Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Russian neural morphological tagging: do not merge tagsets

Movsesyan A. A.
Institute for Information Transmission Problems (Kharkevich Institute)
Russian Academy of Sciences, Moscow, Russia
derise@iitp.ru

Abstract

There are multiple morphologically annotated corpora of Russian available. They have different tagsets and annotation guidelines, which makes them difficult to use together. We proposed a neural morphological tagger for Russian based on multitask learning technique which is able to predict morphological tags of words for different tagsets. We evaluated our model on various corpora and showed that utilising multiple corpora without merging them not only improves tagging performance but allows for scalable indirect conversion between multiple tagsets in all directions. Furthermore, we also showed that treating each corpus separately is more efficient than merging the corpora even if they share the same tagset.

Keywords: morphological tagging, tagset conversion, multitask learning
DOI: 10.28995/2075-7182-2022-21-402-411

Морфологический анализ русского языка на основе нейронных сетей: не объединяйте морфологические стандарты

Мовсесян А. А.
Институт проблем передачи информации РАН им. А. А. Харкевича
Москва, Россия
derise@iitp.ru

Аннотация

Для русского языка доступно множество аннотированных корпусов, снабженных морфологической разметкой. Различия между их морфологическими стандартами и схемами аннотации усложняют их совместное использование. Мы разработали модель морфологического анализатора для русского языка на основе нейронных сетей и многозадачного обучения. Модель позволяет снабжать слова морфологической разметкой для разных морфологических стандартов. Для оценки качества мы использовали ряд корпусов и показали, что использование нескольких корпусов без их слияния не только улучшает качество разметки, но и позволяет косвенно использовать модель для ковертации между несколькими морфологическими стандартами во все стороны, причем модель легко масштабируется на большее число стандартов. Кроме того, мы также показали, что использование каждого корпуса как отдельную единицу более эффективно, чем слияние корпусов, даже тогда, когда корпусы имеют общий морфологический стандарт.

Ключевые слова: морфологический анализ, морфологический стандарт, конвертация, многозадачное обучение

1 Introduction

Morphologically annotated corpora are valuable sources of data for linguistic research and natural language processing (NLP) tasks like morphological tagging and parsing. Such a corpus provides each word with a set of values of morphological categories[1] such as part-of-speech (POS), case or gender.

In the case of the Russian language, many corpora with morphological annotation exist. However, each corpus often has its own unique tagset (Hana and Feldman, 2010; Sharoff et al., 2008, to name a

[1] Throughout the paper we will refer to each unique set of morphological features assigned to a word as a morphological tag.

few) and converting between them without mistakes and information loss is a challenging task. One clear example is morphological analysis contest MorphoRuEval-2017 (Sorokin et al., 2017). The organisers provided four different annotated corpora and automatically converted morphological tags to the Universal Dependencies (UD) v2.0 format (Nivre et al., 2020). But most participants ended up using only one dataset because adding others did not improve the performance of their models, especially models based on deep learning methods.

From a linguistic perspective, merging different corpora allows linguists to widen their research scope. From a statistical perspective, including machine learning and deep learning, more data would allow better performance of morphological processing tasks because it helps with the data sparseness problem.

Tagset conversion is challenging for multiple reasons:

1. Lack of parallel data. Russian corpora have little common texts, which makes it hard to create conversion rules, since each word's tag depends on context. Training a supervised conversion model is also not possible under these circumstances.
2. Inter-annotator agreement. Even if two corpora share the same tagset, they might follow different annotation guidelines because some language phenomena are debatable. These differences might be crucial in terms of performance for the neural taggers. This problem to a lesser extent occurs within a single corpus when different annotators make different decisions because of the flaws in the guideline (Plank et al., 2014). Another challenge occurs when inter-annotator agreement score is high but all annotators make the same error in some cases (Bočarov et al., 2013).
3. Lack of annotated data. Some corpora are small and not representative enough to make plausible conversion results without the use of additional resources.

