автоматическая оценка

Основной принцип, который мы хотели реализовать в ручной оценке, — сделать оценку как можно более простой для асессора, а ее результаты — интерпретируемыми. Мы выбрали вариант ранжирования систем на основе попарных сравнений вариантов перевода. Такой подход отличается от *ранжирования* нескольких

[4] http://translatedby.com

[5] http://translate.yandex.ru/corpus

вариантов перевода асессором — подхода, который используется в рамках экспериментов по оценке WMT. В случае большого количества участвующих систем асессоры каждый раз ранжируют только часть вариантов переводов. На основе частичных рангов не всегда просто получить однозначное полное ранжирование систем [Callison-Burch et al., 2012]. На основании попарных сравнений проще построить общее ранжирование, к тому же попарные сравнения — более простая задача для асессора. Однако такой метод подразумевает больший объем оценки (который все же остается приемлемым в случае небольшого количества участвующих систем). Ниже мы обсуждаем, как можно снизить объем ручной оценки.

В нашем случае асессоры должны были делать попарные сравнения двух предложений — переводов участвующих систем — с образцовым переводом, выполненным человеком. Асессор должен был выбрать лучший из двух вариантов или отметить, что оба варианта эквивалентны. При этом асессор не видел исходное предложение, а только человеческий перевод.

Как было сказано выше, 330 тестовых предложений были задействованы в ручной оценке. Исходная идея состояла в том, чтобы генерировать пары предложений для оценки динамически для оптимизации объема оценки. К сожалению, ограничения используемого инструмента оценки не позволили реализовать такой сценарий. Мы были вынуждены проводить полное сравнение — 28 пар на одно тестовое предложение (для 8 систем, участвовавших в ручной оценке). Изначально задачи по оценке были распределены между 11 асессорами (добровольцами и членами участвующих в кампании команд) с небольшим перекрытием. Задания по оценке были распределены таким образом, чтобы все варианты перевода одного предложения оценивались одним асессором, что предположительно должно приводить к более согласованному ранжированию. Недостаток такого подхода — в том, что члены участвующих команд оценивают, в том числе, результаты работы «своих» систем. Незадолго до срока окончания оценки некоторые асессоры сообщили, что не смогут закончить оценку вовремя. Невыполненные задания были переназначены другим асессорам; дополнительно три новых асессора присоединились к оценке. Таким образом всего в оценке приняло участие 14 человек. Переводы 60 тестовых предложений были оценены с двойным перекрытием (таким образом, для 60×28=1680 пар у нас есть решение двух асессоров). Общий объем оценки составил 10 920 попарных сравнений. По сообщениям асессоров, на оценку одной пары уходило от 30 до 90 секунд, при этом для оценки некоторых сложных предложений требовалось до 5 минут.

Для оценки мы использовали многофункциональный инструмент оценки машинного перевода TAUS DQF в режиме «быстрое сравнение» (*quick comparison*).[6]

На основе оценок асессоров системы можно ранжировать для каждого предложения из тестового набора. В случае равенства очков ранги усреднялись. Например, так выглядят ранги, если системы на позициях 2–4, 7–8 имеют равное количество очков:

1 3 3 3 5 6 7.5 7.5

[6] https://tauslabs.com/dynamic-quality/dqf-tools-mt

Для получения общего ранжирования систем ранги на уровне предложений усреднялись по всем предложениям.

После того, как мы получили все попарные сравнения вариантов перевода, мы смогли провести моделирование динамического формирования пар для сравнения и понять, какой объем оценки можно сэкономить с использованием такой методики. Идея состоит в том, чтобы сначала получить предварительное ранжирование систем (например, на основе автоматических метрик), а потом сортировать этот «массив предложений» с помощью алгоритма сортировки вставками (или его варианта с использованием бинарного поиска).

В дополнение к ручной оценке мы также запустили автоматическую оценку, используя следующие метрики: BLEU [Papineni et al. 2001], METEOR [Banerjee and Lavie, 2005], TER [Snover et al., 2009] и GTM [Turian et al., 2003]. BLEU и METEOR могут рассматриваться как метрики близости машинного перевода образцовому; TER и GTM демонстрируют более высокую корреляцию с объемом необходимого постредактирования [O'Brien, 2011].

4. Результаты

Мы получили результаты от пяти участников, две команды прислали по два прогона. Таким образом, в сумме у нас было семь прогонов для оценки (обозначены P1..P7 в данном отчете), см. краткое описание систем в Табл. 1. Как видно из таблицы, в кампании приняли участие как признанные группы из индустрии и академических организаций, так и молодые команды. В оценку были включены также переводы 947 тестовых предложений четырех онлайн систем (обозначены в отчете OS1..OS4). Таким образом, в автоматической оценке участвовало 11 прогонов, в ручной — восемь (четыре онлайн системы и четыре системы-участницы; в ручной оценке не участвовали прогоны P3, P6 и P7).

Таблица 1. Участники ROMIP MTEval 2013

ID	Краткое описание системы
P1	**Compreno (ABBYY)** http://www.abbyy.ru/science/technologies/business/compreno/
P2	**Pharaon (анонимный участник)** Система на основе Moses SMT, использованы корпуса Яндекса и TAUS.
P3,4	**Balagur (Школа анализа данных)** Система на базе MOSES, использован корпус Яндекса (1M) и новостной корпус (200K), собранный по новостным сайтам.
P5	**ЭТАП-3 (ИППИ РАН)** Система перевода на основе правил, использует составленный вручную словарь примерно со 100 000 входов [Boguslavsky1995]
P6,7	**Pereved (МФТИ)** Система основана на Moses и натренирована на параллельных предложениях, извлеченных из Интернета.

Таблица 2. Результаты автоматической оценки

Метрика/ID	OS1	OS2	OS3	OS4	P1	P2	P3	P4	P5	P6	P7
Все (947 предложений)											
BLEU	0,150	0,141	0,133	0,124	0,157	0,112	0,105	0,073	0,094	0,071	0,073
METEOR	0,258	0,240	0,231	0,240	0,251	0,207	0,169	0,133	0,178	0,136	0,149
TER	0,755	0,766	0,764	0,758	0,758	0,796	0,901	0,931	0,826	0,934	0,830
GTM	0,351	0,338	0,332	0,336	0,349	0,303	0,246	0,207	0,275	0,208	0,230
Новости (759 предложений)											
BLEU	0,137	0,131	0,123	0,114	0,153	0,103	0,096	0,070	0,083	0,066	0,067
METEOR	0,241	0,224	0,214	0,222	0,242	0,192	0,156	0,127	0,161	0,126	0,136
TER	0,772	0,776	0,784	0,777	0,768	0,809	0,908	0,936	0,844	0,938	0,839
GTM	0,335	0,324	0,317	0,320	0,339	0,290	0,233	0,201	0,257	0,199	0,217

Таблица 3. Ранжирование систем на основе ручной оценки (усредненные ранги, от лучших к худшим слева направо)

Все (330 предложений)							
OS3	**P1**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
3,159	3,350	3,530	3,961	4,082	5,447	5,998	6,473
Новости (190 предложений)							
OS3	**P1**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
2,947	3,450	3,482	4,084	4,242	5,474	5,968	6,353
Формальные тексты (140 предложений)							
P1	**OS3**	**OS1**	**OS2**	**OS4**	**P5**	**P2**	**P4**
3,214	3,446	3,596	3,793	3,864	5,411	6,039	6,636
Предварительное ранжирование, сортировка вставками							
P1	**OS1**	**OS3**	**OS2**	**OS4**	**P5**	**P4**	**P2**
3,318	3,327	3,588	4,221	4,300	5,227	5,900	6,118
Предварительное ранжирование, сортировка бинарными вставками							
OS1	**P1**	**OS3**	**OS2**	**OS4**	**P5**	**P2**	**P4**
2,924	3,045	3,303	3,812	4,267	5,833	5,903	6,882

Табл. 2 содержит значения автоматических метрик для всех прогонов участников и четырех онлайн систем. По автоматическим метрикам OS1 лидирует на полном наборе тестовых предложений и на предложениях формальных документов, P1 демонстрирует лучший результат на предложениях новостных документов.

Итоговое ранжирование систем на основе ручной оценки представлено в Табл. 3. Внутри трех групп участников разница между усредненными рангами статистически незначима (по t-тесту Уэлча, уровень значимости $p=0,05$):

(OS1, OS3, P1), (OS2, OS4) и (P2, P4). Система P5 располагается между последними двумя группами. Ранжирование систем сохраняется на подмножествах тестового набора, соответствующих новостям и формальным документам. В отличие от ранжирования на основе автоматических метрик (Табл. 2) OS3 входит в тройку лидеров по результатам ручной оценки. Аналогичным образом P5 ранжируется выше, чем P2 по результатам ручной оценки, в то время как автоматические метрики располагают эти системы в обратном порядке. Это наблюдение еще раз подтверждает факт, что автоматические метрики систематически недооценивают качество систем МП, основанных на правилах [Béchar et al., 2012].

Нижняя часть Табл. 3 содержит результаты моделирования ручной оценки систем с динамическим формированием пар предложений для оценки. Системы были предварительно отсортированы на основе метрики NIST (см. Табл. 2). После этого варианты перевода для одного тестового предложения были ранжированы с помощью алгоритма сортировки вставками на основе имеющихся ручных оценок пар. В результате мы получили ранжирование систем, несколько отличающееся от ранжирования на основе полного набора оценок, т. к. при сортировке мы не использовали «усредненные» ранги.[7] При этом ранжирование можно считать идентичным — с точностью до взаимного расположения статистически различных *групп* систем. Преимущество такого подхода в том, что для ранжирования нам достаточно сделать существенно меньше попарных сравнений. В случае классической сортировки вставками нам понадобилось 5131 сравнений (15,5 на одно тестовое предложение; 56% полного набора попарных сравнений для 330 тестовых предложений и 8 систем); сортировка бинарными вставками показала себя еще лучше: 4327 сравнений (13,1 на предложение; 47% от полного набора сравнений). Предположительно, объем оценок можно снизить еще больше, если предварительно ранжировать системы на уровне отдельных предложений.

Показатели согласия асессоров аналогичны показателям при ранжировании вариантов перевода [Callison-Burch et al., 2012; Callison-Burch et al., 2011]: $\kappa=0{,}34$, $\alpha=0{,}48$. Согласованность повышается, если мы рассмотрим только сравнения трех лучших систем с остальными (т. е. не учитываем сравнения внутри групп): $\alpha=0{,}53$. Аналогично, согласованность падает, если мы учитываем только сравнения внутри группы трех лучших систем: $\kappa=0{,}23$, $\alpha=0{,}33$. Эти результаты согласуются с данными о низкой согласованности асессоров в случае оценки систем примерно одинакового уровня [Callison-Burch et al., 2011].

---

[7] Такое итоговое ранжирование не является полностью независимым от метода предварительной сортировки систем. Например, если предварительная сортировка систематически ранжирует одну систему из двух выше, а ручная оценка систематически считает их равными, то в итоговом ранжировании сохранится порядок предварительной сортировки.

5. Заключение

Это был первый опыт систематической сравнительной оценки систем машинного перевода для направления английский → русский. В будущем мы планируем построить новый тестовый корпус с более широкой жанровой палитрой. Мы постараемся дополнить оценку направлением перевода русский → английский. Мы надеемся привлечь больше участников, в том числе международных, и планируем подготовить «легкую версию» дорожки для студентов и молодых исследователей. Также мы рассмотрим проблему адаптации автоматических метрик оценки к русскоязычным данным. Такая метрика должна учитывать развитую русскую морфологию и свободный порядок слов, о чем говорилось выше. С этой целью мы планируем использовать данные ручной оценки, собранные в 2013 году.

Тестовый корпус, профессиональные переводы, переводы систем-участниц и данные ручной оценки будут доступны по адресу http://romip.ru/mteval/data/.

Благодарности

Мы хотели бы поблагодарить всех переводчиков и асессоров, а также Анну Цыганкову — за координацию проекта, Максима Губина и Марину Некрестьянову — за помощь в организации. Мы благодарны компаниям Яндекс и ABBYY, которые приняли активное участие в подготовке мероприятия и взяли на себя часть расходов, связанных с проведением оценки.

Литература

1. *Bogdan Babych, Anthony Hartley, Serge Sharoff, and Olga Mudraya.* 2007. Assisting translators in indirect lexical transfer. In Proc. of 45 ACL, pages 739–746, Prague.
2. *Satanjeev Banerjee and Alon Lavie.* 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan, June.
3. *Hanna Béchar, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef van Genabith.* 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In Proceedings of COLING'12, Mumbai.
4. *Igor Boguslavsky.* 1995. A bi-directional Russian-to-English machine translation system (ETAP-3). In Proceedings of the Machine Translation Summit V, Luxembourg.
5. *Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F Zaidan.* 2011. Findings of the 2011 workshop on statistical machine translation. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22–64. Association for Computational Linguistics.

6. *Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia.* 2012. Findings of the 2012 workshop on statistical machine translation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 10–51, Montréal, Canada, June.
7. *Marcelo Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stuker.* 2012. Overview of the iwslt 2012 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation (IWSLT), pages 12–34, Hong Kong, December.
8. *John Hutchins,* editor. 2000. Early years in machine translation: Memoirs and biographies of pioneers. John Benjamins, Amsterdam, Philadelphia. http://www.hutchinsweb.me.uk/EarlyYears-2000-TOC.htm.
9. *Philippe Langlais.* 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In Proceedings of Second international workshop on computational terminology (COMPUTERM 2002), pages 1–7, Taipei, Taiwan. http://acl.ldc.upenn.edu/W/W02/W02-1405.pdf.
10. *Sharon O'Brien.* 2011. Towards predicting post-editing productivity. Machine translation, 25(3):197–215.
11. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.* 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22 176 (W0109-022), IBM Thomas J. Watson Research Center.
12. *Serge Sharoff, Elena Umanskaya, and James Wilson.* 2013. A frequency dictionary of Russian: core vocabulary for learners. Routledge, London.
13. *Serge Sharoff.* 2010. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York.
14. *Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz.* 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 259–268, Athens, Greece, March.
15. *Joseph Turian, Luke Shen, and I. Dan Melamed.* 2003. Evaluation of machine translation and its evaluation. In Proceedings of Machine Translation Summit IX, New Orleans, LA, USA, September.
16. *John S. White, Theresa O'Connell, and Francis O'Mara.* 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In Proceedings of AMTA'94, pages 193–205.

КАК МОДЕЛИРОВАТЬ ПОНИМАНИЕ ЕСТЕСТВЕННОГО ЯЗЫКА: ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ СМЫСЛА

Богуславский И. М. (Igor.M.Boguslavsky@gmail.com),
ИППИ РАН, Москва, Россия;
Политехнический университет Мадрида, Мадрид, Испания

Диконов В. Г. (dikonov@iitp.ru),
Иомдин Л. Л. (iomdin@iitp.ru),
Тимошенко С. П. (timoshenko@iitp.ru),
ИППИ РАН, Москва, Россия

Ключевые слова: семантика, формальное представление смысла, онтология знаний, лексические функции, модель «Смысл — Текст», конверсивы

SEMANTIC REPRESENTATION FOR NL UNDERSTANDING

Boguslavsky I. M. (Igor.M.Boguslavsky@gmail.com)
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia;
Universidad Politécnica de Madrid, Madrid, Spain

Dikonov V. G. (dikonov@iitp.ru),
Iomdin L. L. (iomdin@iitp.ru),
Timoshenko S. P. (timoshenko@iitp.ru),
Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia

While mainstream semantic parsing mostly consists in word sense disambiguation, semantic role labeling and assigning WordNet/FrameNet categories, deeper NL understanding requires much more. It includes understanding of the meaning of words, extralinguistic knowledge and is based on a more intricately elaborated representation of this meaning than that provided by standard resources. For example, the semantic model should not only know that *ask for, implore* and *demand* belong to the same REQUEST frame. It should also formally represent the very idea of an incentive speech act (e.g. 'X tells Y that he wants him to do Z') and even the difference

between such request varieties as represented by the words listed. Our aim is to build a semantic analyzer supplied with this kind of semantic knowledge and capable of constructing semantic representations that convey this knowledge and can be used for inferences. However, before constructing a parser, one should define the target representation. The focus of this paper is to propose a semantic representation richer than usually considered. Since the depth of representation is an important decision in language modeling, the topic deserves a detailed discussion. Our paper demonstrates selected NL phenomena untreatable by state-of-the-art parsers and semantic representations proposed for them.

Key words: semantic representation, ontology, Meaning-Text Theory, formal sense representation, lexical functions, semantic parsing

1. Introduction

Mainstream shallow semantic parsing mostly includes named entity recognition, word sense disambiguation, semantic role labeling and assigning WordNet/FrameNet categories (e.g. Shi, Michalcea 2004). For example, the sentence

(1) *Messi scored a goal*

is typically assigned a semantic representation of the type

(2) *'A person named Messi is the agent of a GoalEvent'.*

Deeper NL understanding requires much more. In should include understanding of words' meaning, take into account available world knowledge and, ideally, involve as much inference as possible. In a sense, the level of text understanding is determined by the amount of inferences the cognitive agent can make. For sentence (1) such an understanding would include at least the following explicit data:

(3) (a) 'Messi is the captain of the Argentina national football team and a player of FC Barce-lona' [encyclopedic knowledge],

(b) 'Messi hit the ball, which resulted in the ball being located in the goal of the opposite team; as a result, the score of the team for which Messi was playing increased by 1' [linguistic knowledge: the explicit interpretation of the expression *score a goal*].

