sciendo
DOI 10.2478/jazcas-2021-0054

LINGUISTIC ANNOTATION OF TRANSLATED CHINESE TEXTS: COORDINATING THEORY, ALGORITHMS AND DATA

KIRILL I. SEMENOV[1] – ARMINE K. TITIZIAN[2]
– ALEKSANDRA O. PISKUNOVA[2] – YULIA O. KOROTKOVA[2]
– ALENA D. TSVETKOVA[2] – ELENA A. VOLF[2]
– ALEXANDRA S. KONOVALOVA[2] – YULIA N. KUZNETSOVA[3]
[1] Charles University, Prague, Czech Republic
[2] National Research University “Higher School of Economics”, Moscow, Russia
[3] Lomonosov Moscow State University, Moscow, Russia

SEMENOV, Kirill I. – TITIZIAN, Armine K. – PISKUNOVA, Aleksandra O. – KOROTKOVA, Yulia O. – TSVETKOVA, Alena D. – VOLF, Elena A. – KONOVALOVA, Alexandra S. – KUZNETSOVA, Yulia N.: Linguistic annotation of translated Chinese texts: Coordinating theory, algorithms and data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 590 – 602.

Abstract: The article tackles the problems of linguistic annotation in the Chinese texts presented in the Ruzhcorp – Russian-Chinese Parallel Corpus of RNC, and the ways to solve them. Particular attention is paid to the processing of Russian loanwords. On the one hand, we present the theoretical comparison of the widespread standards of Chinese text processing. On the other hand, we describe our experiments in three fields: word segmentation, grapheme-to-phoneme conversion, and PoS-tagging, on the specific corpus data that contains many transliterations and loanwords. As a result, we propose the preprocessing pipeline of the Chinese texts, that will be implemented in Ruzhcorp.

Keywords: Mandarin, Russian, parallel corpus, Chinese word segmentation (CWS), grapheme-to-phoneme conversion (G2P), PoS-tagging, code-switching detection

1 INTRODUCTION

Linguistic annotation is one of the key concepts for current corpus linguistics. Chinese has its unique aspects of linguistic annotation, namely: absence of word segmentation conventions; Chinese characters (with a high degree of homophony and homography); significant distinctions in the morphosyntactic system between the Chinese and the European languages. All the above-mentioned problems are compounded if a sentence contains loanwords, as their phonological, morphological, and orthographic features usually contradict the standard parameters of Chinese words.

The problem of proper annotation of Chinese texts that contain loanwords and transliterations became crucial for the project of the Russian-Chinese parallel corpus (hereinafter – Ruzhcorp; [1]) – a project within the Russian National Corpus. The collection of texts in Ruzhcorp comprises 1070 documents, and the total number of tokens (Russian and Chinese) is more than 3.5 million. The majority of the texts belongs to the fiction domain (81%), and news articles (11%).

Until recently, the linguistic annotation of Ruzhcorp has been inappropriate. The Chinese word segmentation algorithm (henceforth – CWS) was based on a variant of simple greedy search over a pre-loaded dictionary. This caused problems with the detection of Russian loanwords, as they are usually absent in the dictionaries. The algorithm of pinyin (official romanisation in PRC) attribution assigned all the possible readings to each of the characters. Finally, there was no morphosyntactic annotation (hereinafter – PoS-tagging) for Chinese texts.

Our research was aimed to create a proper pipeline of linguistic annotation of the Chinese texts for Ruzhcorp, that would: i) be consistent regarding the linguistic theory; ii) show appropriate results for the original Chinese texts; iii) show appropriate results for detection of Russian transliterations and loanwords in the Chinese texts. Speaking about the necessary layers of annotation, the pipeline should include CWS, PoS-tagging, and pinyin annotation (hereinafter – G2P from "grapheme-to-phoneme"). Within this article, we are going to provide an overview of aspects i and iii, as, relating to aspect ii, we are relying on the analyses carried out by the research community. In Part 2, we present the theoretical comparison of the CWS and PoS-tagging standards for Chinese. In Part 3, we describe the set of experiments in CWS, PoS-tagging and G2P on the Chinese texts of Ruzhcorp. In Part 4, we propose the final model for Chinese linguistic annotation for Ruzhcorp, based on our theoretical and empirical comparisons.

