EPiC Series in Language and Linguistics

Volume 4, 2019, Pages 11–17

Proceedings of Third Workshop "Computational linguistics and language science"

EPiC Language and Linguistics

Automatic morphological analysis on the material of Russian social media texts

Alena Fenogenova[1], Viktor Kazorin[2], Ilia Karpov[1], and Tatyana Krylova[3]

[1] National Research University Higher School of Economics, Moscow, Russia
alenush93@gmail.com, karpovilia@gmail.com
[2] Clover Group, Moscow, Russia
zhelyazik@mail.ru
[3] Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia
ta-kr@yandex.ru

Abstract

Automatic morphological analysis is one of the fundamental and significant tasks of NLP (Natural Language Processing). Due to special features of Internet texts, as they can be both normative texts (news, fiction, nonfiction) and less formal texts (such as blogs and texts from social networks), the morphological tagging has become non-trivial and an actual task. In this paper we describe our experiments in tagging of Internet texts presenting our approach based on deep learning. The new social media test set was created, that allows to compare our system with state-of-the-art open source analyzers on the social media texts material.

1 Introduction

A great amount of theoretical and practical research is conducted in the field of automatic Russian morphological analysis. Theoretical study mainly faces problems concerning part of speech distinction and classification of grammatical categories [12, 14], while the practical work focuses mainly on literary texts and solving the problems of homonymy [9, 13]. However, nowadays disambiguation is not the sticking point in state-of-the-art systems. As texts that systems process are mainly from the Internet (not only news and fiction but also blogs, social media texts), we face the problem of processing specific text features such as misspelling, slang, emotional remarks (like "ооочень" ("veeery")) etc. The problem that tagging systems have to cope with have become more varied and tricky. The methods applied to this task have changed as well.

Systems that provide automatic morphological analysis of Russian text, allow "train and test" functionality to be tested and compared in research [3, 7]. They show a great variety of approaches to the tagging task, still the recent trends in the Natural language processing field cause the researchers to use deep learning techniques [4, 13].

The current work proposes the deep learning approach for tagging Internet texts. In this paper we create a golden dataset, contained only social media texts and compare system perfor-

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 11–17

mance on it with the results on the test set from the competition MorphoRuEval-2017 Dialogue Shared Task.

2 Related work

Most research works on morphological analysis have mainly focused on standardized literary texts for many years. However, the interest in automatic evaluation of social media texts is growing considerably in recent years. As the nature of social media texts is clearly different from standardized texts [1], Natural Language Processing methods need to be adapted for reliable processing. The recent works had investigated the performance of different taggers and approaches on Twitter data [5], the influence of specific in-domain social media lexicon [10] on the performance, the multilingual approaches for this problem [4] an so on.

An independent evaluation of morphological analysis methods and linguistic tools for Russian [8, 17] created the first standards and datasets, revealed problems with tag sets, addressed the problem of homonymy and rare words. The focus of MorphoRuEval-2017 "Dialogue" the Shared Task [16] as it was introduced by the organizers, was the processing of Internet texts that had a lot of specific features. The new datasets were introduced in universal dependencies (UD) format. The results show a great variety of approaches including the popular deep learning techniques. However, the test dataset of social media was presented only by texts from VKontakte social network. From our point of view this dataset and, consequently, the solutions within the competition framework do not take into account the existing variability of network lexic. Thus, in the current paper we present our deep learning approach for morphological analysis as well the new dataset of social media texts for Russian language.

3 Methodology

3.1 Preprocessing

An Internet text has some specific features that need to be taken into consideration during the processing of data. Our research on 10,000 LiveJournal texts shows that 4.7 sentences out of 10 sentences contain out-of-vocabulary words (OOV), about 3 of 10 sentences contain at least one typographic error. So, the proportion of typos in social network text is quite significant, at least 3%. As the dataset we are working with contain texts from the social networks VKontakte and Facebook, where people write mostly informal messages in comments, the preprocessing of these texts is required. The special social media spell checking module was created for this purpose and applied for datasets.

Firstly, we have converted text to lowercase and applied transformations for irregular patterns: based on the data from the SpellRuEval Shared Task 2016 [15], we have created a list of about 50 transformations (such as *"гру" > "говорю", "ваще" > "вообще", "чота" > "что-то"* etc.). Secondly we have used special dictionaries to exclude words that do not need correction:

- slang dictionary – the dictionary consists of teenager's slang list (from the source - http://teenslang.ru) and manually tagged collection of slang words;
- anglicisms dictionary – a list of anglicisms collected by the method described in work [6];
- dictionary of names – manually tagged collection of name entities;

- A.A. Zaliznjak's Russian Grammatical Dictionary (to exclude vocabulary words)

Next, for words that left after the previous step, the correction module was applied. The module suggests candidate corrections by using Damerau-Levenshtein edit distance to find words that have distance from one to three from the misspelled word. Then, the list of most probable candidates are estimated by passing hypothesis into the trigram language model and the error model. The error model is based on the distribution of the most probable partitions of the typo and the candidate into substrings of letters – the product of conditional probabilities for each part of the typo given the corresponding part of the candidate is the highest.