There are many approaches to this problem. We can divide them into two groups: direct and indirect. Direct approaches are mainly rule-based: for a given word in a source corpus, there is a rule to convert its tag to the target corpus format based on the word's context (including annotation). Although some automated tools exist to provide multi-corpora tagset conversion[2], it is hard to cover all possible patterns using rules, and it requires manual correction, which is time-consuming. For example, in the process of converting syntactically tagged Russian text corpus SynTagRus (Inšakova et al., 2019) to the UD format (Droganova and Zeman, 2016) some sentences were omitted due to differences in the guidelines. Somewhat similar is the task of providing a unified tagset from a number of corpora's tagsets for comparison purposes (standardisation). Such tagsets usually lack some morphological features because of conversion difficulties (Ljaševskaja et al., 2010; Lyashevskaya et al., 2017).

Indirect approaches are usually based on statistical morphological taggers. Such taggers, trained on the target corpus, intrinsically utilise source corpus annotation. These approaches are applicable to both tasks: morphological tagging and tagset conversion. One such approach (and some variations) aimed at tagset conversion trains a tagger to produce the so-called bundled tags (Li et al., 2015). Let T^s and T^t be the set of all possible tags in source and target corpus, respectively. Then the set of all bundled tags is a Cartesian product $T^s \times T^t$. During training, instead of predicting a correct label $t_i^t \in T^t$ the model predicts all labels in the set $\{t_i^t\} \times T^s$ thus making the labels ambiguous. That allows to predict labels from both tagsets at the same time. The authors tested the approach by training a POS tagger on two Chinese corpora. This approach is practically inapplicable to Russian because there are hundreds and thousands of different morphological tags possible in a given corpus compared to a few dozens of POS tags in Chinese, which keeps the Cartesian product small.

As for the Russian language, there is one indirect approach in the literature to our knowledge, and it is based on transfer learning technique (Andrianov and Mayorov, 2017). Namely, the authors trained multiple neural taggers (one tagger per source corpus in the case of multiple source corpora) and used their intermediate layers' outputs as inputs to the main tagger trained on the target corpus.

All those indirect approaches have one essential drawback: scalability. We often need to be able to make the conversion in both directions, and the mentioned approaches are not easy to apply when the number of the target corpora is more than one.

The primary objective of this paper is to show how unrelated Russian morphological corpora can

[2] See, for example, `https://pypi.org/project/russian-tagsets/`

benefit each other on the morphological tagging task in a scalable manner. We train a neural morphological tagger in a multitask learning setting, treating each corpus' annotation separately but sharing the intermediate text representation. We do not use pretrained word embeddings or any other external data besides the corpora. We evaluate our model on a set of Russian corpora and also on the data provided in the MorphoRuEval-2017 contest for comparison. We show that utilising multiple corpora in a multitask setting improves tagging performance on each tagset, but it depends on the size of the corpus. We also show that treating multiple corpora sharing the same tagset separately instead of merging them leads to a better tagging performance.

The paper is organised as follows. Section 2 describes the proposed neural tagger model. Section 3 provides experimental results, which we discuss in section 4. Section 5 concludes the paper.

2 Methods

Our model receives a tokenised sentence in the form of word[3] sequence $\{w_1, w_2, \ldots, w_n\}$ as input features, and predicts a sequence of morphological tags $\{t^j_1, t^j_2, \ldots, t^j_n\}$ for each tagset T^j. We provide detailed description of the model in the next sections.

2.1 Model architecture

The model has three basic blocks:

1. word embeddings
2. encoder layer
3. output layer.