2 THEORETICAL COMPARISON OF THE STANDARDS OF CHINESE LINGUISTIC ANNOTATION

2.1 Chinese word segmentation

The concept of "word" in Chinese is a challenging issue. Firstly, there are no spaces in Chinese, and secondly, a character, not a word, was traditionally considered a linguistic unit. But with the growing necessity of tokenization for different NLP tasks, several segmentation standards were developed – [2]. Every standard tends to focus on one of the language levels: morphosyntax, semantics or lexicology. The comparative table of the standards is shown in Table 1.

Standard (Abbreviation)	Basic principle	Description
GB T 13715-1992	lexicography, semantics	The oldest standard, implemented in mainland China in 1993. The standard lacks theoretical foundation and seems too arbitrary.
Peking University standard (PKU)	lexicography, semantics	Based on GB T 13715-1992 with some rules redefined. Segmentation units are determined by lexical semantics and lexical combinability.

Standard (Abbreviation)	Basic principle	Description
CNS 14366 (hereinafter – CNS)	morphology, syntax, semantics	Taiwanese national standard, implemented in 1999 by Academia Sinica. The standard operates with various linguistic concepts, such as morpheme, affix, dependent word and so on, and its rules are more consistent.
Penn Chinese Treebank standard (CTB)	syntax	Based on X-bar syntax theory, with every constituent that can take X^0 position considered as a word. The standard was created specifically for Chinese treebank, and is compatible with Universal dependencies tagset.
Microsoft Research Asia (MSR)	morphology, syntax, semantics	Developed its word taxonomy: lexical word, morphologically derived word, factoid, new word and named entity, all of them are processed in different way. This standard is not self-sufficient and oriented towards compatibility with others: PKU, CNS and CTB.
Vocabulary standards	lexicography	Not holistic standards, because the main rule for them is to consider as a word every unit that is found in a vocabulary.

Tab. 1. Comparison of CWS standards

As we can see, CNS and CTB standards seem to be more systematic and have a strong theoretical background, which makes them both preferable standards.

2.2 PoS-tagging

Tagsets for automatic annotation of Chinese also vary in criteria for parts of speech distinction and number of categories.

One of the most widely used tagsets is Peking University morphosyntax-based standard (PKU) and its modifications. It includes 26 basic categories and up to 46 subcategories, including denoting semantic and morphological classes within basic categories.

The ICTCLAS tagset was made by the Institute of Computer Science, Chinese Academy of Sciences. This is one of the few standards for Chinese that proposes a hierarchical model of morphosyntactic tags with three levels, where the first one denotes parts of speech, and the two latter denote other categories (primarily semantic ones). This standard is one of the most numerous, with more than 90 different tags.

Chinese National Standard (CNS) has a highly detailed list of about 150 tags. Although the main criterion for selection is morphosyntactic properties, the categories highly depend on semantics as well. Although the explanatory power of this standard is high, due to the number of tags, this standard is difficult to be implemented by automatic taggers.

Universal Dependencies (UD), a syntax-based tagset, offers only 15 to 17 clear categories, which makes it convenient for cross-language annotation, but appears to be less distinctive for Chinese than it should be.

Finally, Penn Chinese Treebank (CTB) 3.0, which was a prototype for Chinese UD, is based on the principle of syntactic distribution and has 33 tags. The moderate number of tags and the principles of their attribution that can be modelled through programming means make CTB the most applicable PoS standard.

