***Вольф Е.А., Короткова Ю.О., Семенов К.И.***

**Автоматическая разметка заимствований из русского языка в китайских текстах: проблемы словоделения и морфопарсинга[[1]](#footnote-1)**

**Аннотация**: Статья посвящена вопросам автоматической аннотации китайских текстов в Русско-китайском параллельном корпусе НКРЯ в двух аспектах: разделении текста на слова и морфологического парсинга. Особенное внимание уделяется предложениям, содержащим фонетические заимствования из русского языка. Результаты исследований планируется применить при улучшении лингвистической разметки в русско-китайском параллельном корпусе НКРЯ.

**Ключевые слова***:*китайский язык, автоматическая сегментация, морфологическая аннотация, проблема слов вне словаря

**Automatic Annotation of the Russian Loanwords in Chinese Texts: Issues in Word Segmentation and PoS-tagging**

**Summary:** The article addresses the issues in the automatic annotation of the Chinese texts in the Russian-Chinese Parallel Corpus of RNC in two aspects: word segmentation, and PoS-tagging. We paid particular attention to the processing of the Russian loanwords that are abundant in the Corpus data. We plan to take the results of the research into consideration while creating the new preprocessing pipeline of the Chinese texts in the corpus.

**Keywords:** Mandarin, Chinese word segmentation (CWS), PoS-tagging, out-of-vocabulary problem (OOV)

1. **Введение**

Задачи токенизации и морфологической разметки китайских текстов связаны с рядом трудностей в области орфографии, фонетики, и морфосинтаксиса. Все вышеперечисленные проблемы усугубляются в случае, если в тексте присутствуют заимствованные слова.

С набором этих проблем столкнулась команда разработчиков Русско-китайского параллельного корпуса НКРЯ (далее – ruzhcorp). Несмотря на то, что на текущий момент ruzhcorp обладает несколькими уровнями разметки для китайских текстов, в т.ч. разделением на слова (далее – CWS от Chinese word segmentation), все они представляют собой достаточно примитивные алгоритмы, выдающие много ошибок при обработке заимствований.

В настоящей статье мы представляем два исследования по улучшению текущего положения дел в разметке китайских текстов в ruzhcorp: экспериментальную работу в области словоделения (Раздел 2) и PoS-тэггинга (Раздел 3).

1. **Исследования в области словоделения**
	1. ***Обзор стандартов и алгоритмов CWS и экспериментального набора данных***

В китайском языке отсутствует кодифицированный набор иероглифов, регулярно используемых при транслитерации заимствований. Более того, в китайской орфографии практически не существует маркеров, указывающих на начало или конец последовательности символов, которые необходимо читать «фонетически», игнорируя семантику иероглифов. Таким образом, одна и та же комбинация полнозначных китайских иероглифов может быть распознана и как набор несущих смысл китайских морфем, и как набор знаков, передающих только звучание.

Учитывая, что в ruzhcorp присутствуют как тексты, содержащие заимствования из русского языка в больших количествах, так и тексты большого размера, состоящие исключительно из «стандартных» китайских слов (например, китайская литература XX в.), задача CWS для корпуса распадается на два блока: во-первых, качественное разделение «стандартного» (т.е. не содержащего заимствований) китайского текста на слова; во-вторых, максимально корректное выделение фонетических заимствований. Для того, чтобы решить обе задачи, мы решили сравнить качество ведущих алгоритмов CWS для задачи выделения заимствований с последующей их доработкой.

В настоящий момент существует несколько популярных стандартов CWS, при этом ни один из них не является общепринятым ни в науке, ни в NLP. Стандарты словоделения могут принципиально различаться, так как они берут за основу различные аспекты и уровни языковой структуры – семантику, морфосинтаксис или лексику. Более того, для большинства из этих стандартов существует более одного алгоритма, реализующего этот стандарт. В нашем исследовании было использовано 17 алгоритмов и их вариантов . Краткое описание каждого стандарта словоделения и реализующих его алгоритмов представлено в Таблице 1.