We used GRU-based (Cho et al., 2014) character-level word embeddings, proven to be effective in various NLP tasks, including morphological tagging (Heigold et al., 2017; Lukovnikov et al., 2017). Each word w_i is represented as a sequence of its characters $\{c_1, c_2, \ldots, c_k\}$. Each character is represented as a one-hot encoded vector over a predefined vocabulary V^{char} and passed to a character embedding layer:

$$c_i^{embed} = W^{embed} \cdot one\_hot(c_i),$$

where $W^{embed} \in \mathbb{R}^{char\_embedding\_size \times |V^{char}|}$. All word's character embeddings are then passed to a unidirectional GRU layer:

$$\begin{aligned}
r_i &= \sigma(W_r c_i^{embed} + b_r + U_r h_{i-1}^{char} + u_r),\\
z_i &= \sigma(W_z c_i^{embed} + b_z + U_z h_{i-1}^{char} + u_z),\\
n_i &= \tanh(W_n c_i^{embed} + b_n + r_i \odot (U_n h_{i-1}^{char} + u_n)),\\
h_i^{char} &= (1 - z_i) \odot n_i + z_i \odot h_{i-1}^{char},\\
h_0^{char} &= 0,
\end{aligned}$$

where σ is the sigmoid function, $W_r, U_r, W_z, U_z, W_n, U_n \in \mathbb{R}^{char\_hidden\_size \times char\_embedding\_size}$ and $b_r, u_r, b_z, u_z, b_n, u_n$ are the bias vectors, respectively. The final hidden state of the character sequence is the word embedding of the word w_i:

$$w_i^{embed} = h_k^{char}$$

As the encoder layer, we chose the Transformer model's encoder. Not only this model showed promising results in various sequence tagging tasks (Devlin et al., 2019) because of its receptive field, but also its architecture allows easier interpretation through visualisation compared to other encoder models including recurrent neural networks. We did not make any changes to the architecture besides hyperparameter tuning (we also did not use the decoder layer of the Transformer) so we refer the readers to the original paper (Vaswani et al., 2017) for more details. The output of the encoder layer is

[3] We treated punctuation marks as words.

$$w_i^{enc} = TransformerEncoder(w_i^{embed}),$$

where $w_i^{enc} \in \mathbb{R}^{d_{model} \times 1}$.

We made $|T|$ output layers where $|T|$ is the number of tagsets (corpora). Each output layer projects each encoder's output to a probability distribution over a predefined set of tags:

$$w_i^{out} = softmax(W_{out}^j w_i^{enc} + b_{out}),$$

where $W_{out}^j \in \mathbb{R}^{|T^j| \times d_{model}}, j = 1, 2, \ldots, |T|$ and b_{out} is the bias vector. The predicted morphological tag in a given tagset for a given word is the tag with the highest probability. See Figure 1 for the graphical representation of the model.

