Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

EVENT2MIND FOR RUSSIAN: UNDERSTANDING EMOTIONS AND INTENTS IN TEXTS. CORPUS AND MODEL FOR EVALUATION

Fenogenova A. S. (alenush@gmail.com)
Sberbank, Moscow, Russia

Tikhonova M. I. (m_tikhonova94@.mail.ru)
Sberbank; National Research University Higher School of Economics, Moscow, Russia

Filipetskaya D. V. (dafi913@yandex.ru)
Moscow Institute of Physics and Technology, Moscow, Russia

Mironenko F. D. (fomius2000@yandex.ru)
Saint Petersburg State University, Saint-Petersburg, Russia

Tabisheva A. O. (anastasiatabisheva@yandex.ru)
National Research University Higher School of Economics, Moscow, Russia

The paper provides a comprehensive overview of the corpus for the Russian language for the commonsense inference task. Namely, we construct event phrases, which cover a wide range of everyday situations with labelled intents and reactions of the event main participant and emotions of other people involved. The dataset consists of two parts: a crowdsourced corpus of 6,756 examples from Russian sources and a translated into Russian part of the original corpus of 23,409 examples. Apart from this, we use the collected data in order to train the event2mind model for the Russian language. The paper presents careful description of the best Russian model and the results of the conducted experiments.

Key words: event2mind, chatbots, emotion detection, intents generation, Natural Language Processing, dialog systems

DOI: 10.28995/2075-7182-2020-19-299-309

EVENT2MIND ДЛЯ РУССКОГО ЯЗЫКА: КОРПУС И МОДЕЛЬ — ПОНИМАНИЕ ЭМОЦИЙ И ИНТЕНДОВ В КОРОТКИХ ТЕКСТАХ

Феногенова А. С. (alenush@gmail.com)
Сбербанк, Москва, Россия

Тихонова М. И. (m_tikhonova94@.mail.ru)
Сбербанк; Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Филипецкая Д. В. (dafi913@yandex.ru)
Москвоский физико-технический институт, Москва, Россия

Мироненко Ф. Д. (fomius2000@yandex.ru)
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Табишева А. О. (anastasiatabisheva@yandex.ru)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Распознавание интента (намерения) субъекта является естественным для живого человека и весьма сложной задачей для компьютера. В данной работе представлен корпус для русского языка для задачи распознавания намерения субъекта: по полученному на вход короткому тексту-событию определяются причины, по которым субъект совершил действие, а также эмоции субъекта и других участников события. Формат корпуса соответствует формату оригинального корпуса на английском языке. Собранный корпус состоит из двух частей: размеченный русский корпус из 6756 примеров и переведенная автоматическим переводчиком с английского отфильтрованная часть английского корпуса из 23 409 примеров. Помимо этого был проведен ряд экспериментов по обучению модели для русского языка и получена модель, сравнимая по качеству с английской. Это доказывает воспроизводимость алгоритма для языков с более сложным по сравнению с английским морфологическим составом.

Ключевые слова: event2mind, чатботы, диалоговые системы, детектирование эмоций, Natural Language Processing

1. Introduction

Common sense reasoning tasks have received significant attention in Natural Language Processing and attempts [2], [8], [12], [14], [17], [18] to solve them have been made in recent years. Such type of automatic pragmatic reasoning can be useful for a wide range of NLP applications that require anticipation of people's reasoning and emotions. For instance, incorporating such model into a dialog system could make its reactions more emphatic and appealing for users. Moreover, commonsense reasoning is often regarded as a necessary step towards human understanding.

Most of the proposed models [2], [12], [17], [18] are supervised and, therefore, require large datasets, often with quite nontrivial markup, for training. Constructing such datasets is difficult and expensive. For English there exist several corpora[1] [12], [15], [16] for commonsense reasoning tasks. For Russian, however, the situation is not so good. The absence of data is one of the main obstacles to adaptation most of the models for the Russian language.

Common sense reasoning problem could be stated in different ways. The one considered in the paper, originally proposed in [12], is formulated as follows: given a short free-form text describing an event (*"PersonX eats breakfast in the morning"*) a model makes reasoning about the agent's intents (*"X wants to satisfy hunger"*), reactions (*"X feels satiated, full"*) and possible reactions of the other event's participants.

In [12] the authors presented their model, *event2mind*, which supports commonsense inference on events phrases. In their work a specific focus was made on modeling stereotypical intents and reactions of people. Another contribution of their work was a crowdsourced corpus that supports commonsense inference in English language, which is publicly available[2].