3 EXPERIMENTS WITH LINGUISTIC ANNOTATION ON RUZHCORP DATA

3.1 Data

Ruzhcorp data have substantial differences from the “standard” Chinese texts, as they contain phonetical borrowings and transliterations, which sum up to several thousand. The majority of these unusual tokens occur either in the texts translated from Russian or in the texts that describe Russian realities. Most of these tokens constitute transliterations of Russian proper names (toponyms and anthroponyms), thus, hereinafter we will focus only on the phonetical transliterations of the proper nouns and will use “loanwords” and “transliterations” as synonyms.

To evaluate the performance of the algorithms that cover features of CWS, PoS-tagging and G2P, we created the datasets on Ruzhcorp data, which share three common features:

1. Separate datasets for fiction (Russian-to-Chinese translations) and news domains (articles in Chinese media about Russia).
2. Sentences in each dataset are balanced (each document does not exceed 8–10% of the dataset) and randomized.
3. Objects in each dataset have common features (Russian and Chinese sentences) and the features specific to this dataset. These peculiarities, as well as the quantitative overview of each dataset, are presented in Table 2.

Dataset[1]	**Size (sentence pairs)**	**Features**	**Used in**	**Purpose**
BOOKS_1/NEWS_1	436/78	(automatically) Extracted Russian proper names + their (manually) extracted transliterations. Only sentences with Russian proper names.	CWS, PoS-tagging	evaluation
BOOKS_2/NEWS_2	688/158		CWS, code-switching	fine-tuning
BOOKS_3/NEWS_3	>800/>400		CWS (future)	fine-tuning

[1] The prefix BOOKS means the data are taken from fiction literature, NEWS – from the news articles. The “size” column values are separated by slash for BOOKS_x and NEWS_x datasets, respectively.

Dataset[1]	Size (sentence pairs)	Features	Used in	Purpose
BOOKS_G2P/ NEWS_G2P	650/700	Manual pinyin annotation of the whole sentences.	G2P	evaluation

Tab. 2. Description of the datasets

3.2 Experiments in CWS

3.2.1 Comparison of the best performing CWS algorithms without fine-tuning

Our first task was to evaluate the performance of different CWS algorithms regarding the identification of transliteration boundaries in the sentences. Firstly, we have tested the following algorithms that are widely used for the CWS task and show high quality on default Chinese texts.

Algorithms	Architecture	CWS standards
Ckiptagger [3]	neural network: bidirectional LSTM and multi-head attention layers	CNS
Stanza [4]		CTB
SpaCy [5]		CTB
Pkuseg [6]	neural network: adaptive online gradient descent	PKU
FastHan [7]	neural network: BERT	PKU, CNS, CTB, MSR (different pretrained variants)
NLPIR [8]	dictionary-based method followed by a k-shortest path routing	dictionary
LTP [9]	neural network: ELECTRA	PKU
UDPipe [10]	neural network: bidirectional GRU	CTB

Tab. 3. Overview of the considered CWS algorithms

To compare the algorithms, we used two datasets – BOOKS_1 and NEWS_1. We applied all the above-mentioned algorithms to the datasets and calculated three metrics for each algorithm: recall, F-score, and our metric (hereinafter – “our”) that penalizes models for both overtokenization (segmentation of one loanword into more tokens) and undertokenization (setting broader boundaries for a loanword than necessary). The original metric was designed because traditional metrics do not properly reflect the boundaries of the tokens, rather aiming at their number in a sentence. The results on the Ruzhcorp data are represented in the following figure.