*Таблица 1.* **Сравнительное описание стандартов CWS**

|  |  |  |
| --- | --- | --- |
| **Стандарт (по** [Emerson 2005]**)** | **Основные критерии** | **Алгоритмы, соответствующие стандарту** |
| Пекинский университет (PKU) | Лексическая семантика и сочетаемость  | PKUSeg [Luo и др. 2019], fastHan // {pku, sxu} [Geng и др. 2020], LTP [Che и др., 2021] |
| Academia Sinica (CNS) | Морфосинтаксис | Ckiptagger [Li и Ma 2019], fastHan // {as, cnc} |
| Microsoft Research Asia | Морфосинтаксис | fastHan // msr  |
| Penn Chinese Treebank (CTB); вариация – UD | Синтаксис | Stanza [Qi и др. 2018], fastHan // {ctb, udc, wtb, zx}, Spacy [Honnibal & Montani 2017], UDPipe [Straka 2018] |
| Словарные стандарты | Наличие слова в словаре | NLPIR [张[Zhang] & 商[Shang] 2019] |

Цель нашего исследования – определить наилучший алгоритм CWS из вышеперечисленных, который будет оптимально справляться с выделением заимствований из русского языка. Для проверки качества алгоритмов мы создали датасеты, состоящие из 408 предложений из художественной литературы и 87 предложений из текстов СМИ, в которых содержались заимствования.

Исследование качества алгоритмов CWS на наших данных включает как количественный, так и качественный их анализ. В силу ограничений по формату мы вынуждены представить здесь лишь наши соображения относительно качественного анализа; подробное описание количественного анализа представлено в [Семенов и др. 2021, в печати]. Отметим лишь, что наиболее эффективными оказались нейросетевые алгоритмы fastHan (в особенности на стандарте CTB), LTP и PKUSeg, дающие качество 80-95% на каждом датасете.

* 1. ***Сравнение алгоритмов CWS на данных корпуса: качественный анализ***

Ошибки, сделанные сегментаторами на наших данных, можно разделить на две большие категории. Первая представляет разделение на слова, которое несомненно нарушает целостность заимствований и некорректно с точки зрения любого алгоритма словоделения. Эти вхождения мы называем «однозначными ошибками». Вторая группа представляет два класса заимствований, к выделению которых различные стандарты CWS подходят по-разному; поэтому данная группа ошибок (по сути, являющаяся не в полном смысле слова «ошибками») названа «неоднозначными вхождениями».

Однозначные ошибки можно разделить на две категории – случаи чрезмерной токенизации (заимствование разделяется на большее, чем следует, количество слов) и недостаточной токенизации (в заимствование включаются соседние иероглифы из «стандартных» китайских слов).

Случаи чрезмерной токенизации можно объединить в несколько групп. Во-первых, часто выделяются иероглифы, выполняющие роль грамматических маркеров, расположенные после заимствований (например, локативный послелог *里, lǐ* или результативный маркер глагола *来, lái*). К более примечательным случаям относятся два типа вхождений. Так, заимствования из трех иероглифов, занимающие именные позиции в китайских предложениях, нередко делятся на два слова – односложное (первое) и двусложное (второе). Мы предлагаем объяснять этот тип ошибок структурой китайского личного имени: большинство китайских антропонимов состоят из односложной фамилии и следующего за ним двусложного личного имени. Так, транслитерация фамилии «Рогожин» разделяется сегментаторами PKUSeg и Ckiptagger на два слова: *罗 luō* (китайская фамилия Ло) и *戈任 gērèn* (может служить личным именем Гэжень).

Другой интересный тип ошибок связан с сегментацией четырехсложных заимствований: нередко они делятся на два двусложных слова. Мы предполагаем, что это следствие частотности двусложных китайских слов, которые, согласно [Wong, Xu 2010: 45], составляют 75% китайского лексикона. Согласно ряду работ, например [Duanmu 1999], тяготение китайского лексикона к двусложным словам объясняется фонотактическими причинами, поэтому мы называем такое поведение «метрической токенизацией». Итак, сталкиваясь с последовательностью из четырех иероглифов, алгоритм может посчитать, что более правдоподобным будет разделение на два двусложных слова, а не на одно четырехсложное.