The main difficulty in adapting event2mind model for the Russian and other languages is that it requires a large corpus of the event phrases with labelled intents and reactions for training. Thus, in order to train a model for any other language besides English it is necessary to construct a dataset.

The main contribution of the paper is a text corpus suitable for event2mind training in Russian which consists of two parts:

1. 6,756 event phrases covering a diverse range of everyday events and situations in Russian,
2. a subset of 23,409 event phrases from English corpus translated via Google translator[3].

In addition to that, we used the obtained corpus in order to train Russian model of event2mind. The article provides a careful description of the conducted experiments with different model's versions and configs and presents the best one.

---

[1] http://nlpprogress.com/english/common_sense.html

[2] https://uwnlp.github.io/event2mind/data/

[3] https://translate.google.ru

2. Dataset

One of the goals of the work was to collect a labelled corpus of short free-form texts, which are further referred to as *event phrases*, in Russian in the format suitable for event2mind training. We define events like authors in the original work and try to collect phrases that contain a diverse range of common everyday events and situations.

2.1. Crowdsourced corpus

First step was to collect a sufficient amount of events. For this purpose we gathered texts from several sources:

1. short episode descriptions of TV serias and soap operas. We manually selected 50 TV serias from the Internet portal KisTV[4] making the focus on series about present everyday life situations. In total descriptions of 50 TV series were used among which are Friends, Sex and the City, Santa Barbary, Univer (Универ), Kitchen (Кухня) and others. We did not use fantasy, science fiction or series about medicine as long as they contain a lot of specific words which are not commonly used.
2. Book summaries from Briefly.[5] Total number of downloaded summaries equals 1,512.
3. Texts from SynTagRus corpus[6] [19], [3]—a subcorpus of Russian National Corpus, with fiction and news, with manual syntactic annotation.

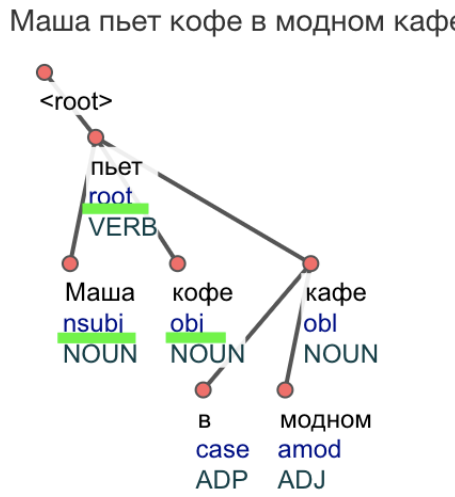

Figure 1: Events in the tree are green-labelled

[4] http://kistv.ru/

[5] http://briefly.ru

[6] https://github.com/UniversalDependencies/UD_Russian-SynTagRus

From these three sources event phrases are extracted. We select event phrases as a combination of a verb predicate with partially instantiated arguments, like in the **Figure 1**. Events from row texts are derived using UDPipe [13] syntactic parser, model trained on Russian Syntagrus, version UD2.4. Namely, we automatically search for syntactic patterns of verb with its arguments in a syntactic tree that satisfy one of the following rules:

- *nsubj* + *root* + *obj*
- *nsubj* + *root* + *iobj*
- *nsubj* + *advmod* + *root*
- *nsubj* + *root* + *case* + *obl*
- etc.

Similar to the original paper we introduce type variables for generalisation of event phrases. In particular, predicate subjects corresponding to pronouns and name entity mentions are replaced with typed variables such as PersonX, PersonY or PersonZ (if there is more than one person in the event). Named Entity Recognition is conducted to replace named entities with the aliases according to the grammatical information from UDpipe. Namely, we have considered all the tokens that match the syntactic pattern above marked with tags PROPN and PRONOUN. Then we work only with text fragments that matched the rule patterns. In addition, phrases which do not contain inanimate subjects are also filtered using grammar information from UDpipe. Following the original work our corpus contains only events with person named entities.

After the depersonalization frequency analysis and Levenshtein distance are used in order to select most common event phrases and to filter nonstandard examples which occur only once. First, we take all the event phrases which occur more than once. For the rest of the data for every pair of event phrases pairwise Levenshtein distance $L(phrase_1, phrase_2)$ is computed and for every pair with $L(phrase_1, phrase_2) \leq 5$ the shorter phrase is chosen for the final dataset. Thus, we obtain 4523 unique event phrases.