To obtain this level of understanding, one should have access to both linguistic and world knowledge and have an inference engine.

A project aiming at this type of semantic analysis has been initiated at IITP RAS. The ultimate goal of this project is to build a broad coverage semantic parser. In this paper, however, we will focus on one aspect of this work — the appropriate semantic

representation and its depth. As opposed to usual routine of discussion we will proceed bottom-up rather than top-down: instead of describing a semantic language and illustrating it by linguistic examples, we will depart from some non-trivial linguistic phenomena which are rarely (if ever) tackled by existing semantic parsers and show how they are represented in our language (Section 2). This material is worth discussing because the depth of representation is a crucial decision in language modeling, which should be taken irrespective of the way in which the parser processes the text. Logically, the target representation precedes the construction of the parser. Once we decide on the range of linguistic phenomena to be covered and devise a formal representation for them, we can choose a strategy of parser building. Our strategy is primarily rule-based. Obtaining semantic representation of the kind proposed below entirely by machine learning methods would require large annotated corpora, which are difficult and expensive to produce. A rule-based system may be viewed as a convenient step towards semi-automatic creation of such a corpus. On the other hand, we believe that knowledge-intensive methods have important advantages over data-driven ones, as far as the transparency and explanatory power is concerned.

Although the main contribution of the paper is theoretical, the feasibility of the representations proposed is confirmed by their being based on the available resources (a lexicon, an ontology, a rule-based engine) that we briefly describe in Section 3. In Section 4 we will present related work, and conclude in Section 5.

2. Selected issues of semantic analysis

2.1. Normalization and paraphrasing

Semantic analysis rules operate on Normalized Syntactic Structure and produce Basic Semantic Structures (SemS, see Section 3 below). One of the first tasks that should be done is the canonization. It includes restoring subjects of non-finite verbs *(I want to run → I want: I run),* processing of ellipsis, comparative constructions and the like. We will illustrate one canonization pattern: elimination of semantically void collocates. This operation is performed by means of a paraphrase generator based on Lexical Functions (LF) (Mel'čuk 1996; Apresjan, Cinman 2002). The paraphrase generator is a system of rules relying on a rich dictionary of lexical functions. In the semantic analyzer, the generator reduces sentences containing collocate LFs to the canonical form without these LFs. Some examples are:

(4) → (5), (6) → (7), (8) → (9).

(4) *John has respect for his teachers / John's teachers enjoy his respect / John treats his teachers with respect*

(5) *John respects his teachers*

(6) *The police gave the protesters an order to disperse / The protesters were ordered by the police to disperse / The protesters received an order from the police to disperse.*

(7) *The police ordered the protesters to disperse.*

(8) *The experts should submit / prepare / make / produce a report on chemical weapons.*

(9) *The experts should report on chemical weapons.*

Strictly speaking, the sentences in these pairs are not fully synonymous. However, the semantic representation we are striving at does not aim to account for all subtleties of meaning. The level of granularity of semantic representations should be determined by the task for which they are constructed. The immediate objective of our semantic representations is to support inference. For this aim, semantic differences that can be observed in these pairs are not relevant. Should an application require finer-grained representations, paraphrasing rules should be made more precise. An example of a subtler representation is given in the next section.

2.2. Semantic definitions of NL words and ontology concepts

Deep NL understanding requires much more elaborated meaning representation than that provided by standard resources. For example, semantic parsers based on FrameNet annotate verbs like *ask for, implore* and *demand* by relating them to the same REQUEST frame. It is true that all the three verbs have the same set of roles. The generalized frame REQUEST allows us to capture this similarity, but we also want to preserve the knowledge about their difference. First, one should make explicit the very idea of the incentive speech act (e.g. 'X tells Y that he wants him to do Z'). This should be made in a formal language so that it could be used for inferences. Second, since our ultimate aim is to model natural language as fully as possible, it is desirable to account for the semantic difference between the varieties of this speech act. There exist many NL speech act types in which the agent informs the addressee that he wants the latter to do something.

Below, we give definitions of three of them: *ask* (as in *He asked to open the window), implore* (as in *They implored her to help)* and *demand* (as in *She demanded an explanation).* Roughly, the difference between *ask* and *demand* is that the one who is asking does not think that the addressee is obliged to fulfill the request, while the one who is demanding assumes that the addressee must do it. Imploring adds to asking the idea that fulfilling the request is very important for the agent so in persuading the addressee to do it he tries to affect his feelings. In the definitions, variables are marked with the ? sign. For brevity, the ontological class to which the variable belongs is encoded by the name of the variable.

(10) *ask for (?Agent1,?Agent2,?Action)* ['?Agent1 tells ?Agent2 that he wants him to do ?Action; ?Agent1 does not think that ?Agent2 must do ?Action']
```
hasAgent(Tell,?Agent1)
hasRecipient(Tell,?Agent2)
```

```
hasObject(Tell,Want)
hasSubject(Want,?Agent1)
hasObject(Want,?Action)
hasAgent(?Action,?Agent2)
hasScope(Negation,Opinion)
hasSubject(Opinion,?Agent1)
hasObject(Opinion,?Action)
hasScope(MustModality,?Action)
```

(11) *implore (?Agent1,?Agent2,?Action)* [='?Agent1 asks ?Agent2 to do ?Action; it is very important for ?Agent1 that ?Agent2 realizes ?Action; ?Agent1 tries to affect the feelings of ?Agent2']

SemS of (11) consists of the SemS of (10) plus the following:

```
hasSubject(Important,?Action)
hasObject(Important,?Agent1)
hasSubject(Degree,Important)
hasValue(Degree,high)
hasAgent(Affect,?Agent1)
hasObject(Affect,Feeling)
hasSubject(Feeling,?Agent2)
```

(12) *demand (?Agent1,?Agent2,?Action)* [='?Agent1 tells ?Agent2 that he wants him to do ?Action; ?Agent1 thinks that ?Agent2 must do ?Action']

```
hasAgent(Tell,?Agent1)
hasRecipient(Tell,?Agent2)
hasObject(Tell,Want)
hasSubject(Want,?Agent1)
hasObject(Want,?Action)
hasAgent(?Action,?Agent2)
hasSubject(Opinion,?Agent1)
hasObject(Opinion,?Action)
hasScope(MustModality,?Action)
hasAgent(?Action,?Agent2)
```

2.3. Converse terms

Natural languages have hundreds of converse terms, i.e. pairs of words that denote the same situation but differ in the syntactic status of their arguments. Obvious examples are *husband — wife, buy — sell, to the right of — to the left of, more — less, better — worse,* etc. Although these words are not synonyms, if we swap positions of arguments we obtain equivalent assertions:

(13) *John is Mary's husband = Mary is John's wife.*

(14) *John bought a house from Mary = Mary sold a house to John.*

(15) *The table is to the right of the window = The window is to the left of the table.*

(16) *John likes physics more than geography = John likes geography less than physics.*

Since converse terms refer to the same situation, it is sufficient for a semantic language and ontologies to contain only one term of the pair. In our semantic language, we have only one correlate for the 'more'/'less' pair — concept MORE. In representing this meaning, we differ from some other approaches (such as e.g. (Nirenburg, Raskin 2004), which treat 'more' as a binary relation: A>B. Our MORE concept has three arguments: A — "what is more?", B — "more than what?", C — "by how much is A more than B?". In sentence (17) the arguments of MORE are: A=John's height, B=Bill's height, C=3 cm.

(17) *John is 3 cm taller than Bill.*

On the other hand, one can opt for having both members of the converse pair. For instance, we represent *husband* and *wife* by different concepts, because these social roles are bound by different conventions and stereotypes which have to be described in the ontology.

An interesting case of converse relations, which as far as we are aware was first introduced in (Boguslavsky 2009), is the relationship between *all* and *only*.

(18) *Here are all my documents*
("for any document x of mine, it is true that x is here").

(19) *Here are only my documents*
("for any *x* that is here, it is true that *x* is my document").

This allows us to have only one semantic unit — a two-place predicate `All`, which covers both *all* and *only*. Here are semantic structures for sentences (22) and (23):

(20) *All the children who guessed the riddle got a prize.*

(21) *Only the children who guessed the riddle got a prize.*

(20a)
```
hasElements(Set,Child)
hasAgent(Guess,Child)
hasObject(Guess,Riddle)
hasAgent(Get,Set)
hasObject(Get,Prize)
hasSubject(All,Set)
hasObject(All,Get)
```

(21a)
```
hasElements(Set,Child)
hasAgent(Guess,Child)
hasObject(Guess,Riddle)
hasAgent(Get,Set)
hasObject(Get,Prize)
hasSubject(All,Get)
hasObject(All,Set)
```

2.4. Evaluation of objects

Evaluation of objects and events plays an enormous role in our life, everyday behavior and common sense reasoning. Therefore the world knowledge modeled by the ontology should contain manifold information on what is good and, bad and for whom. For many situations, we are aware that they are either beneficial or detrimental to the interests of some of their participants. For example, if somebody dies, is sick, late for an appointment, gets ruined, receives a rebuke, or fails an exam, by default this is bad for him. If, on the other hand, he recovers from an illness, gets an award, is promoted or attains his aim, then, again by default, it is beneficial for him. Some situations are estimated differently from the point of view of their different participants. For example, a victory (in a conflict, debate, sports competition, etc.) is beneficial for the winner and adverse for the loser. We will demonstrate that this kind of information can play a role in text understanding. Then we will show how it is incorporated in our Ontology and used for semantic analysis.

Consider sentence (22) and its two possible continuations — (23) and (24).

(22) *In the first tour FC Spartak overwhelmed FC Dynamo.*

(23) *In the second tour FC Zenith suffered the same fate.*

(24) *In the second tour FC Zenith managed to achieve the same thing.*

Both (23) and (24) contain the anaphoric expression *the same* that refers to sentence (22). In both cases, a situation is described that is similar to (22), the only difference being that one of the clubs is replaced with Zenith. In (23) an analogy is drawn between Zenith and Dynamo, and in (24) between Zenith and Spartak. In other words, (23) is unambiguously understood as 'Spartak overwhelmed Zenith', while (24) means that *'Zenith overwhelmed Dynamo'.* It is noteworthy that even though neither (23) nor (24) explicitly specifies the opponent of Zenith, it is "calculated" from the evaluation semantics.

To be able to draw these conclusions, the system should dispose of the following knowledge:

(a) "P is fate suffered by X" implies that P is not in the interests of X;
(a) "X managed to achieve P" implies that P was among X's aims and P is beneficial for X;

(a) "victory of X over Y" is beneficial for X but not for Y.

This knowledge is incorporated into the system as follows:

- The Ontology contains an `Evaluation` concept, which has 4 slots: the agent of the evaluation (`hasAgent`), the object or event under evaluation (`hasObject`), the value of the evaluation (`hasValue`) — good or bad and the beneficiary, i.e. someone for whom the object or event is beneficial or adverse (`hasBeneficiary`).
- This concept is introduced into the description of the concepts which include a default evaluation (cf. examples above). The `WinEvent` concept, which has slots for the winner (`hasWinner`) and for the loser (`hasLoser`) and which covers both a victory and a defeat, is assigned the following properties, among others:

```
hasWinner(WinEvent,?SportAgent1)
hasLoser(WinEvent,?SportAgent2)
hasObject(Evaluation-01,WinEvent)
hasValue(Evaluation-01,good)
hasExperiencer(Evaluation-01,?SportAgent1)
hasObject(Evaluation-02,WinEvent)
hasValue(Evaluation-02,bad)
hasExperiencer(Evaluation-02,?SportAgent2)
```

- A reference to evaluation is included into semantic rules that interpret natural language evaluating expressions. X *suffered the fate of P* contains the component "P is estimated to be bad for X". In our semantic language it is represented as follows:

```
hasObject(Evaluation,P)
hasValue(Evaluation,bad)
hasBeneficiary(Evaluation,X)
```

- Expressions like *X succeeded in / achieved P* include in their definition a reference to P being the aim of X, which in its turn implies that P is beneficial for X:

```
hasObject(Evaluation,P)
hasValue(Evaluation,good)
hasBeneficiary(Evaluation,X)
```

Now, let us see how this knowledge helps interpret sentences (23) and (24). As mentioned above, proposition (22) serves as the antecedent of '*the same',* so theoretically, it can be introduced into the SemS of both (23) and (24) in two different ways:

(24a)
```
hasWinner(WinEvent,Zenith)
hasLoser(WinEvent,Dynamo)
```

(meaning that Zenith beat Dynamo like Spartak beat Dynamo) or

(23a)
```
hasWinner(WinEvent,Spartak)
hasLoser(WinEvent,Zenith)
```

(meaning that Zenith lost to Spartak like Dynamo lost to Spartak).

However, taking into account that the meaning of *Zenith suffers a fate* assigns to Zenith the role of the beneficiary of a negative evaluation, while in the *WinEvent* it is the winner who benefits, version (24a) should be rejected for sentence (23). In a similar way, (23a) is rejected for (24).

3. Semantic analysis in ETAP

The cases analyzed above make part of a small corpus manually annotated with semantic structures. The corpus comprises several hundred sentences which are partly extracted from the articles on football published at various sports portals and partly composed by ourselves. This corpus is used for developing a rule-based semantic analyzer capable of building gold standard structures. As of now, more than a hundred sentences have been processed by the analyzer and assigned correct semantic structures. Once a rule-based analyzer is constructed, it will open the possibility to considerably augment a corpus, which could then be used for refining and evaluating the analyzer, as well as for developing other semantic parsers.

Our analyzer will be described in detail at a later stage when more experiments have been conducted and more data accumulated. Now we will only give a brief sketch of its architecture and resources used.

The analyzer is a new module of the ETAP-3 linguistic processor (see e.g. Apresjan et al. 2003). Before being sent to semantic analysis, the text is subjected to morphological analysis, dependency parsing, and normalization. The semantic analysis consists in two major steps. First, Normalized Syntactic Structures of all sentences are individually transformed into Backbone Semantic Structures (BSemS). At this canonization stage, missing arguments are restored and semantically void collocates are eliminated (see Section 2.1 above). Then all meaningful words are replaced by their definitions. Second, BSemSs are enriched with the world knowledge and the contextual knowledge from the previous text and thus converted to Enhanced Semantic Structures.

Linguistic information is contained in two kinds of resources: the combinatorial dictionary and several sets of rules. World knowledge is contained in the Ontology, while contextual knowledge is stored in the Fact Repository. The ontology we constructed for the analyzer has two sources. We compiled a small domain ontology of football, reusing the existing football ontologies (e.g. http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml). Then we merged it with a general ontology developed on the basis of SUMO (http://www.ontologyportal.org/), which we partially restructured and complemented with a large set of properties. In our analyzer the ontology plays a two-fold role. On the one hand, it is a source of structured information about the world. It is composed of a hierarchy of concepts and instances supplied with properties. Many concepts belong to various classes at a time, so that they inherit properties from multiple sources. On the other hand, the ontology serves as a metalanguage for semantic representation. It is an inventory of semantic units that make up semantic structures.

4. Related work

A popular resource for developing shallow semantic parsers is FrameNet, mentioned above in Section 2.2. Semantic definitions of frames are intended for humans and are not written in a formal language. Therefore, they structures produced by FrameNet-based semantic parsers cannot be used for inference.

There are several directions in which semantic processing relying on ontologies is currently carried out. Our approach to semantic analysis is closely related to the OntoSem approach, with which we share several important ideas, although our linguistic framework is substantially different (Nirenburg, Raskin 2004), (Akshay Java et al. 2006), (Akshay Java et al. 2007), (Raskin, Taylor 2010), (Raskin et al. 2010).

Still another linguistic model underlies a series of papers on FuncGram — an advanced semantic Knowledge Base rooted in the Lexical-Constructional Model (Mairal Usón, Periñán-Pascual 2009, Periñán-Pascual, Arcas-Túnez 2010 a, b, Mairal Usón 2010).

Semantic processing based on OWL-implemented ontologies and Descriptive Logic does not allow accounting for exceptions. Interesting work is being done in order to incorporate common sense reasoning, which is inseparable from defeasible statements (Fahlman 2011), (Carlson et al. 2012). In (Bouayad-Agha et al. 2012a), (Bouayad-Agha et al. 2012b) a two-layer ontology is used for NL generation.

Modern QA systems use ontologies as the core knowledge component. They are often used to annotate original data obtained from the web sites and other sources of unstructured or loosely structured texts. The annotated data is stored in the databases and retrieved to answer the user's questions (Shiyan Ou et al. 2008). (Fernandez et al. 2011), (Cardoso et al. 2010) describe semantically-aware QA working on structured data modeled by an ontology.

There is much research on semantic parsing within the machine learning paradigm. Interesting results have been obtained in supervised and unsupervised semantic parsing in (Ge and Mooney, 2005), (Poon and Domingos, 2009), (Titov and Klementiev, 2011), (Clarke et al. 2010), (Liang et al. 2011). A combination of machine learning and rule-based approaches is used for semantic processing in (Moldovan et al., 2010). However, for the kind of structure we are interested in, no annotated corpora are available.