Явления недостаточной токенизации обычно происходят в в случае, если не отделяется односложная лексема, идущая после длинного заимствования. На первый взгляд, этот процесс кажется обратным чрезмерной токенизации одиночных иероглифов в конце заимствований. Однако подробный анализ выявляет обратное: случаи недостаточной токенизации односложных лексем происходят с полнозначными глаголами (такими как *拿 ná* «брать, поднимать» и *说, shuō* «говорить»), в то время как чрезмерная токенизация происходит со служебными лексемами (например, с копулой *是 shì*).

Обратимся теперь к неоднозначным вхождениям. Первый их тип связан с, пожалуй, единственным указателем на иностранные имена, принятым в китайской орфографии – т.н. срединной точкой. Этот знак (·) употребляется для разделения имени собственного, которое в языке-доноре состояло из нескольких слов (в случае с русскими заимствованиями он будет ставиться, например, между именем и отчеством). Стандарт CNS предписывает разделять слова по этому знаку и считать их разными токенами, в то время PKU и CTB предлагают относиться ко всему транслитерированному кластеру, содержащему срединную точку, как к одному слову. При этом алгоритмы, реализующие эти стандарты, имеют тенденцию работать противоположным образом: UDPipe, Stanza и fastHan //udc, основанные на стандарте CTB, последовательно разделяют слова по срединной точке, а основанный на CNS алгоритм Ckiptagger этого никогда не делает.

Следующий тип неоднозначных вхождений связан с т.н. родовыми словами. Так называются односложные (в большинстве случаев) морфемы, которые нередко употребляются в постпозиции имени собственного с целью его более точного определения. Они занимают неоднозначную позицию с точки зрения как своих синтаксических, так и семантических характеристик. Эту нестабильность отражают и стандарты CWS: одни из них (государственный стандарт КНР, предшественник PKU) считают любые сочетания с родовым словом единым токеном, в то время как другие (например, PKU и CNS) дифференцируют разные родовые слова по разным признакам.

Алгоритмы в большинстве случаев не подчиняются стандартам, однако в их действиях можно найти две тенденции. Первая: чем длиннее родовое слово, тем вероятнее оно будет отделено от имени собственного. Так, двусложный элемент *森林 sēnlín*, «лес» отделяется от предшествующего имени значительно чаще, чем односложный *村 cūn*, «деревня». Вторая тенденция заключается в более частом выделении званий людей (которые занимают позицию родовых слов), нежели в выделении родовых слов, характеризующих топонимы. Так, *将军 jiāngjūn*, «генерал» всегда отделяется от предшествующей фамилии, в то время как *城 chéng* «город» обычно образует единое целое с предшествующим названием.

Необходимо отметить, что две вышеописанные тенденции более всего совместимы с синтаксическим стандартом Penn Chinese Treebank, который предписывает отделять титулы людей и многосложные постпозитивные элементы компаундов, и при этом не отделять односложные постпозитивные морфемы. Это любопытно, так как в данном случае за основу в выделении слов берутся не морфосинтаксические критерии, а фонотактические соображения, ведь тенденция выделять двусложные родовые слова и не выделять односложные вновь напоминает нам о «метрических» тенденциях в китайском языке.

1. **Исследования в области морфосинтаксической аннотации**

Как уже было отмечено, в отличие от других языков при разметке китайских текстов PoS-тэггинг не самая тривиальная задача. Это связано с тем, что существует несколько мнений о том, что такое части речи в китайском языке и как они выделяются (дискуссия по этому вопросу проиллюстрирована, например, в сборнике [Софронов 1989: 37–126]). Исходя из этого, существует большое количество стандартов морфопарсинга для китайского языка. Более того, в ряде систем PoS-тэггинга различают не только морфологические классы слов, но и семантические. Так, имена людей и названия географических объектов могут размечаться различными PoS-тэгами.