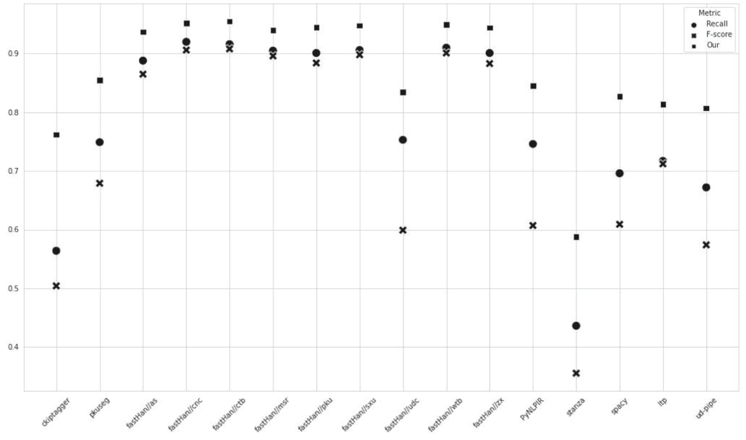

Fig. 1. Results of CWS algorithms on BOOKS_1 data

The quantitative analysis demonstrates, firstly, that despite its widespread, graph-based model, NLPIR was not as good as neural networks. Secondly, monolingual Chinese models turned out to be better than multilingual Stanza and UDPipe. Thirdly, in some cases, the performance of algorithms correlates with the CWS standards (CNS-based Ckiptagger performs worse than PKU-based PKUSeg). However, we cannot state a causal link, because fastHan, which has a CNS variant, performs as well as its PKU or CTB versions.

The qualitative analysis made us discover some inconsistencies within CWS standards and algorithm performance. For instance, the multi-word Russian personal names (e.g, first name and patronymic) are divided only by some segmenters, such as the middle dot – a special symbol "·" in the Mandarin orthography. The standards regard such clusters variously as well: PKU and CNS prescribe not to divide them, while CTB does not. We believe that it is necessary to split the multi-word transliterations by the middle dot, as this symbol is proof that the Chinese speakers are aware of the multi-word nature of these items.

Based on our analysis, we identified fastHan and its CTB-based versions (like fastHan//ctb or fastHan//wtb) as the best algorithm for our Corpus.

The detailed results of the study are represented in [11].

3.2.2 Experiment with fine-tuning FastHan algorithm

Another advantage of fastHan is a built-in fine-tuning function. Thus, we decided to test whether the fine-tuned algorithms would perform better on our data. We fine-tuned the best-performing variants of fastHan based on three main CWS standards: CNS, CTB, PKU. We used BOOKS_2 and NEWS_2 datasets. We passed datasets to CNS-based, CTB-based and PKU-based models, accordingly.

Unexpectedly, the performance of the fine-tuned algorithms after testing on BOOKS_1 and NEWS_1 slightly degraded. The table provides a comparison between models' metrics before fine-tuning and after it.

	Before fine-tuning			**After fine-tuning**		
Metric	**FastHan// PKU**	**FastHan// CTB**	**FastHan// CNS**	**FastHan// PKU**	**FastHan// CTB**	**FastHan// CNS**
Recall	0.9030	0.9180	0.9214	0.8997	0.9064	0.9080
F-score	0.8860	0.9097	0.9077	0.8841	0.8988	0.8902
Our	0.9488	0.9579	0.9554	0.9468	0.9493	0.9493

Tab. 4. Comparison of the FastHan algorithms before and after fine-tuning

We suggest that the main reason was the following: there was a small overlap between the set of proper names in our BOOKS_1+NEWS_1 (test) and BOOKS_2+NEWS_2 (fine-tuning) datasets, as different documents were taken. The overlap comprises only 10 words, which is less than 10% overlap in the test dataset and less than 5% – in the training dataset, thus, the model did not "learn" how to tokenize exact proper nouns in the test dataset.

Currently, we are compiling another dataset – BOOKS_3 and NEWS_3, sharing the same text sample as in a test dataset and being of bigger size, to proceed with experiments in more representative fine-tuning.

3.2.3 Experiment in code-switching detection

Another hypothesis for handling the problem of transliterations was not to fine-tune the CWS models but to use a different module that would be aimed only at code-switching detection, which, in our case, would mean the transliterated Russian proper nouns. As we approach this task, it can be treated as sequence labelling.

To do this, we ascribed labels to the transliterations in BOOKS_2 dataset and trained the LSTM and CRF layers of fastHan algorithm on it. The NEWS_2 dataset was used to check the performance on out-of-domain data.