3.2 Using RNN for tagging

Proposed system is based on the deep learning techniques, the system algorithm is the following. We have defined $s(w_i)$, $f(w_i)$ as stem and flexion of word w_i respectively; the stem and flexion have been gained by the Snowball stemmer. For tagging the one word w_i the stem embeddings get as input the following sequence: $s(w_i\text{-}C)$, ..., $s(w_i)$, ...,$s(w_i+C)$, where C is the number of left/right neighbours. Flexion embeddings get as input the sequence $f(w_i\text{-}C)$, ..., $f(w_i)$, ..., $f(w_i+C)$. For char embeddings the inversion sequence of Nchar word's chars is given; the sequence is cut if the word's length is longer than Nchar or if shorter – the nulls are added. Next, these three sequences by means of corresponding embedding layers are transformed into the vector sequences and these vector representations are given to the Bidirectional LSTM (further in the text – BLSTM).

The usage of the BLSTM is described below:

- in case of a char embedding the layer decodes the word in some vector representation that corresponds to the sense of the word on its char-level;
- in case of a stem sequence the decoding of the semantics is provided due to the context window of the word itself and left/right neighbours (syntactic information);
- the flexion sequence provides the decoding of the word morphology and its context.

Then, the outputs of all BLSTM are concatenated in one vector and this vector is given to the dense layer with the activation function ReLU. Next, the result of this layer is transmitted simultaneously to 13 softmax layers (for classification of every morphological category). During the training procedure the overall loss is minimized by all outputs:

$$L(in, out) = L_1(in, out_1) + \ldots + L_{13}(in, out_{13}), \tag{1}$$

where Li is the categorical cross-entropy.

The full architecture of the described network is presented on the Fig. 1.

In our experiments we have tried different variants of recurrent networks: LSTM, GRU, Simple RNN. LSTM has shown the highest results.

3.3 Embeddings

Every embedding layer represents the lookup table. In the case of the stem and flexion embedding these tables are initialized by the word2vec models pre-trained on the social media corpora (social media texts and electronic dictionary lib.rus.ec provided by the organizers of the MorphoRuEval-2017). Then they were further trained with the whole neural network represented above. The following preprocessing is provided before training the word2vec as well as before giving data as input to neural network:

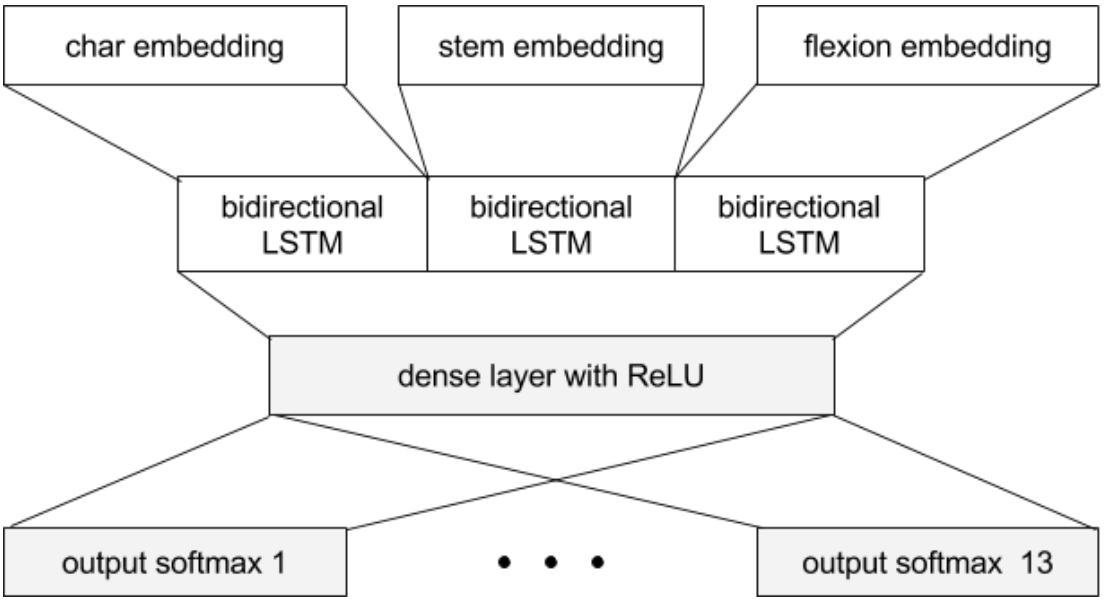

Figure 1: The architecture of BLSTM.