Мы проверили алгоритмы китайского PoS-тэггинга на корректность разметки заимствованных слов, к которым относятся имена людей и названия географических объектов. В качестве данных были взяты те же предложения, что и для оценки качества алгоритмов CWS (Раздел 2). Все предложения были разделены на две группы: содержащие топонимы и содержащие антропонимы (так как в ряде стандартов этим группам присваиваются различные PoS-тэги). Мы рассмотрели ряд алгоритмов, основанных на различных стандартах PoS-тэггинга китайских текстов; все они представлены в Таблице 2.

*Таблица 2.* **Перечисление использованных стандартов и алгоритмов PoS-тэггинга**

|  |  |
| --- | --- |
| Стандарт PoS-тэггинга | Инструмент |
| Chinese national standard (CNS) | Ckiptagger |
| Peking university (PKU) | PKUSeg |
| PKU (модифицированный) | NLPIR, LTP |
| Penn Chinese Treebank (CTB) | fastHan |
| Universal Dependencies + Penn Chinese Treebank (UPOS) | stanza, spacy |

Как и в случае со словоделением, о количественном исследовании подробно рассказано в [Семенов и др. 2021, в печати]. Здесь стоит отметить, что, как в случае со словоделением, явным лидером является fastHan.

Обратимся к качественному анализу ошибок PoS-тэггеров. Самые частотные из них – это присвоение заимствованным существительным PoS-тэги глаголов, прилагательных и наречий. При этом есть случаи, когда можно попробовать интерпретировать подобные ошибки: например, Ckiptagger разметил транслитерацию топонима «Китеж» (*基捷日 jījiérì*) как «слово со значением времени» (отдельная морфологическая категория в китайском, тэг Nb), вероятно, потому, что в составе этого токена присутствует *日 rì* («день, солнце»), составляющая частотная для временных слов.

Наконец, частотна ошибка, когда инструмент путает антропонимы и топонимы. При транслитерации фамилии «Гарин» (*加林 jiālín*) Ckiptagger отнес имя человека к топонимам (тэг Nc), вероятнее всего, из-за родового слова *林 lín -* *«лес»*. NLPIR, наоборот, при транслитерации города «Псков» (*普斯科夫 pǔsīkēfū*) приписывает тэг антропонима nrf. Здесь можно предположить, что алгоритм введен в заблуждение морфемой *夫 fū, «муж»,* которое нередко используется с личными именами или для образования профессий.

**4. Заключение**

В результате нашего исследования мы получили следующие выводы: с точки зрения логичности в области словоделения и морфосинтаксической аннотации наиболее подходящими для нашего корпуса является стандарт Penn Chinese Treebank, реализованный в CWS- и PoS-алгоритмах fastHan.

Учитывая, что финальной целью исследований является улучшение алгоритма лингвистической аннотации ruzhcorp, мы планируем встроить вышеописанные алгоритмы в систему препроцессинга китайских текстов. Тем не менее, нам предстоит доработать ряд алгоритмов (в частности, в области срединных точек), а также провести фундаментальные исследования в области родовых слов и их морфосинтаксической и фонотактической трактовки.

Кроме того, наша рабочая группа продолжает практические исследования и эксперименты в области разметки китайских текстов, а именно: дообучением наилучших моделей fastHan на наборах предложений из ruzhcorp; созданием альтернативного модуля CWS, включающего нейросетевой модуль с определением смены кодов; экспериментами в области параллельной морфологической разметки (привлекая PoS-тэги соответствующих русских предложений).

**Литература**

1. Семенов, К. И., Короткова, Ю. О., Коновалова, А. С., & Вольф, Е. А. (2021) [в печати], Автоматическая лингвистическая разметка китайских текстов, содержащих заимствования: Словоделение, транскрипция, PoS-тэггинг. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies, М.