The results of the experiment are decent, as the table below represents, however, we do not consider them reasonable to add the gained increase to the main pipeline because of lack of training data. Moreover, the performance on OOD data is worse than the original fastHan, thus we conclude that this technique to adjust the quality is needless for our task.

	Fine-Tuning Data (BOOKS_2)			**Test on out-of-domain data (NEWS_2)**		
Metric/Model	FastHan//CNS	FastHan//CTB	FastHan//PKU	FastHan//CNS	FastHan//CTB	FastHan//PKU
Recall	0.8990	0.9296	0.9292	0.7861	0.8181	0.8094
F-score	0.8824	0.9189	0.9226	0.7712	0.8054	0.7882

Tab. 5. Results of code-switching detection experiment

3.3 Experiments in G2P

The main problem of the G2P task for Chinese is that many Chinese characters have multiple phonetic representations depending on the word or syntactic position, so disambiguation of the readings for each character appears to be the key challenge for the task. In general terms, Chinese G2P annotation includes the following steps: word segmentation (and possibly PoS-tagging), obtaining all possible phonetic values for a token, and applying a set of heuristics to choose the most relevant transcription. Therefore, the quality of the pinyin annotation depends on the quality of the previous part(s) of a pipeline, CWS, and PoS-tagging.

In this study, we tested the following G2P algorithms for pinyin annotation. G2pC [12] is based on recurrent neural networks. G2pM [13] is a package with bidirectional LSTM architecture. The Xpinyin [14] model is based on stochastic decision lists using frequencies of pinyin. Pypinyin [15] library uses n-gram statistics and has an in-built collocation dictionary. The G2pC model is the only one to use an external application for CWS and PoS-tagging. Thus, we used the G2pC model with different tools for CWS: PKUSeg, a default model, fastHan and UDPipe.

For the test, we used two manually annotated datasets, BOOKS_G2P and NEWS_G2P, consisting of 1350 annotated sentences. For each character, a pinyin annotation was ascribed. Table 6 presents accuracy scores on the test dataset for each model.

Model	Performance (Accuracy)
G2pC (PKUSeg)	0.7347
G2pC (FastHan)	0.7304
G2pC (UDPipe)	0.7239
G2pM	0.5607
Xpinyin	0.5457
Pypinyin	0.5459

Tab. 6. Comparison of the phonetic annotation results

The best model is G2pC with PKUSeg word segmenter. PKUSeg is pre-trained on several datasets of different domains (medicine, art, etc.) which may help it perform on new data better than other models which are mainly trained on news texts. However, G2pC with fastHan word segmentation shows almost the same performance as the default CWS model.

The detailed results are represented in [16].

3.4 Experiments in PoS-tagging

3.4.1 Comparison of the best performing Chinese PoS-taggers

Regarding PoS-taggers, our first interest was to compare their performance on transliterated toponyms and anthroponyms specifically. For the first PoS-tagging experiment, we examined a group of algorithms represented in the Table below. We can see that almost every tool uses a different tagset, which were compared in 2.2. Regarding the problem of loanwords, we distinguished three groups of tags: for anthroponyms; for toponyms; for more common classes or other lexical classes of the proper names (for example, common nouns or all nouns).

Tool	PoS Tagset	Anthroponyms	Toponyms	More common and related classes
Ckiptagger [3]	Chinese national standard (CNS)	Nb	Nc	Na
PKUSeg [6]	Peking university (PKU)	nr	ns	n, nz
FastHan [7]	Penn Chinese Treebank (CTB)	NR	NR	NN
PyNLPIR [17]	PKU (modified)	nrf	nsf	n, nr, ns, nt, nz
Stanza [4]	Universal Dependencies + CTB (UPOS)	PROPN	PROPN	NOUN
SpaCy [5]	UPOS	PROPN	PROPN	NOUN
LTP [9]	PKU (modified)	nh	ns	n, ni, nz