5. Conclusion

Deep understanding of NL requires more expressive semantic representation than is currently used in most state-of-the-art semantic parsers. It should be equally well-suited for expressing lexical meanings and world knowledge. Such a representation can be built on the basis of RDF-style subject-predicate-object triples. We showed a variety of NL phenomena that are conveniently expressed in such a language. To the best of our knowledge, some of this material is introduced in the computational semantics area for the first time. This is true for semantic definitions of many concepts in 2.2. Our approach to the evaluation topic is also new: it differs from the approaches used in the sentiment analysis domain which are prevalent

today. We showed how lexical meanings can be decomposed, semantically void collocates can be identified and eliminated not affecting the argument structure, converse terms can be properly processed, general semantics modifiers can be contextually interpreted, lexical semantics (including evaluation) can be used in hard cases of anaphora resolution. The semantic language illustrated in this paper is used in the semantic analyzer currently under development within the multifunctional ETAP linguistic processor.

References

1. *Akshay Java* et al. 2006 — Akshay Java, Tim Finin and Sergei Nirenburg. Text understanding agents and the Semantic Web. Proceedings of the 39th Hawaii International Conference on System Sciences, Vol.3, pp. 62b, Kauai HI, 2006.
2. *Akshay Java* et al. 2007 — Akshay Java, Sergei Nirenburg, Marjorie McShane, Timothy Finin, Jesse English, Anupam Joshi. (2007) Using a Natural Language Understanding System to Generate Semantic Web Content. International Journal on Semantic Web and Information Systems, 3(4), pp. 50–74..
3. *Apresjan, Cinman* 2002 — Apresjan, Ju. D., Cinman, L. L. (2002) The Formal Model for Sentence Paraphrasing in NLP systems [Formal'naja model' perifrazirovanija predlozenij dlja sistem pererabotki tekstov na estestvennyx jazykax]. Russian Language in The Scientific Context [Russkij jazyk v naucnom osvescenii]. No. 4, pp. 102–146.
4. *Apresjan* et al. 2003 — Apresjan, Jury, I. Boguslavsky, L. Iomdin, A. Lazursky, V. Sannikov, V. Sizov, L. Tsinman. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. MTT 2003, First International Conference on Meaning — Text Theory (June 16–18, 2003). Paris: Ecole Normale Superieure, pp. 279–288.
5. *Apresjan* et al. 2009 — Apresjan Ju., I.Boguslavsky, L.Iomdin, L.Cinman, S.Timoshenko. Semantic Paraphrasing for Information Retrieval and Extraction. Flexible Query Answering Systems. 8th International Conference, FQAS 2009, Roskilde, Denmark, October 26–28, 2009. Lecture Notes in Artificial Intelligence, Vol. 5822, pp. 512–523.
6. *Boguslavsky* 2009 — Boguslavsky, Igor. Enlarging the Diversity of Valency Instantiation Patterns and Its Implications. Lecture Notes in Artificial Intelligence. Logic, Language, and Computation: 7th International Tbilisi Symposium on Logic, Language, and Computation, TbiLLC 2007, Tbilisi, Georgia, October 1–5, 2007. Bosch, P; Gabelaia, D; Lang, J (Eds.) Springer-Verlag Berlin, Heidelberg, 2009, p. 206–220.
7. *Bouayad-Agha* et al. 2012a — Bouayad-Agha N., Casamayor G., Mille S., Wanner L. Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges. ACM Transactions on Speech and Language Processing, 2012. 9(2).
8. *Bouayad-Agha* et al. 2012b — Bouayad-Agha N., Casamayor G., Mille S., Rospocher M., Serafini L., Wanner L. From Ontology to NL Generation of Multilingual User-Oriented Environmental Reports. Proceedings of NLDB 2012: 17th

International Conference on Applications of Natural Language Processing to Information Systems. Groningen, 2012.
9. *Cardoso* et al. 2010 — Cardoso, N., Dornescu I., Hartrumpf S. and Leveling J. (2010) Revamping question answering with a semantic approach over world knowledge. CLEF Labs 2010, Multiple Language Question Answering, 2010 (MLQA10), Padua, Italy.
10. *Carlson* et al. 2012 — Carlson T., Van Lifferinge S., Holt E., Smith R., Covington M., Potter W. Application of Defeasible Domain-Specific Knowledge to the Description of Gothic Cathedrals in the ARC Project. Proceedings of the 2012 International Conference on Artificial Intelligence. Hamid R. Arabnia et al (eds.) vol. 1, p. 172–178.
11. *Clarke* et al. 2010 — Clarke, J., D. Goldwasser, M. Chang and D. Roth. (2010). Driving Semantic Parsing from the World's Response. Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).
12. *Dornescu* 2009 — Iustin Dornescu. (2009). EQUAL: Encylopaedic Question Answering for Lists. Working notes for the CLEF 2009 Workshop. Corfu, Greece.
13. *Fahlman* 2011 — Scott E. Fahlman. (2011) Using Scone's multiple-context mechanism to emulate human-like reasoning. Proceedings of the AAAI Fall Symposium on Advances in Cognitive Systems, 2011. pp. 98–105.
14. *Ferrandez* et al. 2011 — Oscar Ferrandez, Christian Spurk, Milen Kouylekov, Iustin Dornescu, Sergio Ferrandez, Matteo Negri, Ruben Izquierdo, David Tomas, Constantin Ora-san, Guenter Neumann, Bernardo Magnini, Jose Luis Vicedo. (2011). The QALL-ME Framework: A specifiable-domain multilingual Question Answering architecture. Web Semantics: Science, Services and Agents on the World Wide Web. Vol. 9, Issue 2, July 2011, p. 137–145.
15. *Ge and Mooney* 2005 — Ruifang Ge, Raymond J. Mooney. (2005). A Statistical Semantic Parser that Integrates Syntax and Semantics. Proceedings of the Ninth Conference on Computational Natural Language Learning. Ann Arbor, MI, pp. 9–16, June 2005.
16. *Liang* et al. 2011 — Percy Liang, Michael Jordan, Dan Klein. (2011). Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1, p. 590–599.
17. *Mairal Uson, Perinan-Pascual* 2009 — Mairal Uson, R. y J. C. Perinan-Pascual. "The anatomy of the lexicon component within the framework of a conceptual knowledge base". Revista Española de Lingüística Aplicada 22 (2009), pp. 217–244.
18. *Mel'cuk* 1996 — Mel'cuk, I. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. Wanner, L. (ed.) Lexical Functions in Lexicography and Natural Language Processing,Amsterdam, Philadelphia, pp. 37–102.
19. *Moldovan* et al., 2010 — Moldovan, D., Tatu, M., Clark, Ch. (2010). Role of Semantics in Question Answering. In: Phillip C.-Y. Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (Eds.) Semantic Computing, pp. 373–420.
20. *Nirenburg, Raskin* 2004 — Nirenburg, S., and Raskin, V. (2004). Ontological Semantics. The MIT Press. Cambridge, Mass., London, England.

21. *Periñán-Pascual, Acas-Tunez* 2010a — Perinan-Pascual, J. C. and F. Arcas-Tunez. (2010) Ontological Commitments in FungramKB.. Procesamiento del Lenguaje Natural 44.
22. *Periñán-Pascual, Arcas-Tunez* 2010b — Perinan-Pascual, J. C. and F. Arcas-Tunez. (2010) The architecture of unGramKB. Proceedings of ELRA Conference. Malta.
23. *Periñán-Pascual, Mairal Usón* 2010 — Perinan-Pascual, J. C. and R. Mairal Uson. (2010) "La Gramatica de COREL: un lenguaje de representation conceptual". Onomazein 21. Universidad de Chile.
24. *Poon, Domingos* 2009 — Poon, H., & Domingos, P. Unsupervised semantic parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09 (p. 1).
25. *Raskin, Taylor* 2010 — Victor Raskin, Julia Taylor. Fuzzy Ontology for Natural Language. 29th International Conference of the North American Fuzzy Information Processing Society, Toronto, Canada, July 2010.
26. *Raskin* et al. 2010 — V. Raskin, C.F.Hempelmann, J. Taylor. Application-guided Ontological Engineering. International Conference on Artificial Intelligence, Las Vegas, NE, July 2010.
27. *Shi and* 2004 — Lei Shi and Rada Mihalcea. Open Text Semantic Parsing Using FrameNet and WordNet. Proceedings HLT-NAACL — Demonstrations '04 Demonstration Papers at HLT-NAACL 2004. p. 19–22
28. *Shiyan Ou* et al 2008 — Shiyan Ou, Victor Pekar, Constantin Orasan, Christian Spurk, Matteo Negri. Development and Alignment of a Domain-Specific Ontology for Question Answering. Proceedings of LREC 2008. 26 May — 1 June 2008, Morocco, Marrakech, pp. 2221–2228.
29. *Titov, Klementiev* 2011 — I. Titov, A. Klementiev. A Bayesian Model for Unsupervised Semantic Parsing. Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1, USA, Oregon,.Portland. pp. 1445–1455.

ОПЫТ НАСТРОЙКИ СИСТЕМЫ АВТОМАТИЗИРОВАННОГО ПЕРЕВОДА ПОЛЬЗОВАТЕЛЬСКОГО КОНТЕНТА

Евдокимов Л. В. (Leonid.Evdokimov@promt.ru),
Молчанов А. П. (Alexander.Molchanov@promt.ru)

ООО «ПРОМТ», Санкт-Петербург, Россия

В данной статье описывается опыт компании PROMT по настройке и реализации системы автоматизированного перевода PROMT Deep-Hybrid для интерактивной обработки текстовой информации, представленной на сайте, который представляет собой крупный интернет-ресурс, посвященный туризму и путешествиям. Целью работы было создание решения для перевода пользовательского контента, состоящего из текстов отзывов об отелях, ресторанах и других составляющих современного туристического сектора.

Ключевые слова: машинный перевод, пользовательский контент, автоматизированный перевод, гибридная технология перевода

CREATING AN AUTOMATED SYSTEM FOR TRANSLATION OF USER-GENERATED CONTENT

Evdokimov L. V. (Leonid.Evdokimov@promt.ru),
Molchanov A. P. (Alexander.Molchanov@promt.ru)

PROMT Ltd., Saint-Petersburg, Russia

This paper describes fast implementation of a hybrid automated translation system for processing user-generated content. We report on engine customization for TripAdvisor, the world's largest travel website. Due to the growing potential of the Russian travel market, TripAdvisor created the Russian version of its website and decided to translate all English reviews into Russian. PROMT, a leading provider of industrial MT solutions, was selected as MT vendor for the English-Russian language pair. According to the client's request we had to perform customization within a short period.

All input data represent user-generated content, so we faced several problems while building a large-scale, robust, high-quality engine. We decided to create a solution based on a hybrid machine translation system for the hybrid approach makes possible fast and efficient customization of a translation system with little or none in-domain data.

We automatically crawled a large web-based Russian text corpus of tourist reviews to build a statistical language model for our hybrid translation

system. We analyzed a batch of tourist reviews in English provided by TripAdvisor, created a number of dictionaries, a translation memory and defined translation rules for user-generated content. To handle the problem of various typos and misspellings we added most frequent misspelled words and phrases to the created dictionaries.

We experimented on a test set of tourist reviews in English provided by TripAdvisor. We report on improvements over our baseline system output both by automatic evaluation metrics and linguistic expertise.

Keywords: machine translation, user-generated content, automated translation, hybrid technology

1. Introduction

The fast evolution of computers and the rapid growth of the Internet since the late 1990s made it easier for people to upload, store and share information on the web. Forums, chats and other web-based informational resources led to the emergence of large amounts of the so called 'user-generated content'. User generated content (UGC) is material on websites, and occasionally other media sources, that is produced by users of websites (who are generally amateurs as opposed to professional editors, copywriters etc). In our case the content consists of tourist reviews produced by the users of the tripadvisor.com website.

TripAdvisor is world's largest travel web-based resource. The content is available in 21 languages for 30 countries. Most reviews are presented in English. At the same time, millions of users want to read the reviews in their native language. Human translation cannot be efficient for processing large amounts of UGC. Taking into account the fast growth of Russian travel market, TripAdvisor wanted an efficient automated translation solution for processing UGC.

2. Related Work

Regardless of the growing demand for automated translation of UGC little attention is paid to this topic in the field of machine translation research.

[Flournoy and Callison-Burch, 2000] discuss the possibility of creating a high-quality commercially successful application for real-time automated translation of chat content. The authors note that UGC is characterized by specific repeated colloquial words and phrases and lots of grammar errors. The main task of an MT system is to convey the meaning of the source text, whereas the translation quality is of secondary importance.

[Flournoy and Rueppel, 2010] investigate the development of an automated translation system for Adobe. The authors define three types of UGC:

- user e-mails;
- bug reports and product reviews;
- messages from user forums.

According to the authors, an efficient MT system for processing UGC should have the following features:

- ability to translate large amounts of texts in real time;
- ability to convey source text meaning;
- reliability and robustness (taking into account large volumes and low quality of input data).

[Banjeree et al., 2011] and [Banjeree et al., 2012] present the case-studies of customization of an automated MT system for processing UGC from the Symantec company forum. Authors observe a lot of grammar mistakes and a large number of colloquial words and phrases in the analyzed texts.

[Jie Jiang et al., 2012] report on the customization of an automated translation system for user messages in a multilingual social network. The authors face the following problems:

- a lot of the content is produced by non-native speakers, therefore this content contains many grammatical and syntactic errors;
- the content produced by native speakers contains grammatical and syntactic errors because 1) either the author enters the text too fast and so makes typographical errors, or 2) the author deliberately departs from spelling norms to bring about some linguistic effect.

The reliability and robustness are basic requirements for an automated machine translation system for processing UGC. An MT system for processing UGC should be 1) thoroughly customized for this specific type of content and 2) be able to translate large amounts of text in real time.

3. Aim and Objectives

The main challenge was to achieve high quality of translation. Since manual editing of each review was impossible, the website functionality required a high quality automated translation system that does not require human post-editing. About 80,000 reviews are added to the website weekly, so TripAdvisor required a technically accurate solution for processing large volumes of text. Another client's requirement was to translate the existing content (over 10 million reviews) within a short period. Due to the huge amount of data human post-editing of every single review was impossible. At the same time, UGC is a challenge for MT, since such texts are highly informal and typically contain a significant number of spelling, stylistic and punctuation errors that affect the MT results.

Another important client's requirement was an efficient quality estimation system integrated into the final MT solution. As TripAdvisor wanted to publish high-quality translations only, PROMT had to design an automated quality estimation system with a quality threshold.

The translation results had to contain clear and understandable content. Translation had to meet certain quality criteria, and as manual evaluation of the whole translation volume was impossible, the MT solution had to provide an automatic scoring mechanism for the evaluation of the translated texts.

The tight deadline for developing MT system was another crucial demand made by TripAdvisor.

Website developers wanted a cloud-based server MT solution, that's why we decided to develop a hybrid translation solution based on the PROMT DeepHybrid system (see [Molchanov, 2012]).

4. Statistical and Linguistic Analysis of the Data provided by TripAdvisor

4.1. Initial Data

TripAdvisor provided PROMT with the following data for engine customization:

- TripAdvisor English-Russian glossary (505 entries);
- English-Russian TripAdvisor TMs (~100,000 entries);
- English monolingual text corpus of hotel reviews (~1.2 billion words).

4.2. Domain-Specific Dictionaries

The TripAdvisor English-Russian glossary was converted into a dictionary of the PROMT internal format. We also extracted the most frequent terms and phrases from the English hotel review corpus. We analyzed the translations of these entries and made the necessary corrections and additions to the TripAdvisor dictionary and the baseline PROMT Travel dictionary.

Due to the large amount of misspellings and typos in the text of reviews we decided to create a dictionary with incorrect spelling of frequent English words, e.g.

(1) *couldnt, did'nt, experieince,*

so that the translation system could treat them as known words.

We also created the PROMT TripAdvisor Backgroud dictionary containing frequent travel-related phrases. The dictionaries were then icorporated into the translation system according to their priority: 1) TripAdvisor dictionary (highest priority); 2) TripAdvisor Backgroud dictionary; 3) Travel dictionary; 4) PROMT General dictionary (lowest priority). The priority works as follows: if the word or phrase is missing in the dictionary with the highest priority, the system tracks it in the dictionary with next priority etc.

4.3. Translation Memory

We made a thorough analysis of the English-Russian translation memory provided by TripAdvisor. We decided not to use it for three main reasons: 1) many segments were not domain-relevant; 2) many of them contained lots of different errors

(untranslated and incorrectly translated sentences, segments containing no alphabetic characters etc.); 3) many segments were of adequate quality but not informative for the baseline PROMT system, for example, named entities and geographic names:

(2) *Reno-PropertyOpen-NoDates Salute the white baroque towers of St. Fernando de Noronha and Atol das Rocas Reserves*

Due to the tight schedule we selected a random development set (approximately 10 percent) from the English hotel review corpus provided by TripAdvisor. We used this development set to build a list containing the most frequent in-domain sentences, e.g.

(3) *Highly recommended! The staff was very friendly and helpful.*

etc. These sentences (15K) were processed the following way: 1) the sentences were translated with the baseline PROMT system; 2) the translations were analyzed by our linguists. According to linguistic expertise only 8 % (1200 sentences) contained major syntactic and stylistic errors. These sentences were manually post-edited and integrated into the translation system as a translation memory.

4.4. Target Language Model

A target language model is normally built on the in-domain target texts. In our case, there was no in-domain text corpus in Russian, so we had to create it. We crawled and processed about 27,000 user reviews (80 million words) from different Russian websites dedicated to travelling. These texts were used to build the target language model. The model was integrated into the translation system.

