RuNormAS-2021: a Shared Task on Russian Normalization of Annotated Spans

Denis Zolotukhin
ABBYY
Moscow, Russia
denis.zolotukhin@abbyy.com

Ivan Smurov
ABBYY
Moscow, Russia
ivan.smurov@abbyy.com

Abstract

The paper presents result of the shared task on normalization of spans. The goal of the shared task is introduce datasets and to establish baseline for span normalization in Russian language on different data domains. The shared task introduces two datasets with annotated spans of different nature. The first consists of news articles from Vzglyad newspaper with marked up named entity spans to be normalized by participants. The second contains development strategies of Ministry of Economic Development with more generic spans to be normalized.

On total 6 teams participated in the shared task with best results achieved on the generic track 0.980 exact match and on named entity track 0.981 exact match. Detailed competition description, the data, and evaluation scripts are available at: https://github.com/dialogue-evaluation/RuNormAS.

Keywords: normalization, named entities, shared task, Russian
DOI: 10.28995/2075-7182-2021-20-1245-1250

RuNormAS-2021: Соревнование по нормализации спанов текста

Золотухин Д. Д.
ABBYY
Москва, Россия
denis.zolotukhin@abbyy.com

Смуров И. М.
ABBYY
Москва, Россия
ivan.smurov@abbyy.com

Аннотация

В статье представлены результаты соревнования по приведению в нормальную (начальную) форму спанов текста. Целью соревнования является представление датасетов и установление бейзлайнов на них для задачи нормализации в русском языке на данных разной природы. Для соревнования используется два корпуса: первый состоит из новостных статей газеты Взгляд, на которых требуется нормализовать именованные сущности, второй - из программ стратегического развития Минэкономразвития РФ, на которых требуется нормализовать более общие спаны.

Всего в соревновании приняло участие 6 команд. Наилучшие результаты, достигнутые для дорожки со спанами именованных сушностей, - 0.981 точно совпавших, для дорожки с общими спанами - 0.980. Детальное описание соревнования, текстовые коллекции, инструкция по разметке и скрипты для оценки качества доступны по ссылке: https://github.com/dialogue-evaluation/RuNormAS.

Ключевые слова: нормализация, именованные сущности, соревнование, русский язык

1 Introduction and Related Work

Normalization is a task of bringing the word or collocation to the normal form retaining the agreement inside the span.

While many cases of normalization can be handled by per-word lemmatization, most complicated and interesting examples go far beyond that. E. g. span "кафедры комьпютерной лингвистики" (chair of computational linguistics) should be normalized as "кафедра компьютерной лингвистики"

which differs from per-word lemmas "кафедра компьютерный лингвистика". Thus the correct normalization often requires knowledge of inter-dependency of words in the collocation.

The task is also highly context-dependent. E. g. in Russian span "Иванова" can be normalized differently in the following context "Я вижу Иванова" (I see Ivanov) and in "Иванова вышла" (Ivanova left).

Lemmatization was studied in great detail both for Russian and other languages. Numerous shared tasks were conducted with lemmatization being one of the tasks including, but not limited to SIGMORPHON-2016[12], SIGMORPHON-2017[4], SIGMORPHON-2018[3] and SIGMORPHON-2019[13] across multi-language collections. For Russian specifically the first shared task on lemmatization was held in 2010[10], later followed by MorphoRuEval-2017[9] and GramEval-2020[5].

On contrast normalization of full spans received little to no attention. To our best knowledge the only dataset that can be used for normalization of annotated spans in Russian is Persons-1000[1], that contains annotations of persons with their normal forms. Unfortunately unlike with other named entities, normalization of persons can be reduced to per-word lemmatization in the overwhelming majority of cases.

Thus one can argue that there exists a shortage of available resources for span normalization in Russian. RuNormAS-2021 shared task is our attempt introduce two datasets for this task and to establish reasonable baselines on them with the help of participants.

First track of the shared task is dedicated to named entities normalization. Named entities include Persons, Locations and Organizations. Normalization of such collocations is, in particular, very important to industrial applications of information extraction.

More complicated task is to normalize a broader range of spans of text, not bounded by a label of named entity, such as adjectives, verbs, etc. Second track of the competition consists of such spans. We made an attempt of defining reasonable normal forms for them and asked participants to try to reproduce it.

This paper is organized as follows: section 2 describes the datasets together with the process of mark up. Section 3 describes the shared task set up as well as a simple baseline. Section 4 provides the official shared task results and analysis. Finally, section 5 contains ending remarks.