Tab. 7. Comparison of Chinese PoS-tags that can be classified as borrowings

To compare the algorithms, we used BOOKS_1 and NEWS_1 datasets by taking the sentences, splitting them with CWS algorithms, and applying PoS-taggers. After that, we evaluated the PoS-tags of transliterations. We divided the ascribed PoS-tags into three groups: absolutely correct (when the tagger matches both the part of speech and the semantic class of the word), approximate match, when the tagger chose a morphosyntactically correct annotation but did not ascribe the exact lexical class (for instance, an anthroponym was marked as a toponym or a common noun), and all other cases that are error. The algorithms were evaluated by the F-score metric (Fig. 2).

According to the results, the best tool is fastHan, which has almost 100% correctness. The main errors of all algorithms occurred due to incorrect word segmentation (thus we did not analyse them precisely). Speaking about the mistakes among the correctly segmented words, a notable inaccuracy was marking anthroponyms as toponyms and vice versa. The possible explanation is that such words end with the morphemes that are usually used as semantic markers of the proper names from the opposite groups, so the tagger could decipher them as a generic element (see Conclusions) rather than the last character of the transliteration. The detailed results are represented in [18].

3.4.2 Experiments in parallel PoS-tagging

The method of parallel PoS-tagging is gaining popularity for the multilingual data: among two languages in the parallel corpus, the well-studied standard language for which the task of PoS-tagging is relatively well solved is used as an additional sequence of tags for labelling the under-resourced language. This approach was used either for low-resourced languages or for languages with grammar that differs significantly from European languages.

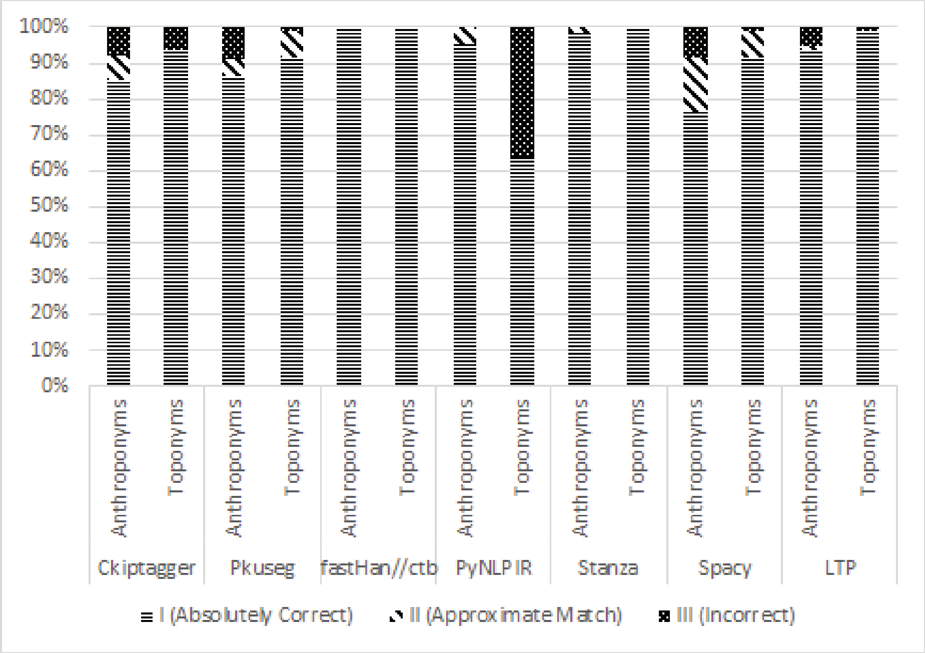

Fig. 2. PoS-taggers' results on our dataset

We implemented this approach to Ruzhcorp data. For that work, we used BOOKS_2 and NEWS_2 datasets, with manual word-to-word alignment. The data were divided at a ratio of 7 to 3 into the training and the test sets. Russian sentences were parsed by a morphological annotator Pymorphy2 and the Chinese were processed by fastHan. Then the two sequences of PoS-tags were given as two inputs for a BiLSTM-based neural network. The accuracy on the test set reached a slightly better score of 0.98 compared to 0.97 produced by a default fastHan model. We consider that an interesting source for further research, however, this approach appears to be excessive in production as it requires word-to-word alignment of data and cannot handle raw text.