2. Софронов, М. В. (Ред.). (1989), Новое в зарубежной лингвистике. Выпуск 22. Языкознание в Китае. М.: Прогресс.

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

4. Duanmu, S. (1999), Stress and the Development of Disyllabic Words in Chinese. Diachronica, 16(1), 1–35. https://doi.org/10.1075/dia.16.1.03dua

5. Emerson, T. (2005), The Second International Chinese Word Segmentation Bakeoff. http://sighan.cs.uchicago.edu/bakeoff2005/

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

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

8. Li, P.-H., & Ma, W.-Y. (2019), CkipTagger. https://github.com/ckiplab/ckiptagger

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

10. Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018), Universal Dependency Parsing from Scratch. https://doi.org/10.18653/v1/K18-2016

11. Straka, M. (2018), UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. 197–207. https://doi.org/10.18653/v1/K18-2020

12. Wong, K.-F., & Xu, R. (2010), Introduction to Chinese natural language processing. Morgan & Claypool Publ.

13.张[Zhang], 华平[Huaping], & 商[Shang], 建云[Jianyun]. (2019), NLPIR-Parser：大数据语义智能分析平台[NLPIR-Parser: An intelligent semantic analysis toolkit for big data]. 语料库语言学[Corpus Linguistics], 6(1), 87–104.

**References**

1. Semenov K.I., Korotkova Y.O., Konovalova A.S., Volf E.A. (2021) [in press], Automatic Annotation of the Chinese Texts that Contain Loanwords: Word Segmentation, Transcription, PoS-tagging [Avtomaticheskaya lingvisticheskaya razmetka kitajskix tekstov, soderzhashhix zaimstvovaniya: Slovodelenie, transkripcziya, PoS-te`gging]. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies, M.

2. Sofronov M.V. (ed.) (1989), New Issues in Foreign Linguistics. Volume XXII: Chinese LanguageScience [Novoe v zarubezhnoj lingvistike. Vy'pusk 22. Yazy'koznanie v Kitae]. Moscow: Progress. 472 p.

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

4. Duanmu, S. (1999), Stress and the Development of Disyllabic Words in Chinese. Diachronica, 16(1), 1–35. https://doi.org/10.1075/dia.16.1.03dua

5. Emerson, T. (2005), The Second International Chinese Word Segmentation Bakeoff. http://sighan.cs.uchicago.edu/bakeoff2005/

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

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

8. Li, P.-H., & Ma, W.-Y. (2019), CkipTagger. https://github.com/ckiplab/ckiptagger

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

10. Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018), Universal Dependency Parsing from Scratch. https://doi.org/10.18653/v1/K18-2016

11. Straka, M. (2018), UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. 197–207. https://doi.org/10.18653/v1/K18-2020

12. Wong, K.-F., & Xu, R. (2010), Introduction to Chinese natural language processing. Morgan & Claypool Publ.

13. Zhang H. Shang J. (2019), NLPIR-Parser: An intelligent semantic analysis toolkit for big data // Corpus Linguistics. Vol. 6(1). pp.87–104.

**Елена Александровна Вольф**

НИУ «Высшая школа экономики» (Россия)

**Elena A. Volf**

Higher School of Economics (Russia)

***E-mail: eavolf@edu.hse.ru***

Кирилл Игоревич Семенов

Kirill I. Semenov

Insitute for Information Transmission Problems of RAS (Russia)

Институт проблем передачи информации РАН (Россия)

***E-mail: kir.semenow@yandex.ru***

Юлия Олеговна Короткова

НИУ «Высшая школа экономики» (Россия)

Yulia O. Korotkova

Higher School of Economics (Russia)

***E-mail: yuokorotkova@edu.hse.ru***

1. Проект выполнен в рамках Конкурса проектных групп для обучающихся НИУ ВШЭ ФГН (название проекта – «Лингвоспецифическая разметка китайских текстов в Русско-китайском параллельном корпусе НКРЯ»). [↑](#footnote-ref-1)