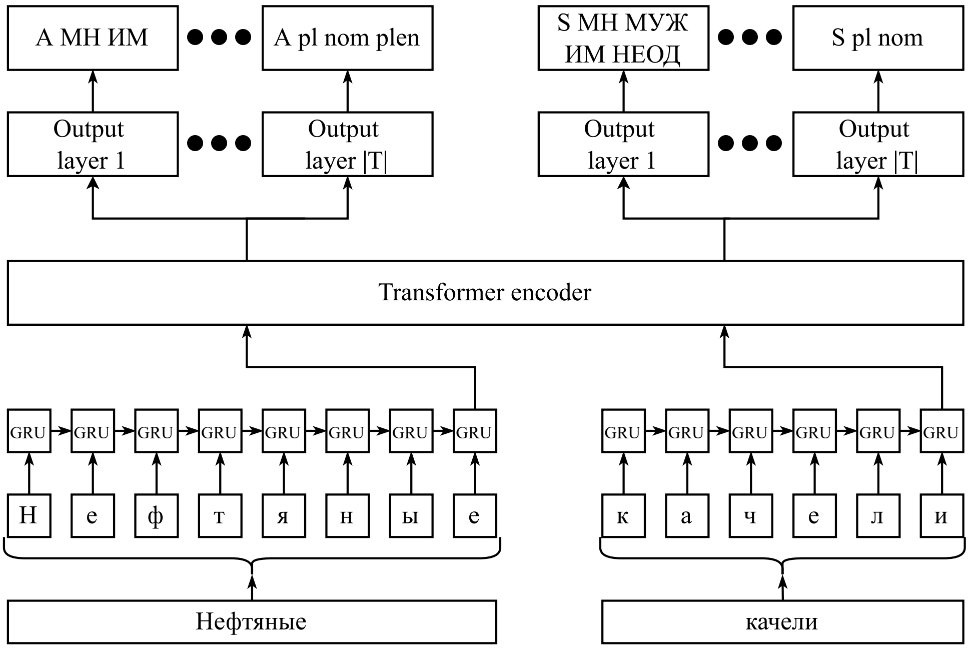

Figure 1: Graphical representation of the proposed neural tagger for the sentence *Neftjanye kačeli.*

2.2 Model hyperparameters

Table 1 shows the hyperparameters we used in our model. We fine-tuned these hyperparameters once and did not change them between the experiments.

Model part	Hyperparameter	Value
Word embeddings	$\|V^{char}\|$	95
	$char\_embedding\_size$	32
	$char\_hidden\_size$	128
Encoder layer	d_{model}	128
	d_{ff}	512
	P_{drop}	0.1
Output layer	$\|T^j\|$	Depends on the corpus

Table 1: Hyperparameters of the proposed model. We did not mention some hyperparameters in the section 2.1 and for them, we either use the notation proposed in the paper (Vaswani et al., 2017) or not mention them at all if we did not make any changes.

We used the Adam optimiser with weight decay (Loshchilov and Hutter, 2019). Its hyperparameters as well as learning rate function are almost identical to (Vaswani et al., 2017) except we chose $warmup\_steps$ to be 10% of the total number of steps.

One problem with our model is the training process. We used cross-entropy as the cost function, but each task (each output layer) has its own cost function and simply adding them up may affect performance since different corpora have different sizes. To overcome this issue, we adopted the approach proposed in (Cipolla et al., 2018). Namely, before adding up, it weighs each cost function by considering the homoscedastic uncertainty of each task.

3 Experiments

We chose eight different corpora to evaluate our model. We divided them into two parts to conduct two different sets of experiments. The first part consists of the manually (re)annotated corpora:

1. Syntactically tagged Russian text corpus SynTagRus (Inšakova et al., 2019). It is a subcorpus of the National Corpus of the Russian language. SynTagRus is supplied with several types of annotation, including fully disambiguated and manually corrected morphological and syntactic annotation.
2. Disambiguated subcorpus of the National Corpus of the Russian language (RNC) (Plungjan and Sičinava, 2004). This subcorpus was manually disambiguated, and it provides full morphological annotation.
3. Russian Universal Dependencies Treebank annotated and converted by Google (GSD)[4]. GSD is a small treebank automatically annotated and converted into UD format. The current version was manually reannotated and provides full morphological and syntactic annotation.
4. Russian Universal Dependencies Treebank based on data samples extracted from Taiga Corpus and MorphoRuEval-2017 and GramEval-2020 shared tasks collections (Taiga)[5]. It includes manually corrected morphological and syntactic annotation.

The second part consists of the corpora provided by the organisers of the MorphoRuEval-2017 contest (Sorokin et al., 2017):

1. UD SynTagRus. It is the SynTagRus corpus automatically converted into UD format.
2. RNC Open. It is a smaller part of the RNC corpus mentioned above being automatically converted into UD format.
3. GICR. It is a morphologically disambiguated part of the General Internet Corpus of Russian (Piperski et al., 2013). It was automatically annotated and then converted into UD format.
4. OpenCorpora. It is a morphologically disambiguated part of the OpenCorpora project[6]. It was manually annotated and then automatically converted into UD format.

We tackled some corpora differently from others. The first difference is how we split the corpora into training, development and test sets. GSD and Taiga corpora have predefined splits, so we left it as is. For SynTagRus and RNC, we used their intersection as test sets and split the remaining sentences randomly so that 10% of the sentences form a development set. For the remaining four corpora, the organisers of the MorphoRuEval-2017 contest provided a shared test set, so we split these corpora into train and development sets with the ratio 9:1, respectively.