The detailed results of the experiment are described in [19].

4 THE FINAL MODULE FOR LINGUISTIC ANNOTATION OF THE CHINESE TEXTS IN RUZHCORP

After reviewing all the results, we merged all the modules into one algorithm of Chinese text processing. Our algorithm consists of CWS, PoS-tagging, and G2P functions and a module with custom rules that split the multi-word transliterations by the middle dot (see 3.2.1). Each module is applied sequentially: the CWS, PoS-tagging, G2P and custom rules are applied one after another and take the outputs of the previous module as inputs.

While working on this algorithm, we had to decide on which standard it should be based and which tools we should use for each task. Our decision is based not only on the idea of using the best standards and tools for each task, but also on the idea of making all the tools work in harmony in our algorithm. In case of CWS and PoS-tagging tasks, there is no problem as both of these modules perform best results using fastHan based on CTB standard. This standard is considered the best for processing Chinese texts because, unlike other standards, it is centred on syntactic structure, which is more relevant to Chinese than morphological and lexico-semantic features. However, considering the G2P task, it is a little more difficult because the best tool for it is G2pC, which uses PKUSeg as a word segmenter and a PoS-tagger. Nevertheless, we decided to implement CTB-based fastHan into G2pC although this implementation performed negligibly worse performance than the default G2pC as was shown in Section 3.3.

G2pC, unlike other tools, takes into consideration CWS and PoS-tagging annotation and uses this information to solve the ambiguity problem, which explains its good performance. All other algorithms take only CWS as input, which logically lowers their results. This allows us to conclude that it is more rewarding to create a "sequential" structure of CWS, PoS-tagging, and G2P modules (each module takes as input the output of the previous module) rather than a "parallel" structure (PoS-tagging and G2P modules take only CWS results independently).

The code-switching detection was not included because it works worse than algorithms for the CWS task on their own. Parallel PoS-tagging showed better results than monolingual PoS-taggers, but it cannot be used for the annotation in our corpus, at least for now, as it requires Russian sentences with a deeper manual markup.

The code is available through this link: https://github.com/ruzhcorp/ruzhcorp_chinese_annotation.

5 CONCLUSIONS AND PERSPECTIVES

The paper presents a comparative analysis of the current situation in Chinese word segmentation, PoS-tagging, and automatic transliteration from both theoretical and experimental sides by using Ruzhcorp data. In terms of theory, the frameworks that fit the Russian-Chinese parallel corpus most are the syntax-based standards of both CWS and PoS-tagging (such as CTB or UD) and that the best G2P predictions are made with the use of information about tokenization and PoS-tags. From the technical perspective, the best algorithms are, firstly, based on the modern neural architectures (namely BERT, ELECTRA and RNN). Secondly, for Chinese-specific tasks like CWS, monolingual algorithms perform better than multilingual ones.

As a result of the set of experiments, we propose an algorithm that includes all three aspects of the Chinese linguistic annotation, and that features both neural and rule-based patterns. To date, all the texts in Ruzhcorp have been re-annotated with this algorithm and are available at the webpage https://ruzhcorp.github.io/.

There are areas of future research in that field. Firstly, our observations show inconsistencies in the detection of the so-called generic elements in Chinese: after a proper noun, a "generic" noun is used in order to denote the type of objects the name refers to. The CWS standards treat this phenomenon in significantly different ways, taking into account phonotactic (length of the generic element) or semantic features. Thus, we find it necessary to provide a specification of the CWS standard for Ruzhcorp, which will include a more consistent approach to generic elements. The second research area is deepening the experiments on parallel linguistic annotation. On the one hand, this can be conducted for scientific purposes, such as parallel PoS-tagging, on the other hand, this is a valuable help for the task of word-to-word alignment, which is rather aimed at corpus-aided language learning.