A language model is a set of n-grams (word sequences of n-length) and their statistical characteristics. The rule-based system may have several translation options for some words and phrases. The language model is a component of the PROMT DeepHybrid system. It is used to score the translation candidates generated by the rule-based component and select the best one according to perplexity score. Perplexity (PPL) is inversely proportional to probability and is calculated for every translation candidate. The lower the PPL is, the better the translation candidate fits the language model.

We called the language model built on the Russian reviews corpus the BigTripAdvisor Language Model. It was integrated into the translation system for TripAdvisor.

4.5. Quality Estimation System

According to the client's requirements, our automated translation system had to be equipped with a quality estimation component. Quality estimation (QE) systems are used to estimate machine translation output quality at run-time. In our case, we had to select the high-quality translated reviews suitable for publishing on the website without human post-editing and reject the low quality ones.

First of all, we had to choose a confidence metric which would be the basic element of our QE system. Due to the tight schedule, we decided to create a simple metric based on PPL. Our experts performed the quality evaluation of 1000 sentences with different PPL scores. The results of this experiment showed that there is a sufficient correlation between the translation quality and the PPL scores (see Figure 1).

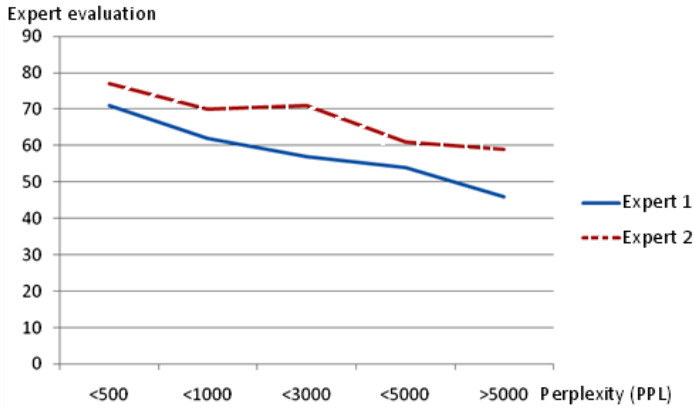


Fig. 1. Correlation between expert evaluation and PPL scores

QE systems normally operate on the sentence level. According to the client's request, our QE system had to estimate the entire text of the reviews. The average review length for the TripAdvisor website is approximately 100 words or three to five sentences. We decided to use the arithmetic mean of the PPL scores for separate sentences of reviews.

According to another request from TripAdvisor, the QE system had to be scaled from 1 to 5 with the accuracy of 0.1. Low-quality translations with PPL over 10,000 received the score equal to 1, high quality translations with PPL under 10 received the score equal to 5. The scaling formula is presented in Figure 2 below.

$$M = \begin{cases} 5, & PPL < 10 \\ \dfrac{4 \cdot (4 - \log_{10} PPL)}{3}, & 10 \leqslant PPL \leqslant 10^4 \\ 1, & PPL > 10^4 \end{cases}$$

Fig. 2. Scaling the PPL scores

We scored the translations of all reviews from the English monolingual corpus provided by TripAdvisor. The number of translations with scores 1 and 5 was less than 0.1 %. The distribution of the QE metric scores is presented in Figure 3.

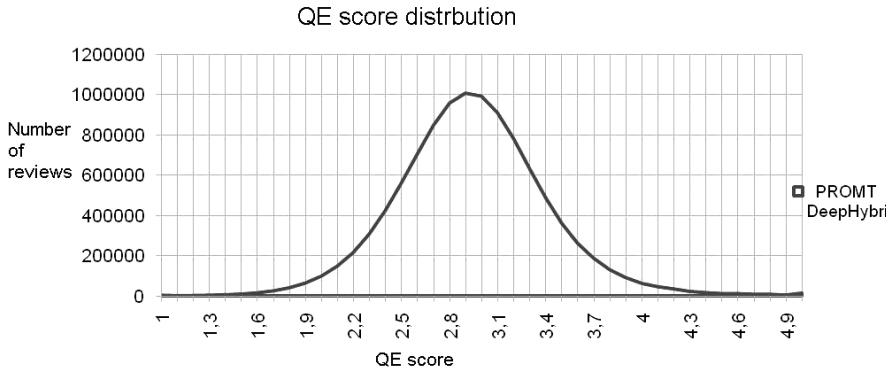


Fig. 3. Quality estimation score distribution

4.6. Deliverables

We developed a reliable, robust, scalable server-based translation solution. The solution was based on the PROMT DeepHybrid translation engine and included the following components:

- English-Russian Dictionaries: 1) PROMT TripAdvisor dictionary containing client-specific terms (approximately 5,600 entries); 2) PROMT TripAdvisor Background dictionary containing domain-relevant terminology (approximately 27,600 entries); PROMT TripAdvisor Geography background dictionary containing geographic names (approximately 48,200 entries).
- Target language model built on the text corpus of reviews in Russian.
- QE system.

5. Translation Quality Evaluation

Tripadvisor provided a parallel corpus (approximately 70K words) of the English reviews and their translations with human post-editing. We used this corpus to evaluate the translation quality. The English revews were translated with: 1) PROMT baseline system; 2) PROMT baseline system with the TripAdvisor dictionaries; 3) fully customized PROMT DeepHybrid system with all components. The BLEU scores are presented in Table 1.

Table 1. BLEU scores and the percentage of unknown words for various PROMT translation system configurations

System	BLEU score	percentage of unknown words
PROMT baseline system	17.12	2.56%
PROMT baseline system + TripAdvisor dictionaries	19.42	2.19%
PROMT DeepHybrid system (PROMT baseline system + TripAdvisor dictionaries + Language model)	20.13	2.16%

Our experts performed lingustic analysis of the PROMT baseline system and the PROMT DeepHybrid system output. 3,291 sentences (78% of the test set) of the PROMT DeepHybrid system output contained changes compared to the PROMT baseline system output. Our experts compared 100 random RBMT and DeepHybrid translations in terms of improvements and degradations. The results showed that the DeepHybrid engine outperforms the RBMT engine according to human evaluation. The experts observed 49 improvements and 9 degradations for the DeepHybrid system output compared to the baseline system output. 42 translations were classified as equivalent.

Examples of translation quality improvements are presented in Table 2. The table also includes the translations of the Google online translation service.

Table 2. Examples of translation quality improvements

№	Source sentence	PROMT Baseline System	PROMT DeepHybrid system	google.translate
1	A big thumbs up to the Kiydan family	Большие большие пальцы до семьи Kiydan	Оценка «отлично» семье Киидэн	Большие пальцы в семье Kiydan
2	Can't wait to go back!!	Не может ждать, чтобы возвратиться!!	Не терпится вернуться снова!!	Не может ждать, чтобы вернуться!
3	The **brakfast** was awsome.	brakfast был awsome.	Завтрак был потрясающим.	Завтраком было потрясающим.
4	The food and **resturant** was very good	Еда и resturant были очень хороши	Еда и ресторан были очень хороши	Еда и ресторан был очень хорош
5	At least the staff were **pleasent**!	По крайней мере, сотрудники были pleasent!	По крайней мере, персонал был приятным!	По крайней мере, сотрудники были **приятно!**
6	Dinner at the hotel was quite expensive and we preferred to eat out, however we ate at the hotel one day when the **menue** included lobster.	Обед в отеле был довольно дорог, и мы предпочли идти куда-нибудь поесть, однако мы поели в отеле однажды, когда menue включал омара.	Ужин в отеле был довольно дорогим, и мы предпочли идти куда-нибудь поесть, однако мы поели в отеле однажды, когда меню включало омара.	Ужин в отеле был довольно дорогим, и мы предпочли пойти куда-нибудь поесть, но мы поели в отеле однажды, когда МЕНЮ включены **омаров**.

6. Conclusions

We created an automated translation solution that fully answered the project objectives and the client's requirements. The entire process of system development and customization took about a month. The solution we created has the following features:

- Fast and efficient translation of large volumes of texts.
- High quality translation.
- Low costs for development and customization of the MT system (compared to the manual translation costs).
- Accurate and efficient quality estimation system.
- The solution was integrated into the TripAdvisor workflow with minimal costs for development and support on the client's side.

We managed to show how an efficient MT solution for translating user-generated content can be developed and customized within a short period and with no parallel in-domain data.

References

1. *Banerjee P., Naskar S. K., Roturier J., Way A., Genabith J.* (2011), "Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling", available at: http://mt-archive.info/MTS-2011-Banerjee.pdf
2. *Banerjee P., Naskar S. K., Roturier J., Way A., Genabith J.* (2012), "Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data?", available at: http://nclt.dcu.ie/mt/papers/Banerjee_EAMT_2012.pdf
3. *Flournoy R., Callison-Burch C.* (2000), "Reconciling User Expectations and Translation Technology to Create a Useful Real-world Application", available at: http://mt-archive.info/Aslib-2000-Flournoy.pdf
4. *Flournoy R., Rueppel J.* (2010), "One Technology: Many Solutions", available at: http://amta2010.amtaweb.org/AMTA/papers/4-05-FlournoyRueppel.pdf
5. *Jiang J., Way A., Haque R.* (2012), "Translating user-generated content in the social networking space", available at: http://amta2012.amtaweb.org/AMTA2012Files/papers/JiangWayHaque.pdf
6. *Molchanov A.* (2012), "PROMT DeepHybrid system for WMT12 shared translation task", available at: http://www.statmt.org/wmt12/pdf/WMT43.pdf

ВЛИЯНИЕ РАЗЛИЧНЫХ ТИПОВ ОРФОГРАФИЧЕСКИХ ОШИБОК НА КАЧЕСТВО СТАТИСТИЧЕСКОГО МАШИННОГО ПЕРЕВОДА

Мещерякова Е. М. (mescheryakova@yandex-team.ru),
Галинская И. Е. (galinskaya@yandex-team.ru),
Гусев В. Ю. (vgoussev@yandex-team.ru),
Шматова М. С. (mashashma@yandex-team.ru)

Яндекс, Москва, Россия

В статье рассматривается рост качества машинного перевода в зависимости от исправления различных типов ошибок в исходном тексте на материале трех языковых пар (англо-русской, немецко-русской и польско-русской). Мы выбрали по 500 случайных пользовательских запросов к сервису машинного перевода, последовательно исправили в них разные типы опечаток и ошибок: отсутствующую диакритику; опечатки любого рода; неправильную пунктуацию и капитализацию; все ошибки. Всего для немецкого и польского языков мы получили по пять тестовых наборов (включая оригинал), для английского — четыре (в нем отсутствует диакритика). Все наборы были протестированы на трех бесплатных статистических системах машинного перевода, и для каждого было измерено значение BLEU.

Исправление всех опечаток дает увеличение BLEU примерно на 10–15% по сравнению с оригинальными запросами. Исправление опечаток и ошибок в пунктуации и капитализации по отдельности дают улучшение примерно на 5–10% в зависимости от языка и особенностей тестового набора. Исправление же только диакритики прироста почти не дает: 0% для немецкого языка и 0,5–1% для польского.

Ключевые слова: статистический машинный перевод, качество машинного перевода, метрика BLEU, опечатки, капитализация, пунктуация

IMPACT OF DIFFERENT TYPES OF SPELLING MISTAKES ON THE QUALITY OF STATISTICAL MACHINE TRANSLATION

Mescheryakova E. M. (mescheryakova@yandex-team.ru),
Galinskaya I. E. (galinskaya@yandex-team.ru),
Gusev V. Yu. (vgoussev@yandex-team.ru),
Shmatova M. S. (mashashma@yandex-team.ru)

Yandex, Moscow, Russia

Errors in the original text will most probably affect the quality of machine translation. It would be interesting to see how different types of errors can influence the translation. To do this, we selected three sets of 500 random queries in English, German and Polish. In each set we corrected different types of errors: 1) missing diacritical marks (except English); 2) all misprints (including diacritics); 3) errors in punctuation and use of capitals; 4) all types of errors listed in 1)–3). As a result we had five sets of 500 queries for German and Polish and four sets for English.

Then we translated all the sets into Russian using three free online statistical machine translation systems and compared their BLEU scores to see how they increase in corrected tests as compared to the original ones.

We also used different types of BLEU: along with the usual one, which treats punctuation signs as words, we used simplified BLEU which disregards punctuation, and also extended BLEU which takes into consideration both punctuation and use of capitals.

We show that in a fully corrected text BLEU increases by approx. 10–15% as compared to original sets. Correcting each of the two main types of errors — misprints and punctuation/capitalization — gives an increase of 5–10% each depending on the language and on the peculiarities of the test sets. On the other hand, correcting only diacritics has very small impact on the translation quality: close to zero in German and 0,5–1% in Polish.

Key words: statistical machine translation, machine translation quality, BLEU, misprints, capitalization, punctuation

1. Введение

Влияние грамотности исходного текста на качество машинного перевода кажется очевидным: орфографические ошибки[1] в оригинале затрудняют распознавание словоформ и их последовательностей и, соответственно, процесс перевода (см., например, Carrera et al. 2009; Plesco, Rychtyckyj 2012). Отсутствие опечаток/орфографических ошибок, правильная расстановка знаков препинания и заглавных букв входит — наряду со стилистическими требованиями — в число условий подготовки текста для машинного перевода (ср. понятие «controlled/standard language»; см. Aikawa et al. 2007 о сравнительном влиянии нескольких стилистических и орфографических факторов на итоговое качество перевода). Соответственно, устранение опечаток может быть частью предобработки текста, предназначенного для любого автоматического анализа (см., например, Shoukry, Rafea 2012 о нормализации арабских текстов из твиттера для сентимент-анализа).

С распространением бесплатных онлайн-сервисов машинного перевода все большую долю в них занимают тексты, которые не подвергаются предварительному редактированию, могут содержать большое количество опечаток, неверную пунктуацию и употребление заглавных букв (Jiang et al. 2012). Добавим сюда случаи перевода интернет-страниц, элементы которых, будучи скопированы в окно перевода, превращаются в набор слов и словосочетаний, не разделенных пунктуацией.

Наконец, отдельную проблему представляет употребление диакритики в тех языках, где она существует: зачастую авторы разного рода «субстандартных» текстов склонны пренебрегать ею. Обычно это происходит при наборе текста на клавиатуре, предназначенной для другого языка и не имеющей нужных символов; таким образом, в отличие от других типов ошибок, возникающих по вине пользователя (по небрежности или неграмотности), диакритика чаще отсутствует по техническим причинам.

Некоторые онлайн-системы перевода помогают пользователю исправить опечатки в исходном тексте — указывая на слово, содержащее ошибку, или даже предлагая варианты исправления. Однако при переводе с иностранного языка воспользоваться такими подсказками, очевидно, сложно: не зная иностранного слова даже на уровне понимания общего смысла, пользователь в большинстве случаев не будет уверен и в его написании. Можно предположить, что обработка исходного текста должна производиться автоматически, являясь составной частью алгоритма машинного перевода.

Интересным кажется узнать, насколько различные типы ошибок в запросе влияют на итоговое качество перевода. Для нашего исследования мы взяли три языковые пары: англо-русскую, немецко-русскую и польско-русскую. Для каждой пары был взят набор из 500 случайных пользовательских запросов к сервису машинного перевода, которые последовательно тестировались:

---

[1] Для наших целей неважно различие между орфографическими и грамматическими ошибками, происходящими от недостаточного знания автором правил данного языка, и случайными опечатками.

- в оригинальном виде, как они задавались пользователями;
- только с исправленной диакритикой (кроме английского);
- с исправленной диакритикой и прочими опечатками/ошибками, включая лишние или недостающие пробелы;
- с исправленной капитализацией и пунктуацией;
- полностью исправленными.

В разделе 2 мы приведем примеры типичных ошибок в пользовательских запросах; в разделе 3 более подробно охарактеризуем тестовые наборы; в разделе 4 опишем последовательность эксперимента и применявшиеся в нем метрики. Наконец, в разделе 5 будут приведены собственно результаты измерений и комментарии к ним.

2. Типичные ошибки в запросах

Приведем типичные примеры ошибок в запросах (все примеры сконструированы нами либо взяты из открытых интернет-источников).

1. Опечатки:

- случайные опечатки, пропуск или перестановка одной или нескольких букв, вставка или пропуск пробела и т. д.: англ. *chanel* → *channel* ‘канал’, *theey* → *they* ‘они’, *sayi ng* → *saying* ‘говоря’; *dont’* → *don’t*; польск. *Warszwa* → *Warszawa* ‘Варшава’;
- отсутствие пробела между словами, разделенными знаками препинания: англ. *I saw him yesterday.He said…* → *I saw him yesterday. He said…* ‘Я видел его вчера. Он сказал…’. Такого рода ошибки часто бывают систематическими (т. е. автор последовательно не ставит пробелы после знаков препинания), и, поскольку многие системы не умеют разделять слова без пробела, приводят к резкому снижению качества перевода.

2. Отсутствие диакритики:

- польск. *zolte zloto* → *żółte złoto* ‘желтое золото’, нем. *mude* → *müde* ‘усталый’ (автор пользуется клавиатурой, не имеющей клавиш для особых символов и диакритических знаков). Пропуск диакритики на отдельных буквах (не систематически во всем запросе), по крайней мере, в рассматриваемых здесь языках встречается редко, потому что символы с диакритикой расположены на отдельных клавишах и обычно не по соседству с соответствующими простыми символами; единственное исключение — *l* и *ł* на соседних клавишах в польской раскладке клавиатуры. Разумеется, неиспользование диакритики возможно только при неформальной переписке.