2.2. Crowdsourcing

In order to annotate raw events with intents and reactions we design a Yandex. Toloka[7] task supplemented with annotation instruction. A snippet of the task is shown in **Figure 2**. For each event phrase we ask:

- *weather it contains a meaningful event,*
- *is it possible to find some reasons that cause such event,*
- *can subjects of the event have some emotions and reactions to it.*

In addition to that we provide possible answers generated by the trial event-2mind model trained only on automatically translated part of the dataset. This model performs poorly, nevertheless in some cases it generates reasonable answers.

In case of positive answer for the first question in the toloka's task we ask annotators to verify trial model's answers and then to give their own variants for possible intents and reactions of the agent (PersonX) and reactions of other event participants if any.

[7] https://toloka.yandex.ru/

We encourage annotators to write more than one variant for every event phrase and, in addition, we get verified or corrected variant from the trial model. Thus, we get at least two valid answers for an example. We set the toloka's task with annotator overlap equal 3. Then we drop the events that were considered to be wrong by at least 2 annotators and set label "none" if in the example there were only intents and no reactions or vice versa.

It should be mentioned that for label standardization we recommend annotators to use infinitive word forms such as хотеть есть (be hungry), быть счатливым (be happy), грустный (sad).

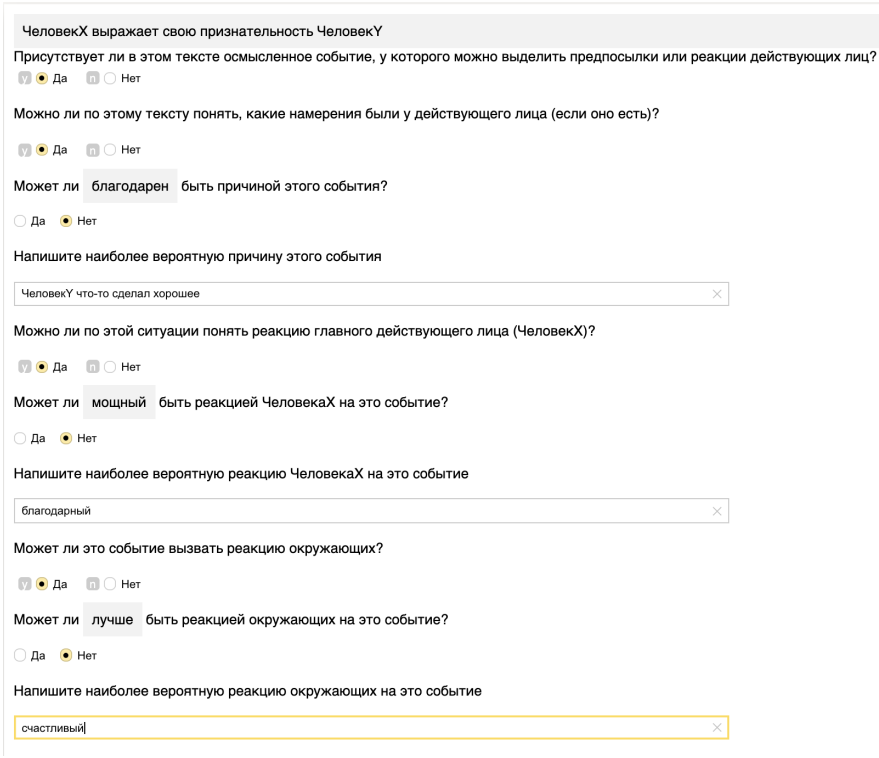

Figure 2: Toloka's markup for event2mind task

As a result we collect almost 7k examples of good Russian events and corresponding intents and reactions for them.

2.3. Translated English corpus

Besides the annotated Russian dataset we prepared an automatically translated into Russian and then cleaned English corpus. The motivation for this was that we did not posses enough resources to annotate as many events in Russian as the authors did. In order to find the dependence between the model's performance and the corpus size

and to estimate the minimum size of Russian dataset sufficient for the model's training we performed a number of experiments. The original English model was trained on different subsets of the original corpus, which were of different size. The results of the experiments are presented in Table 1. It should be mentioned that we were unable to obtain exactly the same scores as in paper though we used the original Allennlp[8] event2mind config of the best model loaded in official framework's repository.

Table 1: English model performance in dependence on the dataset size. Evaluation of intents/reactions is measured similarly to the original paper by *recall*@10 (percentage of times the gold falls within the top 10 decoded; higher is better) on development set

data size	val loss	intent	xreact	oreact
46k (full)	2.52	0.38	0.41	0.65
30k	2.60	0.36	0.39	0.65
20k	2.74	0.32	0.37	0.63
5k	3.22	0.31	0.35	0.55

It could be observed that small dataset of 5,000 examples is obviously not enough for the model to obtain reasonable quality. However, with the set of 30 thousand events the validation loss and intent recall metrics do not critically decrease compared to the full English corpus. Thus, it could be concluded that the annotated Russian corpus of about 7,000 events is not enough for training a model and the data should be to augmented.