The second difference is how we collected grammemes. We used the tagset descriptions provided with SynTagRus, RNC, GSD and Taiga and then omitted all non-inflectional features. For the remaining four corpora, we used only those grammemes which were counted at the testing phase of the MorphoRuEval-2017 contest.

To collect the tagset of a corpus, we followed the following algorithm:

1. Collect each word's tag from a corpus.
2. Exclude unused grammemes from each tag.
3. Remove duplicate grammemes from each tag (in case of annotation errors).

[4] https://universaldependencies.org/treebanks/ru_gsd/index.html

[5] https://universaldependencies.org/treebanks/ru_taiga/index.html

[6] http://opencorpora.org/

4. For SynTagRus and RNC: replace each tag in which any grammatical category has two or more different values with a special "erroneous" tag.
5. Sort grammemes in each tag.
6. Return unique preprocessed tags.

See Table 2 for the detailed statistics of each corpus.

Corpus name	#Sentences	#Words	#Grammemes	#Tags ($\|T^j\|$)
SynTagRus	97138	1685273	45	470
RNC	519726	7961784	62	1285
GSD	5030	98000	52	652
Taiga	17871	197001	54	683
UD SynTagRus	50116	931075	41	237
RNC Open	98892	1344875	41	492
GICR	83148	1086148	41	292
OpenCorpora	38508	457583	41	366

Table 2: Corpora statistics. We treat punctuation marks as words and POS features as grammemes.

We conducted two series of experiments. The first series concerns 4 corpora: SynTagRus, RNC, GSD and Taiga. They have different sizes, tagsets and annotation guidelines. We trained 15 different neural taggers using different subsets of corpora (one tagger for each of the 4 corpora, one tagger for each of the 6 pairs, one tagger for each of the 4 triples and one tagger trained on all 4 corpora) and compared their performance.

The second series of experiments concerns the remaining 4 corpora: UD SynTagRus, RNC Open, GICR and OpenCorpora. These corpora share the same tagset, they are similar in size, but they follow different annotation guideline. We trained and evaluated 15 different neural taggers in the same way as in the first series, but because the tagsets are the same, we were able to train another *combined* tagger using a single merged corpus which consists of all 4 corpora. For that final experiment, we also merged the corpora's tagsets.

Each tagger has the same model architecture described in section 2.1. We trained each tagger for 10 epochs and chose the final parameters based on the best development set performance. We did not use fixed mini-batch size because different sentences vary in size dramatically. Instead, each mini-batch contained some sentences of the same length from the same corpus with the overall limit of 2048 words per mini-batch. Since each corpus has morphological annotation for only one output layer, we froze the weights of other output layers during training, depending on to which corpus the sentences from the current mini-batch belong.

4 Results

To compare the taggers, we used per-word and per-sentence accuracy. The word is tagged correctly if the tag predicted by tagger is the same as in the gold standard (it means that the tags' grammemes also match). The sentence is tagged correctly if each word's tag match with the corresponding tag in the gold standard.

Figure 2 illustrates the per-word tagging accuracy on the test sets for each tagger from the first series of experiments. We arranged the models in ascending order of their joint corpora size. Each line corresponds to the respective output layer, so different lines also correspond to different test sets.

From the results, it is clear that low-resource corpora always benefit from multitask learning scenario when trained jointly with the larger corpora, despite their tagsets and annotation guidelines. The opposite does not hold. However, Taiga and GSD generally benefit from each other, as well as SynTagRus and RNC. We speculate that this might be due to two reasons. The first reason is that these two pairs have comparable corpus size. The second reason is that GSD and Taiga have almost identical tagsets.