3. Отсутствие капитализации и/или пунктуации:

- нем. *sind diese probleme für dich so wichtig* → *Sind diese Probleme für dich so wichtig?* ‘Эти проблемы для тебя так важны?’ (пользователь в переписке или в чате пренебрегает заглавными буквами и знаками препинания; как и в случае с диакритикой, такого рода «послабления» возможны только в неформальных текстах);

- польск. *data i miejsce urodzenia adres nr dokumentu tożsamości/paszportu* → *Data i miejsce urodzenia, adres, nr dokumentu tożsamości/paszportu* ‘Дата и место рождения, адрес, номер документа, удостоверяющего личность / паспорта’ (при копировании содержимого веб-страницы названия полей, которые требуется заполнить, сливаются в единую последовательность слов);
- польск. *Mam na sprzedaż nowy VW. AUTO W STANU BARDZO DOBRYM* ‘Продаю новый VW. Машина в очень хорошем состоянии’ (автор выделяет заглавными буквами важную часть сообщения);
- особый случай представляют собой обращения и приветствия в письмах, которые часто отделяются запятой, а далее идет текст с заглавной буквы и с новой строки; при копировании в окно перевода разрыв строк исчезает, и запятая оказывается перед заглавной буквой: польск. *Witam, Uprzejmie informuję…* ‘Добрый день. С уважением сообщаю…’ (обычная формулировка для официального письма).

Отметим, что в одном запросе могут содержаться и часто содержатся разные типы ошибок.

3. Характеристика тестовых наборов

Каждый из трех тестовых наборов, использованных в эксперименте, состоит из 500 случайных запросов к системе машинного перевода, каждый — длиной не более 1000 символов. Свойства наборов, однако, довольно сильно отличаются, что вызвано, с одной стороны, свойствами языков, с другой — разницей в тематике запросов.

В каждом языке могут быть свои особенности — в том числе орфографические, — которые могут приводить к ошибкам в запросах. Так, немецкий и польский языки различаются по количеству букв с диакритикой: 3 в немецком и 9 в польском; соответственно, игнорирование диакритики в польском языке сильнее искажает текст, и можно предположить, что и степень влияния этого фактора на качество перевода будет выше. С другой стороны, в немецком языке принято писать с большой буквы все имена существительные; соответственно, пренебрежение капитализацией вызывает дополнительные орфографические ошибки.

Тематика рассматриваемых тестовых наборов также достаточно сильно отличается. Так, к примеру, в переводах с немецкого языка около 40 % составляют учебные тексты и упражнения, еще около 25 % — литературные тексты, переводы веб-страниц — 9 %, а переписка любого рода (включая чаты) — лишь 8 %. Напротив, в запросах на перевод с английского языка доля переписки превышает 30 %, доля учебных текстов составляет около 20 %, примерно столько же составляют переводы веб-страниц, а доля литературных текстов — всего около 8 %. В польском же переводы веб-страниц находятся в лидерах — почти 45 %; следом идет переписка (около 40 %), а литературные и учебные тексты вместе составляют около 10 %.

Очевидно, разница в тематике является одной из причин различной средней длины запросов, см. Таблицу 1.

Таблица 1. Средняя длина запроса

Язык	Среднее количество слов в запросе
Английский	17
Немецкий	23
Польский	20

Можно было бы ожидать, что такое различие в тематике скажется на среднем уровне грамотности запросов: литературные тексты (которые, скорее всего, не набираются вручную, а копируются из какого-либо источника в интернете) должны быть орфографически и пунктуационно выверены, в отличие, к примеру, от сообщений в чатах. В этом случае средний уровень грамотности немецких запросов должен быть выше, чем, скажем, английских. Сравним, однако, данные о количестве запросов с ошибками разных типов в Таблице 2:

Таблица 2. Доля запросов с разными типами ошибок в тестовых наборах

Язык	Диакритика	Опечатки (включая диакритику)	Капитализация + пунктуация	Ошибки любого типа
Английский	—	32,4%	38%	53,2%
Немецкий	5%	40,2%	48,8%	67,2%
Польский	12%	36,6%	62,2%	71,4%

Мы видим, что наши ожидания не полностью оправдываются. В английском тестовом наборе, несмотря на большое количество потенциально «ненормативной» переписки, доля ошибочных запросов как в целом, так и по отдельным типам ошибок ниже всего. Количество ошибок в употреблении заглавных букв и в пунктуации в польском языке значительно превышает количество аналогичных ошибок в немецком.

Часть этих различий, тем не менее, можно объяснить. Соотношение ошибок в диакритике в польском и немецком языке в целом соответствует ожиданиям. Большое количество ошибок на капитализацию и пунктуацию в польском языке, очевидно, происходит из-за переводов веб-страниц, отдельные элементы которых при копировании сливаются (см. примеры ошибок выше).

4. Методика эксперимента

Для сравнения результатов в каждом из трех тестовых наборов последовательно исправлялись ошибки различных типов. В результате, для каждого языка, помимо оригинального, были созданы следующие четыре набора:

- с исправленной диакритикой (кроме английского);
- с исправленной диакритикой и прочими опечатками/ошибками, включая лишние или недостающие пробелы;

- с исправленными капитализацией и пунктуацией;
- полностью исправленные.

Всего, таким образом, для немецкого и польского языков у нас было по пять наборов по 500 запросов, для английского — четыре.

Для каждого языка были подготовлены эталоны переводов на русский язык, после чего все наборы были протестированы на трех бесплатных статистических онлайн-системах машинного перевода. Для оценки качества применялась метрика BLEU (Bilingual Evaluation Understudy), широко используемая для оценки статистического машинного перевода.

BLEU основана на сравнении машинного перевода с эталоном, сделанным человеком. Для этого подсчитывается количество последовательностей из n слов (n-граммов), совпадающих в сравниваемом переводе и в эталоне; n обычно берется от 1 до 4. Значение BLEU высчитывается в среднем для всего корпуса переводов (в нашем случае 500 фрагментов для каждого языка) и составляет от 0 до 1 (либо от 0 до 100), где 0 означает, что совпадения отсутствуют, а 1 (100) — что сравниваемый корпус полностью идентичен эталону (см. подробнее о метрике BLEU: Papineni et al. 2002).

В нашем эксперименте использованы три метрики BLEU: а) стандартная (учитывающая пунктуационные знаки как отдельные токены); б) BLEU без учета пунктуации; и в) BLEU с учетом пунктуации и капитализации (т. е. учитывающая также различие строчных и заглавных букв в переводе и эталоне).

5. Результаты эксперимента

В этом разделе приводятся данные по изменению BLEU в трех системах машинного перевода для каждого языкового набора при исправлении ошибок разного типа: опечаток, капитализации и пунктуации, всех ошибок. В скобках указывается прирост значения BLEU по сравнению с исходными запросами.

5.1. Приведем результаты подсчетов стандартного BLEU (с учетом пунктуации)

Таблица 3

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
С 1	28,8	30,7 (+1,9)	30,1 (+1,3)	32,1 (+3,3)
С 2	30,9	33,1 (+2,2)	32,1 (+1,2)	34,5 (+3,6)
С 3	26,6	28,0 (+1,4)	28,9 (+2,3)	30,2 (+3,6)

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Немецкий язык				
C 1	23,9	26,2 (+2,3)	24,2 (+0,3)	26,9 (+3,0)
C 2	22,6	24,4 (+1,8)	23,0 (+0,4)	25,4 (+2,8)
C 3	20,4	21,8 (+1,4)	20,8 (+0,4)	22,2 (+1,8)
Польский язык				
C 1	33,1	35,0 (+1,9)	37,7 (+4,6)	40,0 (+6,9)
C 2	20,9	22,0 (+1,1)	26,1 (+5,2)	27,3 (+6,4)
C 3	20,0	20,6 (+0,6)	24,2 (+4,2)	24,9 (+4,9)

Прокомментируем результаты измерений.

1. В английском языке исправление обоих типов ошибок дает более или менее равномерный прирост качества (возможно, это соотносится с тем, что процент запросов с ошибками каждого из этих типов в английском сравним — 32,4% и 38%, см. Таблицу 1). Правда, в разных системах вклад двух типов ошибок может отличаться: в C1 и C2 сильнее влияние опечаток, в C3 больше влияет капитализация/пунктуация.
2. В немецком обращает на себя внимание существенно большее влияние на рост BLEU исправление опечаток по сравнению с исправлением ошибок капитализации и пунктуации — хотя процент запросов с ошибками второго типа не меньше, а даже несколько больше, чем первого (40,2% и 48,8% соответственно — соотношение близко к тому, которое мы видели в английском наборе). Сходство соотношения во всех системах подтверждает этот результат.
3. В польском языке, напротив, очень сильно влияние капитализации и пунктуации и невелик вклад опечаток. В определенной степени это связано с особенностями тестового набора и большой долей таких ошибок в нем (62,2% запросов), однако это, очевидно, не единственная причина: по сравнению с английским доля запросов с опечатками в польском несколько выше (36,6% против 32,4%), а влияние их исправления на качество ниже во всех системах перевода.

Посмотрим, насколько влияет на качество перевода исправление только диакритики, без прочих опечаток, в польском и немецком языках. Приводятся только значения BLEU при исправленной диакритике и разница с исходными запросами.

Таблица 4

	Немецкий	Польский
C 1	23,8 (−0,1)	33,8 (+0,7)
C 2	22,6 (+0,0)	21,0 (+0,1)
C 3	20,4 (+0,0)	20,1 (+0,1)

Как видно, исправление только диакритики дает очень небольшой эффект. В немецком языке он вовсе нулевой. В польском он отличен от нуля — что, видимо, объясняется бóльшим количеством букв с диакритикой и, соответственно, бóльшим процентом ошибок данного типа (см. Таблицу 2), — но тоже очень невелик. Более или менее заметен он в польском языке только в системе 1, даже несмотря на то, что и общие показатели BLEU для польского языка в ней выше; в процентном отношении прирост BLEU в Системе 1 составляет 2,1%, в Системах 2 и 3 — 0,5%.

Приведем для сравнения результаты измерений по другим метрикам BLEU.

5.2. BLEU без учета пунктуации

Таблица 5

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
C 1	24,4	26,3 (+1,9)	24,9 (+0,5)	26,7 (+2,3)
C 2	26,9	28,8 (+1,9)	26,8 (−0,1)	28,8 (+1,9)
C 3	23,2	24,6 (+1,4)	23,7 (+0,5)	25,2 (+2,0)
Немецкий язык				
C 1	18,5	20,6 (+2,1)	18,7 (+0,2)	21,2 (+2,7)
C 2	17,5	19,1 (+1,6)	17,5 (+0,0)	19,7 (+2,2)
C 3	16,1	17,1 (+1,0)	16,1 (+0,0)	17,3 (+1,2)
Польский язык				
C 1	29,7	31,8 (+2,1)	30,0 (+0,3)	32,2 (+2,5)
C 2	17,1	18,2 (+1,1)	18,4 (+1,3)	19,4 (+2,3)
C 3	16,2	16,9 (+0,7)	16,8 (+0,6)	17,3 (+1,1)

Ожидаемым образом, здесь резко уменьшаются цифры в предпоследней колонке, в немецком языке — вовсе до нуля. Оставшиеся цифры, очевидно, показывают реальный вклад исправления капитализации и пунктуации в качество перевода как такового. В то же время влияние исправленных опечаток в немецком и польском языках при таком способе измерения несколько повышается.

5.3. BLEU с учетом капитализации и пунктуации

Таблица 6

	Неисправленные запросы	Исправлены всеопечатки	Исправлены капитализация и пунктуация	Исправлены все ошибки
Английский язык				
C 1	26,7	28,4 (+1,7)	28,8 (+2,1)	30,6 (+3,9)
C 2	29,3	31,4 (+2,1)	30,8 (+1,5)	33,2 (+3,9)
C 3	24,9	26,2 (+1,3)	27,6 (+2,7)	28,8 (+3,9)
Немецкий язык				
C 1	21,7	24,0 (+2,3)	22,4 (+0,7)	24,9 (+3,2)
C 2	21,3	23,0 (+1,7)	21,9 (+0,6)	24,1 (+2,8)
C 3	19,1	20,4 (+1,3)	19,6 (+0,5)	21,0 (+1,9)
Польский язык				
C 1	30,1	31,7 (+1,6)	36,5 (+6,4)	38,7 (+8,6)
C 2	19,5	20,5 (+1,0)	25,0 (+5,5)	26,3 (+6,8)
C 3	18,3	18,9 (+0,6)	23,1 (+4,8)	23,9 (+5,6)

Заключение

Данная работа посвящена влиянию орфографической правильности исходного текста на качество статистического машинного перевода. Мы показали, что исправление всех — орфографических и пунктуационных — ошибок может дать прирост качества по метрике BLEU примерно на 10–15 % по сравнению с неисправленным текстом. Что касается отдельных классов ошибок: орфографических, с одной стороны, и в пунктуации и капитализации — с другой, вклад каждого из этих классов может различаться в зависимости от свойств языка и особенностей запросов. Мы также рассмотрели влияние на перевод отдельного типа ошибок — отсутствия диакритики, которое может возникать не только по небрежности пользователя, но и по техническим причинам. Вклад исправления этого типа ошибок в качество перевода оказался невелик — впрочем, он в большей мере, чем другие, зависит от конкретного языка.

Было бы интересно проверить полученные результаты на других языковых парах. Кроме того, в дальнейшем необходимо исследовать влияние на качество перевода других типов ошибок — в первую очередь синтаксических.

Литература

1. *Aikawa T., Schwartz L., King R., Corston-Oliver M., Lozano C.* Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. MT Summit XI, 10–14 September. Proceedings. Copenhagen, 2007, pp. 1–7.
2. *Shoukry A., Rafea A.* Preprocessing Egyptian dialect tweets for sentiment mining. AMTA-2012: Fourth workshop on computational approaches to Arabic script-based languages. Proceedings. San Diego, 2012, pp. 47–56.
3. *Jiang J., Way A., Haque R.* Translating user-generated content in the social networking space. AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas. Proceedings. San Diego, 2012.
4. *Carrera J., Beregovaya O., Yanishevsky A.* (2009). Machine Translation for Cross-Language Social Media, available at: http://www.promt.com/company/technology/pdf/machine_translation_for_cross_ language_social_media.pdf
5. *Papineni K., Roukos S., Ward T., Zhu W. J.* (2002). BLEU: a method for automatic evaluation of machine translation. ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. Proceedings. Stroudsburg, 2002, pp. 311–318.
6. *Pesko C., Rychtyckyi N.* Machine Translation as a Global Enterprise at Ford. AMTA-2012: the Tenth Biennial Conference of the Association for Machine Translation in the Americas. Proceedings. San Diego, 2012.

КОНТЕКСТНО-ЗАВИСИМЫЙ ПЕРЕВОД СЛОВАРЯ ОЦЕНОЧНЫХ СЛОВ ПРИ ПОМОЩИ ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ

Уланов А. В. (alexander.ulanov@hp.com),
Сапожников Г. А. (gsapozhnikov@gmail.com)

Hewlett-Packard Labs Russia, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: анализ мнений, оценочные слова, машинный перевод, классификация

CONTEXT-DEPENDENT OPINION LEXICON TRANSLATION WITH THE USE OF A PARALLEL CORPUS

Ulanov A. V. (alexander.ulanov@hp.com),
Sapozhnikov G. A. (gsapozhnikov@gmail.com)

Hewlett-Packard Labs Russia, St. Petersburg State University, St. Petersburg, Russia

The paper deals with multiligual sentiment analysis. We propose a method for projecting an opinion lexicon from a source language to a target language with the use of a parallel corpus. We can make sentiment classification in a target language using an opinion lexicon even if we have no labeled dataset. The advantage of our method is that it captures the context of a word and thus produces a correct translation of it. We apply our method to the language pair English-Russian and conduct sentiment classification experiments. They show that our method allows creating high-quality opinion lexicons.

Keywords: opinion mining, sentiment analysis, opinion words, machine translation

1. Introduction

Sentiment analysis is one of the most popular information extraction tasks both from business and research prospective. It has numerous business applications, such as evaluation of a product or company perception in social media. From the standpoint of research, sentiment analysis relies on the methods developed for natural language processing and information extraction. One of the key aspects of it is the opinion word lexicon. Opinion words are such words that carry opinion. Positive words refer to some desired state, while negative words — to some undesired one. For example, "good" and "beautiful" are positive opinion words, "bad" and "evil" are negative. Opinion phrases and idioms exist as well. Many opinion words depend on context, like the word "large". Some opinion phrases are comparative rather than opinionated, for example "better than". Auxiliary words like negation can change sentiment orientation of a word.

Opinion words are used in a number of sentiment analysis tasks. They include document and sentence sentiment classification, product features extraction, subjectivity detection etc. [12]. Opinion words are used as features in sentiment classification. Sentiment orientation of a product feature is usually computed based on the sentiment orientation of opinion words nearby. Product features can be extracted with the help of phrase or dependency patterns that include opinion words and placeholders for product features themselves. Subjectivity detection highly relies on opinion word lists as well, because many opinionated phrases are subjective [14]. Thus, opinion lexicon generation is an important sentiment analysis task. Detection of opinion word sentiment orientation is an accompanying task.