- the token dictionary is built
- the tag "undefined" is assigned to the tokens with low frequency

After the training procedure, the token *"null"* is added in embedding with the corresponding vector that is randomly initialized and used for padding. If the left or right context has the length less than C, *"null"* is assigned to all missing tokens. The lookup table for char embedding is initialized by the numbers generated in range from -0.05 to 0.05 of uniform distribution.

4 Evaluation

4.1 Datasets

For evaluation we have used two following datasets: 1) datasets from MorphoRuEval 2) social media dataset.

Four sources of annotated data were presented by the Shared Task for close track as training sets: RNC Open, GICR corpus with the resolved homonymy, OpenCorpora.org data and SynTagRus.

Extra data from the MorphoRuEval, which we use to train the distributional semantics model, are the following: 1) VKontakte data - about 500k tokens 2) Twitter data - about 300k tokens.

To unify the representation of marked data, the organizers decided to use the conll-u format, as the most common and convenient. The format of the Universal Dependencies (UD) [11] (further UD) with some specifications, agreed between the participants and organizers (UD 1.4 as well as 2.0 have been applied) was chosen to unify morphological tags. As our research was done with an evaluation of the MorphoRuEval-2017 Shared Task, we had to build our system

Table 1: Results on the datasets from MorphoRuEval-2017. All scores represent accuracy per tag

Dataset	MorphoRuEval best (close)	MorphoRuEval best(open)	BLSTM	TreeTagger
News	93,99%	97,37%	92.69%	81,17%
VK	92.42%	**96,62%**	92.31%	77.46%
Fiction	94.16%	97,45%	92.1%	74.34%

Table 2: Results on the social media dataset.

System	Accuracy per tag	Accuracy per sentence
TreeTagger	75.29%	23.81%
BLSTM (without spell)	89.88%	58.27%
BLSTM	92.48%	61.04%

and carried out experiments paying special attention to the data, agreements and specifications of this competition.

Additionally, we have created our own golden dataset that contains only sentences from users of social networks Facebook and VKontakte. These texts as opposed to MorphoRuEval-2017 golden standard contains a significant amount of misspellings, slang and proper names. We have collected texts, tokenized them by Mystem version 3.0, convert to UD.2 format and carefully checked automatically and manually by the experts. The new golden dataset contains about 5000 tokens. The final dataset can be acquired from github[1] repository.

4.2 Experiments and results

We have carried out a set of experiments within the framework of the competition on the test data provided by the organizers. We have trained our system on the train data from GICR and tested system on three MorphoRuEval's test datasets: 1) news data (lenta.ru) 2) vk.com data 3) fiction texts.

Additionally, we have also tested Baseline system of the competition – TreeTagger[2]. This is an open source tagger, based on the HHM, which uses a binary decision tree to estimate transition probabilities. We have trained TreeTagger on the GICR train data as well.

The system's outputs were evaluated by the competition's scripts and the following scores were estimated: 1) part of speech tag and full tag accuracy pro item 2) part of speech tag and full tag accuracy pro sentence.

The results of our system's performance, baseline performance and output of the best competition system from close and open tracks are presented in the Table 1 for all three MorphoRuEval-2017 test sets.

The results of the Baseline-system TreeTagger and our neural network system, tested on our new social media dataset are presented in Table 2.

Most of the system errors appeared in non-vocabulary lexicon, so the gradual increase of the training collection should reduces the number of errors. However our attempts to train the model on the joint RNC, GICR, OpenCorpora, SynTagRus collections decrease system performance. This may be caused by the specific features of the given data or some discrepancy

[1]https://github.com/lab533/SocialMediaCorpora
[2]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

in format. We consider this problem can be partially solved by means of active learning and/or the use of a more sensitive error propagation model.

5 Discussion

The most complicated cases in part of speech tags are pronouns and conjunctions that are often confused with determinatives and particles, respectively. Adverbs and parenthesis as well as proper names and nouns are very frequent as well. However, on the MorphoRuEval-2017 in the evaluation script organizers have skipped or specified most of such cases by the predetermined lists. Verbs or adjectives instead of nouns are more crucial as they are usually the cases of homonymy that are really tricky.

Among the categories of full tags the Case nominative and accusative are expectedly the most complex one. Animacy is also the frequent cause of taggers problems, as well as category Number – the slang words and proper names in social media texts cause such errors.