Another finding is that despite the single-task learning models show a clear trend “more data — better

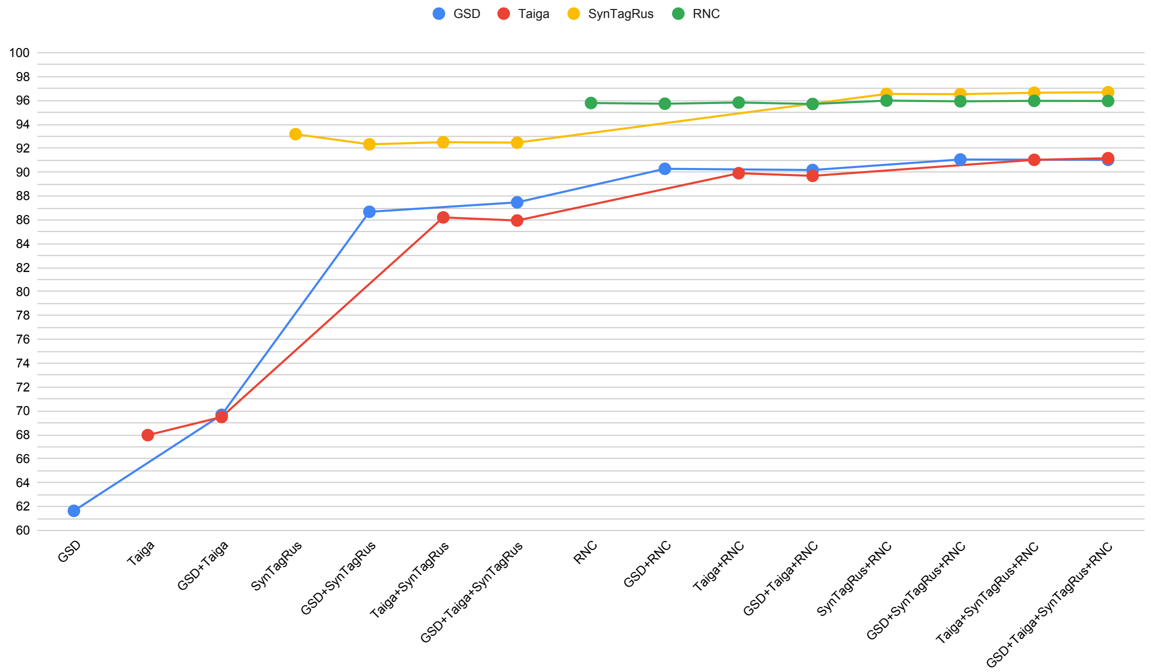

Figure 2: Per-word tagging accuracy (%) on the test sets for each model from the first series of experiments. Each line shows accuracy with respect to its output layer. The models are arranged in ascending order of their joint corpora size.

performance", the SynTagRus corpus shows the best overall performance. We believe that this is because SynTagRus has relatively small tagset, and it suffers less from the data sparseness problem.

One limitation of our comparison is the fact that we fine-tuned the model architecture's hyperparameters using the SynTagRus+RNC pair, which might be the reason why these two corpora benefit from each other. At the same time, the best performance for each test set provide the largest or the second-largest model in terms of joint corpora size. This contrasts with the paper (Mishra, 2019): the authors utilised a similar multitask learning approach to do POS tagging of English tweets, but did not improve the results for all corpora compared to a single-task learning approach.

Figure 3 illustrates the performance of the taggers from the second series of the experiments in the same manner as in Figure 2. This chart has two key differences from the previous one. The first difference is that the performance of the model for a given tagset does not depend on the tagset's corpus size at all: the largest corpus is RNC Open, and it performs poorly compared to the UD SynTagRus and GICR corpora. This appears to be the case of annotation guidelines differences. Since all these models share the same test set, the results show which corpus' annotation guideline is closer to the test set's one. This agrees with the fact that according to (Sorokin et al., 2017) the test set is the GICR subcorpus.

The second difference is the fact that here each corpus benefits from all others. This does not contradict the previous findings because all these corpora have comparable size. One exception which is visible on the UD SynTagRus line has already been explained: using data from the GICR corpus leads to better performance.

The best performance of the second series of experiments achieved by the largest model with the GICR tagset prediction layer. We compared our best model with the models provided by the participants of the MorphoRuEval-2017 contest in a closed setup, since we did not use any extra resources. We also added our combined model mentioned in section 3 into comparison. The results are shown in Table 3.