Opinion lexicon generation task can be solved in several ways. The authors of [12] point out three approaches: manual, dictionary-based and corpus-based. The manual approach is precise but time-consuming. The dictionary based approach relies on dictionaries such as WordNet. One starts from a small collection of opinion words and looks for their synonyms and antonyms in a dictionary [10]. The drawback of this approach is that the dictionary coverage is limited and it is hard to create a domain-specific opinion word list. Corpus-based approaches rely on mining a review corpus and use methods employed in information extraction. The approach proposed in [9] is based on a seed list of opinion words. These words are used together with some linguistic constraints like "AND" or "OR" to mine additional opinion words. Clustering is performed to label the mined words in the list as positive and negative. Part of speech patterns are used to populate the opinion word dictionary in [21] and Internet search statistics is used to detect semantic orientation of a word. Work [7] extends the mentioned approaches and introduces a method for extraction of context-based opinion words together with their orientation. Classification techniques are used in [2] to filter out opinion words from text. The approaches described were applied in English. There are some works that deal with Russian. For example, paper [4] proposes to use classification. Various features, such as word frequency, weirdness, and TF-IDF are used there.

Most of the research done in the field of sentiment analysis relies on the presence of annotated resources for a given language. However, there are methods

which automatically generate resources for a target language, given that there are tools and resources available in the source language. Different approaches to multilingual subjectivity analysis are studied in [14] and [1] and are summarized in [3]. In one of them, subjectivity lexicon in the source language is translated with the use of a dictionary and employed for subjectivity classification. This approach delivers mediocre precision due to the use of the first translation option and due to word lemmatization. Another approach suggests translating the corpus. This can be done in three different ways: translating an annotated corpus in the source language and projecting its labels; automatic annotation of the corpus, translating it and projecting the labels; translating the corpus in the target language, automatic annotation of it and projecting the labels. Language Weaver[1] machine translation was used on English-Roman and English-Spanish data [3]. Classification experiments with the produced corpora showed similar results. They are close to the case when test data is translated and annotated automatically. This shows that machine translation systems are good enough for translating opinionated datasets. It is also confirmed by the authors of [19] when they used Google Translate[2], Microsoft Bing Translator[3] and Moses[4].

Multilingual opinion lexicon generation is considered in the recent paper [19] that presents a semi-automatic approach with the use of triangulation. The authors use high-quality lexicons in two different languages and then translate them automatically into a third language with Google Translate. The words that are found in both translations are supposed to have good precision. It was proven for several languages including Russian with the manual check of the resulting lists. The same authors collect and examine entity-centered sentiment annotated parallel corpora [20].

In this paper we develop the idea of multilingual sentiment analysis. We propose a method for projecting an opinion lexicon from a source language to a target language with the use of a parallel corpus. We apply it to the language pair English-Russian having a collection of a parallel and a pseudo-parallel review corpora. The method is evaluated against the baseline, which is a translation of the opinion word lexicon with Goolge Translate. Sentiment classification experiments are conducted to evaluate the quality of the lexicons. The advantages of our method are the following. It captures the context of opinion words thus producing correct translations. It doesn't require a machine translation tool, as in [19] or a bilingual dictionary as in [14]. However, machine translation tool may be employed in the absence of parallel corpus or for better recall. The opinion lexicon is needed only in one language, unlike in work [19] where 2 lexicons are required.

---

1 http://www.sdl.com/products/automated-translation/

2 http://translate.google.com/

3 http://www.bing.com/translator

4 http://www.statmt.org/moses/

2. Approach

The idea of our approach is to use a parallel corpus to construct an opinion lexicon in a target language, given that there is an opinion lexicon in a source language. A parallel corpus is a text with its translation to the target language. We suppose that it contains opinionated sentences. An opinion lexicon is a set of words carrying opinion. It is not necessarily divided into positive/negative or other groups. The opinion lexicon for the target language is extracted from the parallel corpus by translating the words from the opinion lexicon in the source language. The algorithm of the method is as follows:

1. Collect a corpus of parallel reviews, align sentences
2. Compute word lexical translation probabilities
3. Collect opinion words translations and normalize them

Let us consider the mentioned steps in greater details. The task of parallel corpus acquisition and preparation is a well-studied area of research [8]. One collects or crawls data that is available in different languages. Parallel documents are determined by some identifier, e.g. name, time, or specific number. Documents are split into sentences by the sentence splitter, paragraphs are kept preserved. The resulting text is processed by the sentence aligner. A parallel corpus with opinionated texts can be obtained from the sites that post reviews in different languages (manually translated). Usually, such reviews are editorial. They contain opinionated text; however opinion words there tend to be more polite than in forums or user reviews. The size of the corpus is less important than the coverage of words from the source opinion lexicon. In the absence of a natural parallel corpus, a pseudo-parallel corpus can be used [20], which is a text along with its translation done by an automatic translation system.

Lexical translation probabilities of words are computed on the aligned corpus:

$$p_s(t) \text{ and } p_t(s),$$

where *t* is a word in the target language, *s* is a word in the source language. Lexical translation is a translation of a word in isolation. To compute it, one has to count how many times a certain word was translated into different options within the aligned sentences. The ratios of these counts and the count of that word represent the distribution of lexical translation probabilities. This operation is performed in both translation directions, i. e. $t \to s$ and $s \to t$.

Opinion word translations are collected for a given opinion word list in the source language. Correct translation of a source opinion word is determined as follows:

$$\exists t, s: p_t(s) = \max_i p_i(s) \;\; and \;\; p_s(t) = \max_j p_j(t)$$

In other words, to make translation of a source word, we choose a word with a maximum translation probability and check that it translates to the same word with a maximum probability as well. The translated words are normalized.

3. Experiments

3.1. Opinion lexicon projection

We conducted several experiments to validate the proposed approach. Two parallel datasets are used in our experiments. The first one consists of Russian and English reviews collected from the Mobile Review site[5]. We downloaded all pages from the English editorial of the site. Then we downloaded Russian versions of these pages using English links without the token "-en". We will refer to this dataset as to "MR".

The second one consists of the first 5,000 lines from the reviews of books, cameras and films taken from ROMIP 2011 sentiment analysis dataset [5] and 1,000 lines of iPhone4 reviews from Yandex Market[6] along with their Russian translation produced by Google Translate. We will refer to it as to "ROMIP-GT". The datasets are split into sentences with Freeling[7] and aligned with Microsoft Bilingual Sentence Aligner [18]. After the above mentioned, the aligned "MR" contains 579,559 Russian and 726,798 English words, the aligned "ROMIP-GT" contains 714,533 Russian and 820,241 English words. We use GIZA++ [15] for creating word lexical translation tables. English opinion word lists are downloaded from Bing Liu's homepage[8]. There are 4,818 negative and 2041 positive words. We will refer to this list as to "BL" dictionary. Mystem[9] is used to normalize the Russian words. They are transformed to singular, masculine, nominative, present time forms.

We produce 4 opinion lexicons in Russian in total. During lexicons construction we remove all words containing spaces and minuses, and which are shorter than 3 symbols. "BL-GT" lexicon contains translated and normalized opinion words from "BL". "BL-GT filtered" lexicon was constructed in the following way. Words from "BL" were translated to Russian and then back to English using Google Translate. We collected only those Russian translations that produced English translation equal to its English original.

"MR" lexicon is created by application of our method to "MR" parallel corpora. "ROMIP-GT" lexicon is created using our method with the "ROMIP-GT" dataset. "ROMIP-GT merged" lexicon is produced in the following way. We applied our method to 3 subsets of "ROMIP-GT", i.e. books, films and cameras. Then the resulting lists were merged. The number of opinion words in each lexicon is listed in Table 1. Table 2 shows intersections of the lexicons.

---

5 http://mobile-review.com/

6 http://market.yandex.ru/

7 http://nlp.lsi.upc.edu/freeling/

8 http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

9 http://company.yandex.ru/technologies/mystem/

Table 1. Opinion words number

Lexicon	Positive	Negative	Total
BL (English)	*2,041*	*4,818*	*6,859*
BL-GT	1,443	3,067	4,510
BL-GT filtered	907	2,037	2,944
MR	163	182	345
ROMIP-GT	706	1,311	2,017
ROMIP-GT merged	1,057	1,812	2,869
Union:	1,993	4,040	6,033

The lexicon "BL-GT" is the biggest with almost 4.5 thousand words. However, it is less than the original list by 34%. This is due to the fact that some words were translated to the same surface form (27%), due to phrases removal (they contain spaces) and due to normalization. There is a small portion of untranslated words as well. "BL-GT filtered" is almost a half of the original dictionary. It is interesting to see, however, that so many words are translated from English to Russian and back to English with the original form.

"MR" lexicon that was produced from the Mobile Review parallel corpus is rather small. This is because it contains a different English lexicon than the opinion word list "BL". The "MR" texts were written by a limited number of persons, while the opinion lexicon "BL" contains contributions from a lot of people.

Interestingly, "ROMIP-GT merged" is 30% bigger than "ROMIP-GT" and is almost as big as "BL-GT filtered". Table 2 suggests that "ROMIP-GT merged" has 1222 or 45% of words in common with "BL-GT filtered". This is because the words in the latter case were translated in isolation while in the first case they were translated within the context.

We can get as many as 6,033 opinion words if we merge all lists, which is 89% of the original English list.

Table 2. Opinion words intersection

Intersection		Words		
		pos	neg	total
MR	ROMIP-GT merged	118	88	206
MR	BL-GT	132	178	310
ROMIP-GT merged	BL-GT	626	1,006	1,632
ROMIP-GT merged	BL-GT filtered	436	786	1,222

We made a manual assessment of the lexicons. Table 3 shows their precision. "BL-GT filtered" is the most accurate. This can be explained by the fact that it contains just the right English words translated unambiguously without context. Also, we compared "MR" and "ROMIP-GT" lists. The first was derived from professional reviews, the second from user reviews. It is interesting to note that "MR" contains "specific" opinion words and "ROMIP" contains emotional words.

Table 3. Precision by manual assessment

Lexicon	Precision
BL-GT	0,79
BL-GT filtered	0,87
MR	0,76
ROMIP-GT	0,83
ROMIP-GT merged	0,82

3.2. Document Sentiment Classification

The number of words in the list doesn't mean its quality. We conducted several experiments to benchmark the produced opinion word lists. We decided not to check the words manually, but to use them in the real-world task, that is sentiment classification. The experiments are performed on the annotated part of ROMIP 2011 dataset [5]. It contains reviews of books, films and cameras. There are 750 positive and 124 negative review instances.

Counting the number of positive and negative words is the most straightforward way to text sentiment classification [13]. The one with the greater number of opinion words wins. The work [17] suggests that it is better to consider the presence of an opinion word in text rather than the number of appearances. We implement both approaches. We will refer to the first as to "Frequency voc" and to the second as to "Binary voc".

Supervised approaches to text sentiment classification were studied by Pang et al. [17]. We use a linear perceptron classifier with two types of feature computation: term frequencies and delta TF-IDF. The latter was proposed by Martineau et al. [11] and proven to be efficient for sentiment classification in Russian [16]. The experiment results of these methods were obtained after performing 10-fold cross validation. These results act as a base line of supervised classification that requires an annotated dataset. We compare them with dictionary-based classification that does not require class labels to train, because it has negative and positive words. Therefore, results of supervised classification are considered as a higher bound for a dictionary based.

Table 4. Experiment results

Lexicon	Method	MicroP	MicroR (Acc)	MacroR	MacroF1
	Perceptron	**0.84**	**0.84**	0.59	0.60
	Perceptron + TfIdf	**0.84**	**0.84**	**0.62**	**0.63**
Romip-GT	Binary Voc	0.76	0.68	0.59	0.58
	Frequency Voc	0.79	0.72	0.59	0.59
Romip-GT merged	Binary Voc	**0.84**	0.80	0.59	0.61
	Frequency Voc	**0.86**	**0.82**	0.59	**0.61**

Lexicon	Method	MicroP	MicroR (Acc)	MacroR	MacroF1
BL-GT	Binary Voc	0.65	0.60	**0.62**	0.54
	Frequency Voc	0.73	0.69	0.59	0.56
BL-GT filtered	Binary Voc	0.78	0.78	0.59	0.58
	Frequency Voc	0.77	0.72	0.58	0.58
MR	Binary Voc	0.67	0.52	0.50	0.49
	Frequency Voc	0.66	0.53	0.51	0.50

The experiment results are represented in Table 4. The binary approach provides the same weight to all of the words. Low performance of the binary approach as compared with the frequency approach means that the lexicon is of low quality. It may contain common words that can be found in the text (that rarely speak about subjectivity). So we can say that "BL-GT" is rather dirty. "ROMIP-GT merged" gives the best performance among the opinion lexicons. It has the same number of words as "BL-GT filtered", but the performance of the "ROMIP-GT merged" is higher, so we can say that its quality for sentiment classification is better. It is because the words in "ROMIP-GT merged" were translated with the use of context unlike the words in "BL-GT filtered". "BL-GT filtered" shows better results in manual assessment, but worse results in classification. We can explain this by the fact that "ROMIP-GT merged" contains such words that out of context may seem not opinion words or words that are more often used in user reviews as compared with words from "BL-GT filtered".

We supposed that the increase in the classification performance could be due to the fact that we used a part of the big dataset ROMIP 2011 to retrieve "ROMIP-GT merged", and the labeled dataset that was used for classification was also a part of ROMIP 2011. However, it turned out that the intersection between these parts did not exceed 1%, and it couldn't lead to the significant increase of the classification performance.

We use our lexicons as a list for feature selection as in [6], and train a linear perceptron classifier. It produces nearly the same results both for "ROMIP-GT merged" and "BL-GT filtered". This experiment shows that "BL-GT filtered" contains enough words that can be used as classification features. However, it also contains common words that have low weight in the supervised classifier, which does not happen when this lexicon is used in vocabulary classification.

4. Conclusion

We proposed a novel method for opinion lexicon projection from a source language to a target language with the use of a parallel corpus. The method was applied to different datasets and evaluated against the baseline. The quality of created lexicons was evaluated in sentiment classification benchmark. The experiments showed that the lexicons are of high quality. They can be used for sentiment annotation of a corpus in a target language as well.

Out future work is related to enhancement of the method and conducting more experiments. We plan to work with opinion phrases, investigate other translation

options instead of the most probable ones. We will apply our method to other language pairs, apart from English-Russian. Additionally, it will be interesting to explore how the method can be applied to other tasks, such as subjectivity lexicon projection and, more general, multilingual projection of document features.

References

1. *Banea C., Mihalcea R., Wiebe J., Hassan S.* Multilingual subjectivity analysis using machine translation. EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008.
2. *Breck, E., Y Choi, and C. Cardie.* Identifying expressions of opinion in context. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-2007), 2007.
3. *Carmen Banea, Rada Mihalcea, and Janyce Wiebe.* Multilingual Sentiment and Subjectivity, in Multilingual Natural Language Processing, editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.
4. *Chatviorkin Ilya, Lukashevich Natalia.* Automatic Extraction of Domain-Specific Opinion Words. Proceedings of the International Conference Dialog, 2010.
5. *Chatviorkin Ilya, Braslavski Pavel, Lukashevich Natalia.* Sentiment analysis track at ROMIP 2011. Proceedings of the International Conference Dialog, 2012.
6. *Dang Y., Zhang Y., Chen H.* A lexicon-enhanced method for sentiment classification: An experiment on online product reviews, IEEE 2010.
7. *Ding, X., B. Liu, and P. Yu.* A holistic lexicon-based approach to opinion mining. In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.
8. *Eisele A., Chen Y.* MultiUN: A Multilingual Corpus from United Nation Documents. In Language Resources and Evaluation, 2010.
9. *Hatzivassiloglou, V. and K. McKeown.* Predicting the semantic orientation of adjectives. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997), 1997.
10. *Hu, M. and B. Liu.* Mining and summarizing customer reviews. In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), 2004.
11. *J. Martineau and T. Finin.* Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media, San Jose, CA, 2009.
12. *Liu, Bing.* Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. 2nd ed. 2011, XX, 622 p.
13. *Melville P., Gryc W., and Lawrence R.* Sentiment analysis of blogs by combining lexical knowledge with text classification. KDD 2009.
14. *Mihalcea R., Banea C. and Wiebe J.* Learning Multilingual Subjective Language via Cross-Lingual Projections, in Proceedings of the Association for Computational Linguistics (ACL 2007), Prague, June 2007.

15. *Och F. J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19–51 March 2003.
16. *Pak A., Paroubek P.* Language independent approach to sentiment analysis (Limsi participation in romip'11) Proceedings of the International Conference Dialog, 2012.
17. *Pang B., Lee L.* Thumbs up? Sentiment Classification using Machine Learning Techniques, In Proc. of the conference on the Empirical Methods 2002
18. *Robert C. Moore*. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (2002), pp. 135–144
19. *Steinberger J., Ebrahim M., Ehrmann M., Hurriyetoglu A., Kabadjov M., Lenkova P., Steinberger R., Tanev H., Vázquez S., Zavarella V.* Creating Sentiment Dictionaries via Triangulation. Decision Support Systems, May 2012.
20. *Steinberger J., Lenkova P., Kabadjov M., Steinberger R., van der Goot E.* Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In Proceedings of the International Conference Recent Advances in Natural Language Processing, 2011.
21. *Turney, P.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2002), 2002.

STATISTICAL MACHINE TRANSLATION WITH LINGUISTIC LANGUAGE MODEL

Zuyev K. A. (konst@abbyy.com),
Indenbom E. M. (Eugene_I@abbyy.com),
Yudina M. V. (Maria_Yu@abbyy.com)

ABBYY, Moscow, Russia

Stemming from traditional "rule based" translation a "model based" approach is considered as an underlying model for statistical machine translation. This paper concerns with training on parallel corpora and application of this model for parsing and translation.