For this purpose we create a supplementary dataset from English corpus automatically translated into Russian. The original dataset is translated via Google translator. The English corpus contains data from several sources: ROC Story training set [11], the GoogleSyntactic N-grams [4], the Spinn3r corpus [5], and idioms. However, we take only examples from the source ROC Story as it has the highest annotation agreement statistics according to the original research [12] and correspondingly it contains more clear events which are less exposed to the automatic translation errors.

The translated data is further filtered and phrases containing English or transcribed phrases are removed. Finally, the data is checked by reviewers and sentences which were translated wrongly are deleted.

It should be noted that the translated corpus is of poorer quality compared to the crowdsourced one due to the imperfections of the automatic translation and the lack of coherence between the phrases and the labels.

2.4. Final dataset

The final Russian corpus[9] is a union of the Russian crowdsourced dataset and the automatically translated part of English corpus which contains 30,165 events in the event2mind format.

[8] https://allennlp.org/

[9] https://github.com/Alenush/russian_event2mind/tree/master/dataset

3. Model and experiments

The event2mind model aims to generate three entity-specific pragmatic inferences (subject's intent, subjects's reaction, and others events participants' reactions) given an event phrase in free-form text. First, the input is encoded as a vector $h^E \in \mathrm{R}^H$. This vector is further used to predict the output which consists of three sequences of words. Event2mind is a neural encoder-decoder model. The system is multitask learning, simultaneously minimizing the loss for all three decoders at each iteration.

RNN decoder generates the textual description. The event phrase embedding h^E is set as the initial state h_{dec} of three decoder RNNs, which then output the intent/reactions one word at a time (using beam-search at test time). An event's intent sequence ($v_i = v_i^{(0)}, v_i^{(1)}, \ldots$) is computed by the following formula:

$$v_{(t+1)i} = softmax(W_i\ RNN(v_{(t)i}, h_{(t)i,dec}) + bi).$$

Thus, a model can successfully compose embedding representations of previously unseen events and reactions. Though the event sequences are typically rather short (3.6 tokens on average), event2mind model still benefits from the BiRNN's ability to compose words.

In the conducted experiments we used the full final Russian dataset (both translated and crowdsourced parts). In order to make translated and annotated example distribution more uniform the dataset was shuffled before training the model. It should be mentioned that we also experimented with training a model only on the translated part of corpus, however, it did not perform well. In the original article the authors performed experiments with different encoders and decoders but the most promising one were found to be BiRNN with GRU and vector of size 100 in encoder and sequence decoder. We tried LSTM and GRU in encoder and different embeddings models: such as fasttext [1], [7] and word2vec [9], [10] from RusVectores source.[10] The results are presented in the **Table 2**.

Table 2: Scores on Russian dataset with different model configurations. Evaluation of intents/reactions is measured similarly to the original paper by *recall*@10 (percentage of times the gold falls within the top 10 decoded; higher is better) on development set. Best models by the average *recall*@10 are highlighted in bold.

vectors	architecture	val loss	intent	xreact	oreact
araneum fasttext	LSTM	0.9704	0.818	0.725	0.92
araneum fasttext	GRU	0.9695	0.819	0.725	0.92
ruscorpora fasttext	LSTM	0.9508	**0.821**	**0.725**	**0.9195**
ruscorpora fasttext	GRU	0.9517	**0.822**	**0.7255**	**0.9195**
araneum word2vec (skipgram)	LSTM	0.916	0.816	0.707	0.924
araneum word2vec (skipgram)	GRU	0.919	**0.827**	**0.727**	**0.92**
ruscorpora word2vec (skipgram)	LSTM	0.9205	0.816	0.70	0.92
ruscorpora word2vec (skipgram)	GRU	0.917	**0.825**	**0.725**	**0.923**

[10] https://rusvectores.org/ru/models/

From the obtained results the following conclusions could be made:

- word2vec embeddings perform a little better than fasttext ones,
- ruscorpora fasttext embeddings perform better compared with the araneum ones,
- with word2vec embeddings GRU shows better results than bidirectional LSTM.

The best model[11](areneum word2vec + GRU) repeats the English results on the Russian data. **Table 3** shows several examples of the model's work on real event phrases. It could be seen that the model's predictions for intents are sensible and reflect common knowledge and reasoning. Still person's reactions are of poorer quality. It may be explained by the fact that the set of emotions depends on a larger piece of text than one short phrase. Therefore, event phrase without the context is not enough for making precise decisions about peoples' feelings or emotions.