ACKNOWLEDGEMENTS

Our work is supported by the Commission for the Support of Educational Initiatives of the Faculty of Humanities of HSE University (project title – "Language-Specific Markup of Chinese Texts in the Russian-Chinese Parallel Corpus of RNC").

References

[1] Semenov, K. I., Kuznetsova, Y. N., and Durneva, S. P. (2020). Russian-Chinese parallel corpus of RNC: Problems and perspectives. Proceedings of the 10th International Conference "Russia and China: History and Perspectives for Cooperation", pages 633–640.

[2] Emerson, T. (2005). The Second International Chinese Word Segmentation Bakeoff. Accessible at: http://sighan.cs.uchicago.edu/bakeoff2005/.

[3] Li, P.-H., and Ma, W.-Y. (2019). CkipTagger. Accessible at: https://github.com/ckiplab/ckiptagger.

[4] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Association for Computational Linguistics (ACL) System Demonstrations. Accessible at: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[5] Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Accessible at: https://spacy.io/.

[6] Luo, R., Xu, J., Zhang, Y., Ren, X., and Sun, X. (2019). PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. Accessible at: http://arxiv.org/abs/1906.11455.

[7] Geng, Z., Yan, H., Qiu, X., and Huang, X. (2020). fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP. Accessible at: http://arxiv.org/abs/2009.08633.

[8] Zhang, H., and Shang, J. (2019). NLPIR-Parser: An intelligent semantic analysis toolkit for big data. Corpus Linguistics, 6(1), pages 87–104.

[9] Che, W., Feng, Y., Qin, L., and Liu, T. (2021). N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models. Accessible at: http://arxiv.org/abs/2009.11616.

[10] Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207. Accessible at: https://doi.org/10.18653/v1/K18-2020.

[11] Semenov, K. I., Korotkova, Y. O., Volf, E. A., and Konovalova, A. S. (2021). Automatic Annotation of the Chinese Texts that Contain Loanwords: Word Segmentation, Transcription, PoS-tagging. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies, Supplementary volume, pages 1081–1095. Accessible at: http://www.dialog-21.ru/media/5420/_-dialog2021supvol.pdf.

[12] Cai, Z., Yang, Y., Zhang, C., Qin, X., and Li, M. (2019). Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features. Accessible at: https://arxiv.org/abs/1907.01749.

[13] Park, K., and Lee, S. (2020). g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset. Accessible at: http://arxiv.org/abs/2004.03136.

[14] Luo, E. (2020). Xpinyin. Accessible at: https://github.com/lxneng/xpinyin.

[15] Huang, H. (2020). pypinyin. Accessible at: https://github.com/mozillazg/python-pinyin.

[16] Konovalova, A. S., and Tsvetkova, A. D. (2021). Comparative analysis of grapheme-to-phoneme models for the Russian-Chinese parallel corpus. Program book of Buckeye East Asian Linguistics Forum 4, pages 28–30. Accessible at: https://cpb-us-w2.wpmucdn.com/u.osu.edu/dist/6/3609/files/2021/03/BEALF-4_Program_Book_2021-3-5.pdf.

[17] Roten, T. S. (2018). PyNLPIR PoS tagset. Accessible at: https://pynlpir.readthedocs.io/en/latest/api.html.

[18] Semenov, K. I., Korotkova, Y. O., and Volf, E. A. (2021). Automatic Annotation of the Russian Loanwords in Chinese Texts: Issues in Word Segmentation and PoS-tagging. Proceedings of Corpora 2021 International Conference. 14 pages [in press].

[19] Konovalova, A. S. (2021). Automatic POS-tagging for Chinese Using Parallel Data [BA thesis]. Higher School of Economics. 82 pages.