Preface

Statistical machine translation has made a significant breakthrough in machine translation within past decade. Due to availability of huge parallel corpora and increased raw computational power it turned out that rather simple statistical methods rival (and beat from commercial point of view) the traditional rule based methods with foundation on years of linguistic research. Nevertheless, the further advances in statistical machine translation are considered to be related with more linguistically-rich models. Even such a commodity tool as Moses provides support for using parsing information in translation process.

Statistical Machine Translation — a short overview

In statistical machine translation target sentences are produced from sentences by so-called "noisy-channel" — a filter, which modifies input into output. The design of true filter is unknown but can be modeled by assuming some parametric model. The model's parameters can be tuned and the structure can be validated by comparing behavior of model and "true filter". In case of machine translation the existing parallel corpora provide possible input and outputs for the modeled filter.

Originally models for statistical machine translation were very simple — a sequence of words. Then, to model the context dependency of translation, the phrase models and hierarchical phrase models were introduced [4]. It turned out that more complex models (with richer parametric space) are hard to trained. So parse trees are used to restrict possible phrases and labels familiar to linguists such as NP, VP are used to guess hierarchical phrases [6]. Actually now this model is a context free transduction grammar.

Although linguistic notions are used, little linguistic research is in place. Instead, the corpora marked-up with parse trees are used to train parsers.

Proposed approach

Language model used in our approach is based on well-established concepts of (noncomputational) linguistics. For the more detailed description, see [1]. Here is a brief summary of the model.

We represent a sentence by an HPSG-style tree. We distinguish between *surface* and *semantic structure*. Surface structure is language dependent, while semantic structure is deemed as universal.

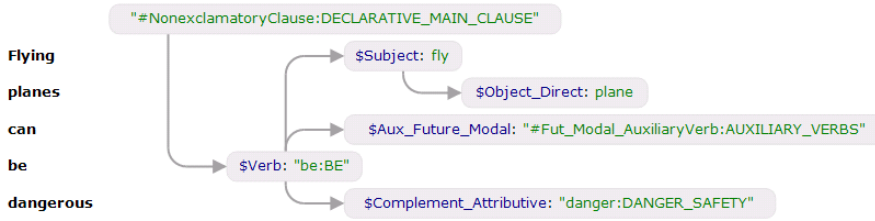
Therefore semantic structure is the "model" for translation process. Nodes of the tree (constituents) are normally formed from the words of the sentence. The constituent bears syntactic and semantic features. One of the most important features is *lexical class* — the representation of the meaning of the word. The meaning for our system is the position within our *semantic hierarchy*.

The *semantic hierarchy* (SH) — thesaurus-like hierarchical tree. It consists of universal nodes that represent different semantic concepts — semantic classes (SCs), which are filled with lexical items of natural languages — lexical classes (LCs). The main principle of organizing information within our hierarchy is the inheritance principle: higher nodes denote general notions, while their descendants denote more specific meaning and inherit main semantic and syntactic characteristics (these characteristics we call model) from their ancestors. Units of universal semantic information in our system are called *semantemes* — some of them are added in the hierarchy explicitly, others (for example, semantemes representing grammatical information such as tense, voice etc.) are computed during parsing.

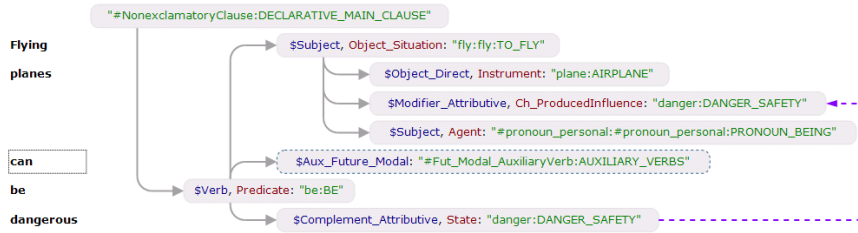
Dependencies between different units in the hierarchy are described in terms of semantic relations or *semantic slots* (which partly correlate to semantic roles, see [Fillmore 1968], for example). Semantic relations are also part of universal semantic structure and are language-independent. Dependencies between constituents on the surface syntactic level are called *surface slots* which are language-dependent. The correspondence between surface and semantic relation is called *diathesis*.

Along with tree dependencies, constituents can be linked with *non-tree relations* such as conjunction, anaphora, control and movement.

Syntactic structure



Semantic structure



For this model we have developed descriptions of semantic hierarchy, syntactic paradigms (surface slots with government, agreement, order restrictions and relations of slots and grammatical features. The descriptions distinguish between allowed and not allowed structures. There is no much emphasis on disambiguation of allowed structures.

Now we can reconsider translation process as conversion from source text to target via source surface structure, semantic structure, target surface structure to target text.

Since the model is ambiguous, we can treat this process as probabilistic and try to estimate conditional probabilities of the model features.

The probabilistic model includes:

- Lexeme & POS ngram probabilities
- Lexical class probability
- Lexicalized surface dependency probability
- Lexicalized semantic dependency probability
- Surface to sematic slot mapping
- Surface slots ngram probabilities
- Lexical classes co-occurrence probabilities
- Lexical class translation probability
- Surface slot translation probability

We use Bayesian approach to construct probability from different components. Taking into consideration the unprepared part of the audience of the conference we provide explanations instead of formulas.

Lexeme & POS ngram probabilities

This is a traditional language model, except that we take lexeme+part of speech instead of words. It is used to guide search on initial stages of parsing and in cases of incomplete parse trees. Currently we use 3-grams.

Lexical class probability

Lexical class probability differs significantly between various domains (e.g. meanings of word “file<noun>” in such domains as Law, Manufacturing or Information technologies). Thus, for the whole text we detect possible domains and calculate conditional

probability of different meaning of the words (lexical classes) for the Bayesian mixture of domains. For example, if we try to determine SC for the source lexeme *file* in the text for which we have established domain Information technologies, it is more realistic to choose the LC "file:FILE" (file as set of related data in computer). On the opposite, of we deal with the text labeled as Manufacturing domain, it is more probable that we have "file:FILE_AS_TOOL" ("a hand tool which is used for rubbing hard objects").

Processing of the whole text slightly improves precision of analysis and translation in comparison to sentence by sentence mode.

Lexicalized dependency probability

Lexicalized dependency probability (either surface or semantic) is a probability of the dependency link in the parse tree conditioned on lexical classes of parent and child. Currently there are ca. 500 dependency labels and more than 100K lexical classes. It means we have to learn more than $5 \times 1{,}012$ parameters.

Although many combinations are prohibited by the model, still their number is huge in comparison to the volume of available parallel corpora (~1G of words).

We use hierarchy to approximate parameters.

For example, if we try to determine the correct SC for *run* in the sentence like "I need to run the clock", we receive information from our hierarchy that *clock* is a device (the hole path up the tree is CLOCK: TIMEPIECE: DEVICE_FOR_MEASURING_AND_COUNTING: DEVICE), and we know that the class DEVICE is statistically good combined with the class " TO_ACTIVATE", so it is more reliable to choose "run:TO_ACTIVATE".

Lexicalized dependency probability is crucial for determination of the correct parsing tree and disambiguation of word senses.

Surface to sematic slot mapping

To select semantic slot for surface slot at analysis and to select surface slot for semantic slot at synthesis we collect co-occurrence data for surface and semantic slots.

Lexical classes co-occurrence probabilities

Domain depended lexical class probability provides only a rough adaptation to a particular large-scale domain. There are words, which senses do not correlate with easy identifiable domains or are indistinguishable within one, or there is no much text to identify domain and the dependency context is neutral. For example, in the sentence "Washington criticized Syria." we need to distinguish between the city and the surname (this difference does not influence translation, but is important for other applications of parsing). In this case co-occurrence of classes can help determine the right analysis if from the training data we know that Washington as a person had little to do with Syria.

Co-occurrence of classes is computed for siblings in dependency tree, for all words with limited neighborhood and for conjuncted words. Just as for the dependencies the number of parameters is quadratic to number of class. Here the approximation with hierarchy is used as well.

Lexical classes translation probability

Although the model was originally planned to have rich semantic features (*semantemes*) for differentiation between synonyms of one semantic class across languages, in practice we augmented it with conditional probability of synonym in target language for the give synonym in source language.

Surface slot translation probability

In theory, surface slot selection at target language must be guided by source semantic slot and features of child and parent constituents. But it turns out that it is not possible to take into account all cases in the model. Thus we use as well probabilistic model which estimates target diathesis probability by source surface slot and complexity of child subtree.

Hierarchical approximation of lexicalized pairwise correlations

Here we present our method of computing co-occurrence statistics in case of lack of data by using semantic hierarchy.

The co-occurrence we need to compute is conditional probability

$$\log \frac{P(A \cap B)}{P(A)P(B)},$$

where A and B are two lexical classes. In case we have enough data we can use counts to calculate this value

$$\log \frac{N(A \cap B)N}{N(A)N(B)},$$

But for many class pairs $N(A \cap B)$ is either very small (which makes very unreliable estimations) or zero. The required probability can be decomposed with the use of hierarchy: , where — is *i*th ancestor of A

$$\prod_{\substack{n=0\ldots L-1 \\ m=0\ldots K-1}} \frac{N(A^{(n)} \cap B^{(m)})N(A^{(n+1)} \cap B^{(m+1)})}{N(A^{(n)} \cap B^{(m+1)})N(A^{(n+1)} \cap B^{(m)})},$$ where $A^{(i)}$ — is i^{th} ancestor of A

Thus we can use counts of events for the classes in higher levels of hierarchy. These counts of superclasses are larger and give more accurate estimates of probability.

Training the probabilistic model

To train the model we have to have correct parse trees to estimate probabilities of model components. There is no such resource of adequate size. To cope with this problem we use parallel corpora and the parse trees are “hidden variables”.

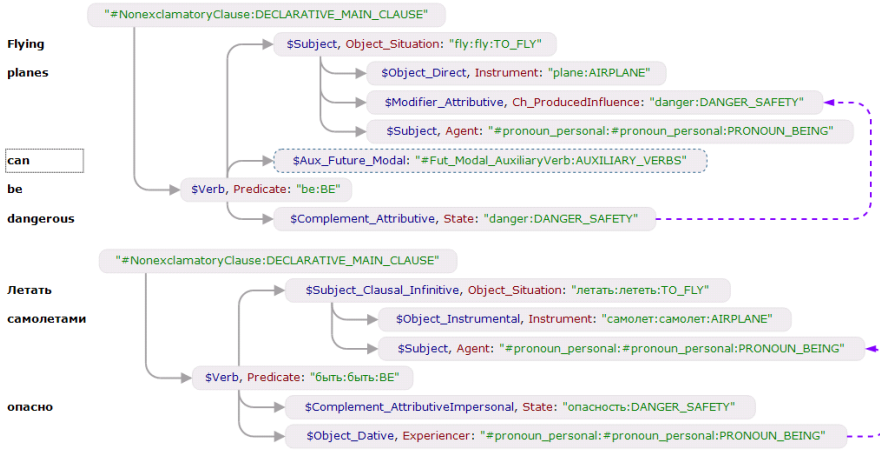
To make it work we need to have alignment of trees and a way to generate aligned parse trees. Alignment model is very simple — we condition the probability of alignment on, distance within hierarchy, on whether there are the same dependencies in aligned trees, and on the order of aligned constituents. To correctly handle lists of out-of-dictionary words (for example named entities) we also compute transliteration distance for such words.

To guess about hidden variable, that is presumably correct parse trees, we modified our parsing algorithm in the following way:

- We align two dependency graphs and attribute more weight to aligned constituents and links.
- We generate parse trees from the two graphs. They are generated by order of diminishing probability of parse structure to be correct and to produce the available translation.
- We align pairs of parse trees and select best trees (both by parsing and alignment quality).
- For further parameter estimation we utilize several generated trees to mitigate overfitting to erroneous parsing results.

See the example below on how the universal semantic structure and the parallel analysis help disambiguated classical case.

Parallel semantic structures



Resent research is concentrated on computing probabilistic model parameters for other linguistic descriptions such control, movement and ellipsis.

We also experimenting with non-Bayesian estimation of parameters, since Bayesian approach assumes independence of features which is hard to achieve.

Out of model translation

It is not feasible to cover complex, huge and dynamic languages by manual model. Two problems that we see are:

- There are too many words.
- Many contextual translations go across the hierarchy.

To cope with the first problem we have introduced a special lexeme for unknown words. We predict the morphological features of unknown (to our system) word by making hypothesis about its flexion. Unknown word lexeme is mapped to different places in the hierarchy, thus we also try to guess the rough meaning of the word, e.g. person, action, artifact.

At present, we either transliterate the unknown word or keep them untranslated. We could as well mine possible translation from alignments of parallel corpora.

The second problem is that some words in some context are translated in the adjacent or sometimes very far lexical classes of hierarchy (e.g. power plant — [электро] станция). As with phrase-based statistical machine translation we automatically capture regular out-of-hierarchy translations and use them as collocations. In comparison to phrases in SMT and collocations in traditional dictionaries our collocations are parse tree fragments. For more about mining the collocation, see [7].

To achieve the good quality of translation, comparable to popular online services, the system should be trained on huge, kept up-to-date internet corpora. Currently we train our system on roughly 10^8 sentences.

Evaluation

Internal evaluations

Internal evaluation is performed on several parallel and marked-up corpora.

We use modified BLEU to estimate translation quality. To our opinion, this variant of BLEU is more suitable for flective languages — only 1-gramms are matched literally, while higher-order n-gramms are reduced to lemmas. Absolute BLUE-score is very dependent to the corpora and to the system. For us it is 0.15-0.20. We rely on it to control incremental changes in the model and the algorithm.

Some corpora are partially marked-up with surface and semantic dependencies and lexical classes. We control the sentence level precision which is within 60-80%.

We also have small internal stand-out corpora to manually estimate and compare the translation quality with other systems.

External evaluations

It is hard to compare parsing performance of different systems if they are based on different linguistic principles. Anyway such attempt has been done at previous Dialog conferences. In [3] the part of speech disambiguation was tested (which indirectly correlates with parsing performance if parsing is used for this purpose). In [2] the parsing structures of different systems have been manually compared with a certain degree of freedom to match different approaches to the syntax. In both evaluations the system has shown good results.

This year the translation quality is estimated by range of automatic scores and by manual translation. We have achieved good results in both comparisons. The system was run in per-sentence mode without utilizing surrounding context. Although this context was available we were not able to use due to technical problems.

Conclusion

The development of the system and its good results in evaluations proves the plausibility of the linguistically oriented model-based approach to natural language processing. Due to the universality of the model, it can be used in many NLP tasks. Trained on the parallel corpora it can then perform translation, parsing, word sense disambiguation.

References

1. *Anisimovich, Druzhkin, Minlos, Petrova, Selegey, Zuev,* 2012. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. Proceedings of Dialog 2012 (pp. 80–103), Moscow, Russia.
2. *Toldova S. Ju.* et al. NLP evaluation 2011–2012: Russian syntactic parsers. Proceedings of Dialog 2012, Moscow, Russia.
3. *Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A., Garejshina A.,Grishina Ju., D'yachkov V., Ionov M., Koroleva A., Kudrinsky M., Lityagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., Koval' S.* (2010), Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka [NLP evaluation: Russian morphological parsers], in Computational linguisticsand intellectual technologies. Proceedings of International Workshop Dialogue' 2010, Vol. 9 (16), Moscow, pp. 318–326.
4. *David Chiang.* 2005. A hierarchical phrase-based model for statistical machine translation. In Proceeding of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 263–270.
5. *Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst.* 2007. Moses: Open source toolkit for statistical machine translation. In 45th Annual Meeting of the Association for Computational Linguistics.
6. *Philipp Koehn, Barry Haddow, PhilipWilliams, and Hieu Hoang.* 2010. More linguistic annotation for statistical machine translation. In Proceedings of the Joint FifthWorkshop on Statistical Machine Translation and MetricsMATR, pages 115–120.
7. *Novitskiy, V. I.* Automatic retrieval of parallel collocations / V. I. Novitskiy // Pattern Recognition and Machine Intelligence / Ed. by S. Kuznetsov, D. Mandal, M. Kundu, S. Pal. — Vol. 6744 of Lecture Notes in Computer Science.— Moscow, Russia: Springer, 2011. — July. — pp. 261–267

Abstracts

RESEARCH OF LEXICAL APPROACH AND MACHINE LEARNIG METHODS FOR SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com), **Klekovkina M. V.** (klekovkina.mv@gmail.com), **Kotelnikov E. V.** (kotelnikov.ev@gmail.com), **Pestov O. A.** (oleg.pestov@gmail.com), Vyatka State Humanities University, Kirov, Russia

Methods and approaches used by the authors to solve the problem of sentiment analyses on the seminar ROMIP-2012 are described. The lexical approach is represented with the lexicon-based method which uses emotional dictionaries manually made for each domain with the addition of the words from the training collections. The machine learning approach is represented with two methods: the maximum entropy method and support vector machine. Text representation for the maximum entropy method includes the information about the proportion of positive and negative words and collocations, the quantity of interrogation and exclamation marks, emoticons, obscene language. For the support vector machine binary vectors with cosine normalization are built on texts. The test results of the described methods are compared with those of the other participants of the ROMIP seminar. The task of classification of reviews for movies, books and cameras is investigated. On the whole. The lexical approach demonstrates worse results than machine learning methods , but in some cases excels it. It is impossible to single out the best method of machine learning: on some collections maximum entropy method is preferable, on others the support vector machine shows better results.