2 Data

Data consist of two parts:

- A sample of news articles from vz.ru - Vzglyad newspaper
- Strategic planning programs of development for Russian regions from Ministry of Economic Development of Russia

First corpus consists of news articles from Vzglyad newspaper. Articles for train part of the data are drawn from 2018, 2019 and 2020, and for test part from 2015. We first marked up named entities with the help of open-sourse automatic NER tagger Slovnet [2] in order obtain an estimate of all named entities in texts. The corpus data was sampled from all articles with an effort to minimise the intersection of sets of named entities between train and test in order to (at least partially) prevent data leaks. After the data is sampled we annotate named entities with the help of accessors and then normalize them (the process of mark up is the same for the 2 datasets and is described below).

Second corpus is based on the data used for RuREBus task[11]. It includes documents annotated for relation extraction task. We use only those parts of annotations that contain the spans of entities in text. Since the entities in RuREBus are non-traditional and their spans can be not only noun phrases, but also verb phrases, adjective phrases etc, this corpus is well-suited to check the normalization of a broader set of spans than the first corpus. All entities spans were normalized with the help of accessors.

Table 1 depicts the sizes of both datasets.

The process of mark up for both datasets was as followed: each span was manually annotated by one accessor with the guidelines described in the following section. In order to verify the consistency of

[1] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000

[2] https://github.com/natasha/slovnet

	vz.ru	Economic Development
Texts (train)	2493	188
Texts (test)	399	30
Spans (train)	39573	54360
Spans (test)	11461	8455

Table 1: Dataset sizes

the markup we gave a portion of data (approximately 10k spans for each corpus) to two accessors later verified by a moderator. The resulting inter-annotation agreement is over 0.998 for both corpora.

We have also explored sample of data in order to estimate the share of spans that cannot be normalized by lemmatization alone. We found, that at least 30% of entities in our datasets require more complex approach.

2.1 Annotation Guidelines

For Named Entities the definition of normal form is trivial. For generic spans on the other hand it is a lot more complicated. We restrict the set of variants for the main word in normalized form to the following:

- Noun in the nominative case (with retained number):
- Adjective in the nominative case and full form (with retained gender):
- Verb in the infinitive (with retained reflexivity):
- Pronoun
- Number
- Particle or predicative

Then we use a set of rules to guide the normalization of Generic track:

- If a span consists of parts, that are compositionally connected, then each part must be normalized independently.
- All dependent words must be consistent with the main word, and changes to the form of the dependent words should only be made if it is necessary to align the dependent words with the form of the main word
- Verbs, participles, and adverbs must be converted to the infinitive (except when the participle is dependent on the noun in the context of the phrase)

We do not include the following transformations into normalization in both tracks:

- Changing the order of components in the entitty (for example name and surname)
- Decoding of the abbreviations
- Transforming numbers into words and visa versa
- Correction of spelling errors

3 Shared Task Set Up

3.1 Tracks and metrics

First track (Named) includes named articles from vz.ru, and the second one (Generic) consists of texts from Ministry of Economic Development.

For both tracks we check if normalization of each span is the same as the result submitted by the participant. We use fraction of exact matches as metric. We do not take into account diactric signs (primarily this is referred to e - $\ddot{\mathrm{e}}$) in both tracks.

Due to technical artifacts in named entity markup it was decided to soften the criteria for the Generic track. When measuring the accuracy we did not consider the case of the letters and spaces between words.

3.2 Baseline

We introduced a baseline solution for both tracks. It uses normalizer from Natasha project[3]. Normalizer uses neural models for morphology tagger, (weak) syntax parser and named entity recognition module. Natasha normalizes named entities taking into account syntax dependencies. Table 3.2 present accuracy achieved by the baseline.

	Generic	Named
Baseline	0.77315	0.88805

Table 2: Baseline accuracy on both tracks

The resulting scores are quite low, even considering that a good share of examples is already in the normalized form. Also, as expected, generic spans appear to be harder to normalize with Natasha normalizer (partially because Natasha is trained to normalize named entities).

4 Results and Analysis

	Generic	Named
ksmith	**0.98013**	0.98115
qbic	0.97907	**0.98150**
eindenbom	0.97575	0.9792
king_menin	0.96452	0.95750
Baseline	0.77315	0.88805
fateev.da	0.77303	0.88970
shkunkov.a	-	0.76798

Table 3: Final results.

Table 3 presents the official results of the shared task.

Here we can see, that best results far outperform baseline solution, meaning that participants were able to find solutions that cover complicated examples. One can also notice that although accuracy on named entities is higher, the difference in score from generic spans is marginal. We believe this shows that approaches to normalizing named entities - a more classic task - can be also applied to broader range of text spans. At the same time, top scores do not greatly exceed 0.98 which means there is still space for work in this area.