Table 3: Examples of model's commonsense inference

Event	Intent	PersonX's reaction
PersonX выпил кофе	пить, разбудить, проснуться	счастливый,гордый, довольный
(PersonX drank some coffee)	(to drink, to awaken, to wake up)	(happy, proud, satisfied)
PersonX позвал на свидание PersonY	любовь, привлечь, благодарен	довольный, счастливый, облегчение
(PersonX invited PersonY for a date)	(love, attract, grateful)	(satisfied,happy,relief)
PersonX идет в школу	учить, веселиться, чтобы получить образование	гордый,грустный, довольный
(PersonX goes to school)	(to teach, to have fun, to get education)	(proud, sad, satisfied)

4. Conclusion

In the paper a new Russian corpus for event2mind task was introduced. Our corpus supports learning representations over a diverse range of events and reasoning about the likely intents and reactions of previously unseen events. In addition to that the results of the experiments with the Russian model are described and the code for the best model is provided. It was demonstrated that architecture works for more grammatically complicated language than English and it is still performing commonsense inference on textually described everyday events. The dataset and Russian model are provided in the repository[12].

[11] https://github.com/Alenush/russian_event2mind/tree/master/model

[12] https://github.com/Alenush/russian_event2mind

In the future we plan to increase the model's performance by experimenting with different language models such as BERT and ELMo, for example. In addition to that we have noticed that the number of subject's reactions is quite limited and, therefore, it could be regarded as a classification problem. Thus, we plan to train a separate classification model on top of BERT or ELMo embeddings.

References

1. *Bojanowski, Piotr, et al.* "Enriching word vectors with subword information". Transactions of the Association for Computational Linguistics 5 (2017): 135–146.
2. *Devlin, Jacob, et al.* "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805 (2018).
3. *Droganova, Kira, Olga Lyashevskaya, and Daniel Zeman.* "Data conversion and consistency of monolingual corpora: Russian UD treebanks". Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway. No. 155. Linköping University Electronic Press, 2018.
4. *Goldberg, Yoav, and Jon Orwant.* "A dataset of syntactic-ngrams over time from a very large corpus of english books". (2013).
5. *Gordon, Andrew S., and Reid Swanson.* "StoryUpgrade: Finding Stories in Internet Weblogs". ICWSM. 2008.
6. *JBenko, V., & Zakharov, V. P.* "Very large Russian corpora: new opportunities and new challenges. In Computational linguistics and intellectual technologies (pp. 79–93) (2016).
7. *Joulin, Armand, et al.* "Bag of tricks for efficient text classification". arXiv preprint arXiv:1607.01759 (2016).
8. *Liu, Quan, et al.* "Combing context and commonsense knowledge through neural networks for solving winograd schema problems". 2017 AAAI Spring Symposium Series. 2017.
9. *Mikolov, Tomas, et al.* "Efficient estimation of word representations in vector space". arXiv preprint arXiv:1301.3781 (2013).
10. *Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig.* "Linguistic regularities in continuous space word representations". Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.
11. *Mostafazadeh, Nasrin, et al.* "A corpus and evaluation framework for deeper understanding of commonsense stories". arXiv preprint arXiv:1604.01696 (2016).
12. *Rashkin, Hannah, et al.* "Event2mind: Commonsense inference on events, intents, and reactions". arXiv preprint arXiv:1805.06939 (2018).
13. *Straka, Milan, Jan Hajic, and Jana Straková.* "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing". Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016.
14. *Trinh, Trieu H., and Quoc V. Le.* "A simple method for commonsense reasoning". arXiv preprint arXiv:1806.02847 (2018).

15. *Wang, Alex, et al.* “Glue: A multi-task benchmark and analysis platform for natural language understanding”. arXiv preprint arXiv:1804.07461 (2018).
16. *Wang, Alex, et al.* “Superglue: A stickier benchmark for general-purpose language understanding systems”. Advances in Neural Information Processing Systems. 2019.
17. *Yang, Zhilin, et al.* “Xlnet: Generalized autoregressive pretraining for language understanding”. Advances in neural information processing systems. 2019.
18. *Zellers, Rowan, et al.* “Swag: A large-scale adversarial dataset for grounded commonsense inference”. arXiv preprint arXiv:1808.05326 (2018).
19. *Дяченко, П. В., et al.* “Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус)”. Труды Института русского языка им. ВВ Виноградова 6 (2015): 272–300.