SEMANTIC REPRESENTATION FOR NL UNDERSTANDING

Boguslavsky I. M. (Igor.M.Boguslavsky@gmail.com), Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; Universidad Politécnica de Madrid, Madrid, Spain, **Dikonov V. G.** (dikonov@iitp.ru), **Iomdin L. L.** (iomdin@iitp.ru), **Timoshenko S. P.** (timoshenko@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

While mainstream semantic parsing mostly consists in word sense disambiguation, semantic role labeling and assigning WordNet/FrameNet categories, deeper NL understanding requires much more. It includes understanding of the meaning of words, extralinguistic knowledge and is based on a more intricately elaborated representation of this meaning than that provided by standard resources. For example, the semantic model should not only know that *ask for, implore* and *demand* belong to the same REQUEST frame. It should also formally represent the very idea of an incentive speech act (e.g. 'X tells Y that he wants him to do Z') and even the difference between such request varieties as represented by the words listed. Our aim is to build a semantic analyzer supplied with this kind of semantic knowledge and capable of constructing semantic representations that convey this knowledge and can be used for inferences. However, before constructing a parser, one should define the target representation. The focus of this paper is to propose a semantic representation richer than usually considered. Since the depth of representation is an important decision in language modeling, the topic deserves a detailed discussion. Our paper demonstrates selected NL phenomena untreatable by state-of-the-art parsers and semantic representations proposed for them.

ROMIP MT EVALUATION TRACK 2013: ORGANIZERS' REPORT

Braslavski P. (pbras@yandex.ru), Kontur labs; Ural Federal University, Russia,
Beloborodov A. (xander-beloborodov@yandex.ru), Ural Federal University, Russia,
Sharoff S. (s.sharoff@leeds.ac.uk), University of Leeds, Leeds, UK,
Khalilov M. (maxim@tauslabs.com), TAUS Labs, Amsterdam, Netherlands

The paper presents the settings and the results of the ROMIP 2013 machine translation evaluation campaign for the English-to-Russian language pair. The quality of generated translations was assessed using automatic metrics and human evaluation. We also demonstrate the usefulness of a dynamic mechanism for human evaluation based on pairwise segment comparison.

SENTIMENT ANALYSIS TRACK AT ROMIP 2012

Chetviorkin I. I. (ilia2010@yandex.ru), **Loukachevitch N. V.** (louk_nat@mail.ru),
Lomonosov Moscow State University, Moscow, Russia

In 2012, Russian Information Retrieval Seminar (ROMIP) continued the investigation of sentiment analysis issues. Along with the last year's tasks on sentiment classification of user reviews we proposed two new tasks on sentiment classification of news-based opinions and query-based extraction of opinionated blog posts. For all tasks new test collections were prepared. The paper describes the characteristics of the collections, track tasks, the labeling process, and evaluation metrics. We summarize the participants' results and describe our simple approach for sentiment extraction task.

COMBINING HMM AND UNIT SELECTION TECHNOLOGIES TO INCREASE NATURALNESS OF SYNTHESIZED SPEECH

Chistikov P. G. (chistikov@speechpro.com), **Korolkov E. A.** (korolkov@speechpro.com),
Talanov A. O. (andre@speechpro.com), Speech Technology Center ltd, St. Petersburg, Russia

We propose a text-to-speech system based on the two most popular approaches: statistical speech synthesis (based on hidden Markov models) and concatenative speech synthesis (based on Unit Selection). TTS systems based on Unit Selection generate speech that is quite natural but highly variable in quality. On the other hand, statistical parametric systems produce speech with much more consistent quality but reduced naturalness due to their vocoding nature. Combining both approaches improves the overall naturalness of synthesized speech. To reduce variability of Unit Selection results, we calculate a statistical generalization of the speaker's intonation. We created a methodology of voice model building in order to solve the task of speech parameterization. The model is a set of HMM models whose state parameters are clustered to provide good quality of synthesized speech even under conditions of insufficient training data. MFCC coefficients, pitch, energy and duration values are used as fundamental features. Objective and subjective experiments show that our method increases the naturalness of synthesized speech.

CREATING AN AUTOMATED SYSTEM FOR TRANSLATION OF USER-GENERATED CONTENT

Evdokimov L. V. (Leonid.Evdokimov@promt.ru),
Molchanov A. P. (Alexander.Molchanov@promt.ru), PROMT Ltd., Saint-Petersburg, Russia

This paper describes fast implementation of a hybrid automated translation system for processing user-generated content. We report on engine customization for TripAdvisor, the world's largest travel website. Due to the growing potential of the Russian travel market, TripAdvisor created the Russian version of its website and decided to translate all English reviews into Russian. PROMT, a leading provider of industrial MT solutions, was selected as MT vendor for the English-Russian language pair. According to the client's request we had to perform customization within a short period. All input data represent user-generated content, so we faced several problems while building a large-scale, robust, high-quality engine. We decided to create a solution based on a hybrid machine translation system for the hybrid approach makes possible fast and efficient customization of a translation system with little or none in-domain data. We automatically crawled a large web-based Russian text corpus of tourist reviews to build a statistical language model for our hybrid translation system. We analyzed a batch of tourist reviews in English provided by TripAdvisor, created a number of dictionaries, a translation memory and defined translation rules for user-generated content. To handle the problem of various typos and misspellings we added most frequent misspelled words and phrases to the created dictionaries. We experimented on a test set of tourist reviews in English provided by TripAdvisor. We report on improvements over our baseline system output both by automatic evaluation metrics and linguistic expertise.

USING SEMANTIC FILTERS IN APPLICATION TO BOOK REVIEWS SENTIMENT ANALYSIS

Frolov A. V. (anton_frolov@rco.ru), **Polyakov P. Yu.** (pavel@rco.ru),
Pleshko V. V. (vp@rco.ru), RCO LLC, Moscow, Russian Federation

The paper studies the use of fact semantic filters in application to sentiment analysis of book reviews. The tasks were to divide book reviews into 2 classes (positive, negative) or into 3 classes (positive, negative, and neutral). The main machine learning pitfalls concerning sentiment analysis were classified and analyzed.

USING STATISTICAL METHODS FOR PROSODIC BOUNDARY DETECTION AND BREAK DURATION PREDICTION IN A RUSSIAN TTS SYSTEM

Khomitsevich O. (khomitsevich@speechpro.com), **Chistikov P.** (chistikov@speechpro.com),
Speech Technology Center Ltd, St. Petersburg, Russia

The paper deals with statistical methods for predicting positions and durations of prosodic breaks in a Russian TTS system. We use CART and Random Forest classifiers to calculate probabilities for break placement and break durations, using grammatical feature tags, punctuation, word and syllable counts and other features to train the classifier. The classifiers are trained using a large high-quality speech database consisting of read speech. The experimental results for prosodic break prediction shown an improvement compared to the rule-based algorithm currently integrated in the VitalVoice TTS system; the Random Forest classifier shows the best results, although the large size of the model makes it more difficult to use in a commercial TTS system. To make the system more flexible and deal with the remaining break placement errors, we propose combining probabilities and rules in a working TTS system, which is the direction of our future research. We observe good results in experiments with predicting pause durations. A statistical model of break duration prediction has been implemented in the TTS system in order to make synthesized speech more natural.

TESTING RULES FOR A SENTIMENT ANALYSIS SYSTEM

Kuznetsova E. S. (knnika@yandex.ru), GK "Geostream", Moscow, Russia,
Loukachevitch N. V. (louk_nat@mail.ru), **Chetviorkin I. I.** (ilia2010@yandex.ru),
Lomonosov Moscow State University, Moscow, Russia

The paper is devoted to testing rules useful for sentiment analysis of Russian. First, we describe the working principles of the POLYARNIK sentiment analysis system, which has an extensive sentiment dictionary but a minimal set of rules to combine sentiment scores of opinion words and expressions. Then we present the results achieved by this system in ROMIP-2012 evaluation where it was applied in the sentiment analysis task of news quotes. The analysis of detected problems became a basis for implementation of several new rules, which were then tested on the ROMIP-2012 data.

BILINGUAL SPEECH RECOGNITION WITHOUT PRELIMINARY LANGUAGE IDENTIFICATION

Lyudovyk T. V. (tetyana.lyudovyk@gmail.com),
Pylypenko V. V. (valeriy.pylypenko@gmail.com), International Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine

We present an approach to speaker-independent recognition of large-vocabulary continuous speech characterized by code-switching between Ukrainian and Russian. The approach does not require language boundary detection or language identification. Special speech and text corpora are not needed to train acoustic and linguistic models. The approach takes into account peculiarities of phonetic systems of Russian and Ukrainian languages. A cross-lingual speech recognition system is developed. A previously developed acoustic model of Ukrainian speech serves for both languages. A set of HMM-models representing 54 Ukrainian phonemes and several non-speech units such as breath, fillers and silence are used. Bilingual linguistic model is trained on a set of Ukrainian and Russian texts. Pronunciation lexicon combines word

forms in both languages. Phonemic transcription of Russian word forms are generated using Ukrainian phonemes. Recognition post-processing can be applied to smooth recognized word sequences by using a dictionary containing Ukrainian and Russian words which sound equally but are written differently. The proposed approach can be applied to the recognition of bilingual speech with between-phrase and within-phrase code-switching. Developed cross-lingual speech recognition system was tested on Ukrainian, Russian, and Ukrainian-Russian speech of one bilingual speaker. Preliminary results show that the proposed approach could achieve a good performance. Accuracy of mixed speech recognition is lower only by 3–7% as compared with monolingual speech recognition accuracy.

EXTRACTING PRODUCT FEATURES FROM REVIEWS WITH THE USE OF INTERNET STATISTICS

Marchuk A. A. (aamarchuk@gmail.com), St. Petersburg State University,
Ulanov A. V. (alexander.ulanov@hp.com), Hewlett-Packard Labs Russia,
Makeev I. V. (ilya.makeev@gmail.com), Saint Petersburg State University of Information Technologies, Mechanics and Optics, **Chugreev A. A.** (artemij.chugreev@gmail.com), St. Petersburg Polytechnical University, Russia

The paper studies the task of extracting product features from reviews. We consider this task as a classification problem and propose a number of classification features. These features are computed using different statistics returned by queries to Yandex search engine, the Internet library and the Russian National Corpus. To justify our approach, we create and manually label a product features dataset, compute the proposed classification features and conduct classification experiments. The results produced by various classifiers applied to different subsets of the data show the feasibility of our approach. We also look at the usefulness of the proposed classification features.

AUTOMATIC EVALUATION OF MACHINE TRANSLATION QUALITY

Màrquez L. (lluism@lsi.upc.edu), TALP Research Center Software Department Technical University of Catalonia

This paper contains an extended abstract of the invited talk presented, with the same title, at Dialogue 2013, the 19th International Computational Linguistics Conference. The presentation will cover several works carried out at the Technical University of Catalonia (UPC) in collaboration with several researchers and colleagues. I would like to especially mention the following: Jesús Giménez, Meritxell Gonzàlez, Lluís Formiga and Laura Mascarell. Sincere thanks to all of them.

USING BASIC SYNTACTIC RELATIONS FOR SENTIMENT ANALYSIS

Mavljutov R. R. (m-ceros@yandex.ru), **Ostapuk N. A.** (nataxane@yandex.ru), Yandex, Moscow, Russia

The paper describes a rule-based approach to sentiment analysis. The developed algorithm aims at classifying texts into two classes: positive or negative. We distinguish two types of sentiments: abstract sentiments, which are relevant to the whole text, and sentiments referring to some particular object in the text. As opposed to many other rule-based systems, we do not regard the text as a bag of words. We strongly believe that such classical method of text processing as syntactic analysis can considerably enhance sentiment analysis performance. Accordingly, we first parse the text and then take into account only the phrases that are syntactically connected to relevant objects. We use the dictionary to determine whether such a phrase is positive or negative and assign it a weight according to the importance of the object it is connected with. Than we calculate all these weights and some other factors and decide whether the whole text is positive or negative. The algorithm showed competitive results at ROMIP track 2012.

IMPACT OF DIFFERENT TYPES OF SPELLING MISTAKES ON THE QUALITY OF STATISTICAL MACHINE TRANSLATION

Mescheryakova E. M. (mescheryakova@yandex-team.ru),
Galinskaya I. E. (galinskaya@yandex-team.ru), **Gusev V. Yu.** (vgoussev@yandex-team.ru),
Shmatova M. S. (mashashma@yandex-team.ru), Yandex, Moscow, Russia

Errors in the original text will most probably affect the quality of machine translation. It would be interesting to see how different types of errors can influence the translation. To do this, we selected three sets of 500 random queries in English, German and Polish. In each set we corrected different types of errors: 1) missing diacritical marks (except English); 2) all misprints (including diacritics); 3) errors in punctuation and use of capitals; 4) all types of errors listed in 1)–3). As a result we had five sets of 500 queries for German and Polish and four sets for English. Then we translated all the sets into Russian using three free online statistical machine translation systems and compared their BLEU scores to see how they increase in corrected tests as compared to the original ones. We also used different types of BLEU: along with the usual one, which treats punctuation signs as words, we used simplified BLEU which disregards punctuation, and also extended BLEU which takes into consideration both punctuation and use of capitals. We show that in a fully corrected text BLEU increases by approx. 10–15% as compared to original sets. Correcting each of the two main types of errors — misprints and punctuation/capitalization — gives an increase of 5–10% each depending on the language and on the peculiarities of the test sets. On the other hand, correcting only diacritics has very small impact on the translation quality: close to zero in German and 0,5–1% in Polish.

ATEX: A RULE-BASED SENTIMENT ANALYSIS SYSTEM PROCESSING TEXTS IN VARIOUS TOPICS

Panicheva P. V. (ppolin86@gmail.com), EPAM Systems, Saint-Petersburg, Russia

ATEX is a rule-based sentiment analysis system for texts in the Russian language. It includes full morpho-syntactic analysis of Russian text, and highly elaborated linguistic rules, yielding fine-grained sentiment scores. ATEX is participating in a variety of sentiment analysis tracks at ROMIP 2012. The system was tuned to process news texts in politics and economy. The performance of the system is evaluated in different topics: blogs on movies, books and cameras; news. No additional training is performed: ATEX is tested as a universal 'ready-to-use' system for sentiment analysis of texts in different topics and different classification settings. The system is compared to a number of sentiment analysis algorithms, including statistical ones trained with datasets in respective topics. Overall system performance is very high, which indicates high usability of the system to different topics with no actual training. According to expectations, the results are especially good in the 'native' political and economic news topic, and in the movie blog topic, proving both to share common ways of expressing sentiment. With regard to blog texts, the system demonstrated the best performance in two-class classification tasks, which is a result of the specific algorithm design paying more attention to sentiment polarity than to sentiment/neutral classes. Along these lines areas of future work are suggested, including incorporation of a statistical training algorithm.

EVALUATION OF NATURALNESS OF SYNTHESIZED SPEECH WITH DIFFERENT PROSODIC MODELS

Solomennik A. I. (solomennik-a@speechpro.com), Speech Technology Ltd., Minsk, Belarus,
Chistikov P. G. (chistikov@speechpro.com), Speech Technology Center Ltd, St. Petersburg, Russia

Obtaining natural synthesized speech is the main goal of modern research in the field of speech synthesis. It strongly depends on the prosody model used in the text-to-speech (TTS) system. The paper deals with speech synthesis evaluation with respect to the prosodic model used. Our Russian VitalVoice TTS is a unit selection concatenative system. We describe two approaches to prosody prediction used in VitalVoice Russian TTS. These are a rule-based approach and a hidden Markov model (HMM) based hybrid approach. We conduct an experiment for evaluating the naturalness of synthesized speech. Four variants of synthesized speech depending on the applied approach and the speech corpus size were tested. We also included natural speech

samples into the test. Subjects had to rate the samples from 0 to 5 depending on their naturalness. The experiment shows that speech synthesized using the hybrid HMM-based approach sounds more natural than other synthetic variants. We discuss the results and the ways for further investigation and improvements in the last section.

CONTEXT-DEPENDENT OPINION LEXICON TRANSLATION WITH THE USE OF A PARALLEL CORPUS

Ulanov A. V. (alexander.ulanov@hp.com), **Sapozhnikov G. A.** (gsapozhnikov@gmail.com),
Hewlett-Packard Labs Russia, St. Petersburg State University, Russia

The paper deals with multilingual sentiment analysis. We propose a method for projecting an opinion lexicon from a source language to a target language with the use of a parallel corpus. We can make sentiment classification in a target language using an opinion lexicon even if we have no labeled dataset. The advantage of our method is that it captures the context of a word and thus produces a correct translation of it. We apply our method to the language pair English-Russian and conduct sentiment classification experiments. They show that our method allows creating high-quality opinion lexicons.

STATISTICAL MACHINE TRANSLATION WITH LINGUISTIC LANGUAGE MODEL

Zuyev K. A. (konst@abbyy.com), **Indenbom E. M.** (Eugene_I@abbyy.com),
Yudina M. V. (Maria_Yu@abbyy.com), ABBYY, Moscow, Russia

Stemming from traditional “rule based” translation a “model based” approach is considered as an underlying model for statistical machine translation. This paper concerns with training on parallel corpora and application of this model for parsing and translation.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 12 (19). 2013

Том 2. Доклады специальных секций

Ответственный за выпуск **А. А. Белкина**
Вёрстка **К. А. Климентовский**

Подписано в печать 14.05.2013
Формат 152 × 235
Бумага офсетная
Тираж 250 экз. Заказ № 553

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9