Methods used to solve both tasks are diverse. The best solution in the Generic track and the second place in the Named track (**ksmith**) uses deterministic rule-based approach with help of PullEnti SDK[6]. The winner in the Named Entity track and the second place in the Generic (**qbic**) uses data-driven approach: it fine-tunes RuBERT[7] (a Russian version of BERT[2]) using lemmatization head: for each word the model predicts which part should be removed and added in order to obtain normalized result (the approach is inspired by the author's participation in GramEval-2020[1]). The forth one in both tracks (**king_menin**) decided to use very general method and tuned RuGPT-3[4] a Russian version of GPT-3[8].

We are glad to see that different approaches to the task were able to demonstrate competitive results. We think that two possible conclusions can be induced from this fact. First, the current state-of-the-art is reasonably well defined since it can be achieved by using vastly different methods. Second, there isn't a universal solution to this task and so we will see development of work in this area.

5 Conclusion

Annotated span normalization is a task of converting an annotated span text to its normal form. This task is non-trivial: it requires the preservation of the correct syntactic structure of the span as well as highly

[3] `https://github.com/natasha`

[4] https://huggingface.co/sberbank-ai/rugpt3xl

context dependent. Because of the former in many cases it cannot be reduced to lemmatization.

While the task is important for industrial applications of information extraction there exist virtually no corpora on this task for Russian. The only dataset is Persons-1000 and it contains only spans of persons that can be reduced to lemmatization in an overwhelming majority of cases.

In this paper we address this issue by introducing two datasets: first containing normal forms of classical named entities (i. e. persons, locations and organizations) and the second based on RuReBus corpus with normal forms of more general spans.

We further explore these corpora by creating a simple baseline and by conducting a shared task on both corpora. The six teams participating in the shared task established state-of-the art for normalization of annotated spans: 0.981 for named entities and 0.980 for general spans.

We hope our corpora and RuNormAS-2021 shared task will create a reference point for normalization useful for both academic and industrial systems and jump-start further development in the field.

References

[1] Anastasyev D. G. Exploring Pretrained Models for Joint MorphoSyntactic Parsing of Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. 19. — 2020. — P. 1–12.

[2] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[3] The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection / Ryan Cotterell, Christo Kirov, John Sylak-Glassman et al. // Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. — 2018. — P. 1–27.

[4] CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages / Ryan Cotterell, Christo Kirov, John Sylak-Glassman et al. // CoNLL SIGMORPHON. — 2017. — P. 1.

[5] GRAMEVAL 2020 Shared Task: Russian Full Morphology and Universal Dependencies Parsing / O. N. Lyashevskaya, T. O. Shavrina, I. V. Trofimov, N. A Vlasova // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. 19. — 2020. — P. 553–569.

[6] Kozerenko Elena Borisovna, Kuznetsov Konstantin Igorevich, Romanov Dmitrii Aleksandrovich. Semantic processing of unstructured textual data based on the linguistic processor PullEnti // Informatika i Ee Primeneniya [Informatics and its Applications]. — 2018. — Vol. 12, no. 3. — P. 91–98.

[7] Kuratov Yuri, Arkhipov Mikhail. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // CoRR. — 2019. — Vol. abs/1905.07213. — 1905.07213.

[8] Language Models are Few-Shot Learners / Tom Brown, Benjamin Mann, Nick Ryder et al. // Advances in Neural Information Processing Systems. — Vol. 33. — Curran Associates, Inc., 2020. — P. 1877–1901. — Access mode: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian / Aleksei Sorokin, Tatiana Shavrina, Olga Lyashevskaya et al. // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. 16. — 2017. — P. 297–313.

[10] NLP evaluation: Russian morphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. / Olga Lyashevskaya, Irina Astaf'eva, Anastasia Bonch-Osmolovskaya et al. // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue". — Vol. 9(16). — 2010. — P. 318–326.

[11] Rurebus-2020 shared task: Russian relation extraction for business / V.A. Ivanin, V.V. Ivanov,

T.V. Batura et al. // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”. — Vol. 19. — 2020. — P. 416–431.

[12] The SIGMORPHON 2016 shared task—morphological reinflection / Ryan Cotterell, Christo Kirov, John Sylak-Glassman et al. // Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. — 2016. — P. 10–22.

[13] The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection / Arya D McCarthy, Ekaterina Vylomova, Shijie Wu et al. // Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology. — 2019. — P. 229–244.