Our best model performance is comparable to the performance of the contest participants' models,

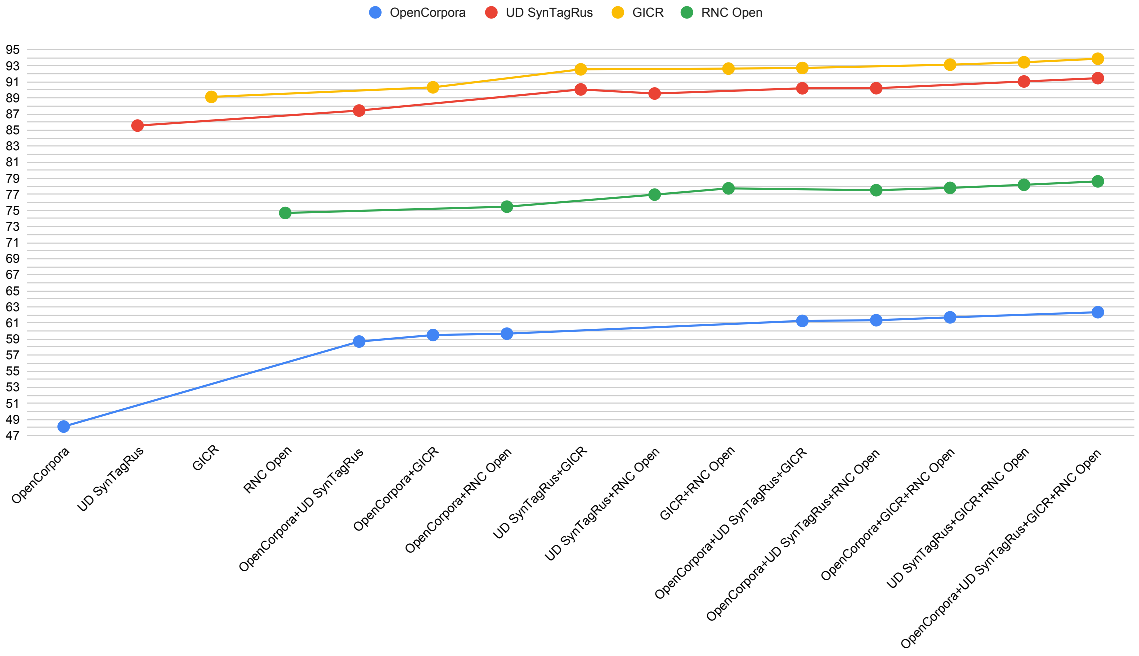

Figure 3: Per-word tagging accuracy (%) on the test sets for each model from the second series of experiments (excluding the combined model). Each line shows accuracy with respect to its output layer. The models are arranged in ascending order of their joint corpus size.

Model name	Per-word accuracy, %	Per-sentence accuracy, %
OpenCorpora+UD SynTagRus+ +GICR+RNC Open (GICR tagset output)	93.88	62.58
Combined	91.25	52.65
MSU-1	93.39	65.29
IQUMEN	93.08	62.71
Sagteam	92.64	58.40
Aspect	92.57	61.01

Table 3: Comparison of our models with the top 4 models provided by the participants of the MorphoRuEval-2017 contest on the test set in a closed setup.

although we did not use any dictionaries or hand-crafted features. We achieved the best per-word accuracy and third best per-sentence accuracy. The comparison with the combined model provides supporting evidence that even corpora with a shared tagset may perform poorly when merged together because of the differences in the annotation guidelines.

5 Conclusion

In this paper, we proposed a multitask learning based approach to Russian neural morphological tagging, which effectively utilises multiple corpora with different tagsets or annotation guidelines. To our knowledge, we for the first time applied the multitask learning technique in terms of predicting tags from different tagsets to the task of morphological tagging of Russian texts.