The results show that spell checking preprocessing (with an accurate processing of proper names) of the text gained the positive growth rate of almost 3%. However, despite the fact that Internet texts especially comments of users from social networks are full of typos and emotional patterns (like *"пожаааалуйста" ["pleeeease"]*), testset of VKontakte data logically contains such items, while the golden set from competition surprisingly almost does not.

6 Conclusion

In this paper we describe our experiments in tagging of Internet texts and present our approach based on deep learning techniques. We have test out system on the dataset we have created and on the data provided by the organizers of MorphoRuEval-2017 Shared task. The results of our system are comparable with the top systems performance from the competition for tagging and morphological disambiguation tasks.

We have proved that preprocessing of the texts increase the quality. We have also tried the multiclass (about 300 classes) morphology features classification, which is not worth implementing than training each morphological feature separately.

Acknowledgements

The reported study is supported by RFBR grant № 16-29-09614 “Principles and technologies of machine processing for social network texts in Russian” and Russian Academic Excellence Project ’5-100’.

References

[1] Baldwin T. et al.: How noisy social media text, how different social media sources? //Proceedings of the Sixth International Joint Conference on Natural Language Processing, p. 356-364, (2013)

[2] Cotterell R., Kirov J., Sylak-Glassman D., Yarowsky J., Eisner, Hulden: The SIGMORPHON 2016 shared task—morphological reinflection. In Proc. of the 2016 Meeting of SIGMORPHON. (2016)

[3] Dereza O., Kayutenko D., Fenogenova A.: Automatic morphological analysis for Russian: A comparative study. In Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. (2016)

[4] Ghosh S., Ghosh S., Das D.: Part-of-speech Tagging of Code-Mixed Social Media Text. EMNLP, pp. 90, (2016)

[5] Gimpel K. et al.: Part-of-speech tagging for twitter: Annotation, features, and experiments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2., Association for Computational Linguistics, p. 42-47, (2011)

[6] Fenogenova A., Karpov I., Kazorin V.: A General Method Applicable to the Search for Anglicisms in Russian Social Network Texts. AINL FRUCT 2016 conference proceedings and IEEE Xplore, (2016)

[7] Kuzmenko E.: Morphological analysis for Russian: Integration and comparison of taggers. International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham. (2016)

[8] Lyashevskaya O., Astaf'eva I., Bonch-Osmolovskaya A., Garejshina A., Grishina J., D'jachkov V., Ionov M., Koroleva A., Kudrinsky M., Lityagina A., Luchina E., Sidorova E., Toldova S., Savchuk S., and Koval' S.: NLP evaluation: Russian morphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2010. Vol. 9 (16). pp. 318-326, (2010)

[9] Muller T., Schmid H., Schutze H.: Efficient higher-order CRFs for morphological tagging. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, p. 322-332, (2013)

[10] Neunerdt M. et al.: Part-of-speech tagging for social media texts. Language Processing and Knowledge in the Web. Springer, Berlin, Heidelberg, p. 139-150, (2013)

[11] Nivre J., Marneffe, Ginter, et al.: Universal Dependencies v1: A Multilingual Treebank Collection. LREC, (2016)

[12] Pesetsky D.: Russian case morphology and the syntactic categories. MIT Press (2013)

[13] Plank B., Sogaard A., Goldberg Y.: Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. arXiv preprint arXiv:1604.05529, (2016)

[14] Sichinava D.: Parts of speech. [Chasti rechi. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki]. Moscow (2011)

[15] Sorokin A., Baytin A., Galinskaya I., Shavrina T.: Spellrueval: the first competition on automatic spelling correction for Russian. Proceedings of the Annual International Conference "Dialogue" №. 156 (2016)

[16] Sorokin A., Shavrina T., Lyashevskaya O., Bocharov B., Alexeeva S., Droganova K., Fenogenova A., Granovsky D.: MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. (2017)

[17] Toldova S., Sokolova E., Astafiyeva I., Gareyshina A., Koroleva A., Privoznov D., Sidorova E., Tupikina L., Lyashevskaya O.: Ocenka metodov avtomaticheskogo analiza teksta 2011-2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011-2012: Russian syntactic parsers]. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012. Vol. 11 (18). Moscow: RGGU. pp. 797-809, (2012)

[18] Tsuruoka. Y.: Stochastic Gradient Descent Training for L1-regularized Log-linear Models with Cumulative Penalty, (2009)

[19] Eger S., Gleim R., Mehler A.: Lemmatization and morphological tagging in German and Latin: A comparison and a survey of the state-of-the-art. Proc. of LREC (2016).