We showed that the effectiveness of morphological tagging depends on corpora size, tagset size and annotation consistency. Our findings help to better understand how tagset conversion affects performance of NLP tasks.

Our model is able to indirectly make tagset conversion in a scalable way taking into account differences in the morphological annotation guidelines, but full morphologically annotated corpora conversion does not end there. Such corpora often have other differences, including tokenisation and lemmatisation scheme. This may constitute the object of future studies.

Acknowledgements

This research has been partially supported by the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793.

References

Ivan Andrianov and Vladimir Mayorov. 2017. Transfer learning for morphological tagging in Russian. // *2017 Ivannikov ISPRAS Open Conference (ISPRAS)*, P 58–63.

V. V. Bočarov, S. V. Alekseeva, D. V. Granovskij, E. V. Protopopova, M. E. Stepanova, and A. V. Surikov. 2013. Crowdsourcing morphological annotation. // *Papers from the Annual International Conference "Dialogue" (2013)*, volume 1, P 109–114.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. // *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, P 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. // *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 7482–7491.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kira Droganova and Daniel Zeman. 2016. Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies. Technical report, ÚFAL MFF UK.

Jirka Hana and Anna Feldman. 2010. A positional tagset for Russian. // *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Georg Heigold, Guenter Neumann, and Josef van Genabith. 2017. An extensive empirical evaluation of character-based morphological tagging for 14 languages. // *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, P 505–513.

Evgenija Inšakova, Leonid Iomdin, Leonid Mitjušin, Viktor Sizov, Tat'jana Frolova, and Leonid Cinman. 2019. SynTagRus segodnja [SynTagRus today]. *Trudy russkogo jazyka im. V. V. Vinogradova*, P 14–40.

Zhenghua Li, Jiayuan Chao, Min Zhang, and Wenliang Chen. 2015. Coupled sequence labeling on heterogeneous annotations: POS tagging as a case study. // *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, P 1783–1792, Beijing, China, July. Association for Computational Linguistics.

O. N. Ljaševskaja, I. Astaf'eva, A. Bonč-Osmolovskaja, A. Garejšina, Ju. Grišina, V. D'jačkov, M. Ionov, A. Koroleva, M. Kudrinskij, A. Litjagina, E. Lučina, E. Sidorova, and S. Toldova. 2010. NLP evaluation: Russian morphological parsers. // *Papers from the Annual International Conference "Dialogue" (2010)*, P 318–326.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. // *International Conference on Learning Representations*.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. // *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, P 1211–1220, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Olga Lyashevskaya, Victor Bocharov, Alexey Sorokin, Tatiana Shavrina, Dmitry Granovsky, and Svetlana Alexeeva. 2017. Text collections for evaluation of Russian morphological taggers. *Jazykovedny Casopis*, 68(2):258–267.

Shubhanshu Mishra. 2019. Multi-dataset-multi-task neural sequence tagging for information extraction from tweets. // *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, P 283–284.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. // *Proceedings of the 12th Language Resources and Evaluation Conference*, P 4034–4043, Marseille, France, May. European Language Resources Association.

Alexander Piperski, Vladimir Belikov, Nikolay Kopylov, Vladimir Selegey, and Sergey Sharoff. 2013. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. // *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, P 24–29.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. // *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, P 742–751.

V. A. Plungjan and D. V. Sičinava. 2004. Nacional'nyj korpus russkogo jazyka: opyt sozdanija korpusa tekstov sovremennogo russkogo jazyka [Russian National Corpus: Experience in creating a corpus of texts of the modern Russian language]. // *Trudy meždunarodnoj konferencii «Korpusnaja lingvistika-2004»*, P 216–238.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. // *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Aleksei Sorokin, Tatiana Shavrina, Olga Lyashevskaya, Victor Bocharov, Svetlana Alexeeva, Kira Droganova, Alena Fenogenova, and Dmitry Granovsky. 2017. MorphoRuEval-2017: an evaluation track for the automatic morphological analysis methods for Russian. // *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017)*, volume 1, P 297–313.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.