Автоматическая лингвистическая разметка китайских текстов, содержащих заимствования: словоделение, транскрипция, PoS-тэггинг

Александра Сергеевна Коновалова
НИУ «Высшая школа экономики», Москва, Россия
askonovalova@edu.hse.ru

Елена Александровна Вольф
НИУ «Высшая школа экономики», Москва, Россия
eavolf@edu.hse.ru

Кирилл Игоревич Семенов
Институт проблем передачи информации РАН, Москва, Россия
kir.semenow@yandex.ru

Юлия Олеговна Короткова
НИУ «Высшая школа экономики», Москва, Россия
yuokorotkova@edu.hse.ru

Аннотация

В статье описываются проблемы лингвистической аннотации китайских текстов в Русско-китайском параллельном корпусе НКРЯ (далее – Корпус) и пути их решения. Особенное внимание уделяется проблеме обработки заимствований из русского языка. Представлено описание экспериментов в трех аспектах лингвистической разметки: словоделения, фонетической аннотации (G2P) и морфологической аннотации (PoS-тэггинг). Также описано создание датасетов, разработанных на основе данных Корпуса, которые могут быть использованы в дальнейших исследованиях нестандартных текстов на китайском языке. Полученные результаты исследования планируется применить для переразметки и дальнейшей обработки текстов в Корпусе.

Ключевые слова: автоматическая сегментация; автоматическая транскрипция; морфологическая аннотация; проблема слов вне словаря; автоматическое определение смены кодов

DOI: 10.28995/2075-7182-2021-20-1081-1094

Automatic Annotation of the Chinese Texts that Contain Loanwords: Word Segmentation, Transcription, PoS-tagging

Aleksandra S. Konovalova
Higher School of Economics, Moscow, Russia
askonovalova@edu.hse.ru

Elena A. Volf
Higher School of Economics, Moscow, Russia
eavolf@edu.hse.ru

Kirill I. Semenov
Insitute for Information Transmission Problems of RAS, Moscow, Russia
kir.semenow@yandex.ru

Yulia O. Korotkova
Higher School of Economics, Moscow, Russia
yuokorotkova@edu.hse.ru

Abstract

The article tackles the problems of linguistic annotation of the Chinese texts presented in the Russian Chinese Parallel Corpus of RNC (hereafter – our corpus), and the ways to solve them. Particular attention is paid to the Russian loanwords in the texts, as they, firstly, are abundant in our corpus, secondly, are of interest as the cases of both out-of-vocabulary and code-switching problems. We describe our experiments in three fields, namely, word segmentation, grapheme-to-phoneme conversion, and PoS-tagging. In order to test the algorithms on our specific data, we created our own datasets based on the corpus, which can be precious for the following research in the field of processing the non-standard Chinese texts. As the main aim of the research is to improve the quality of the annota-

tion in our corpus, we plan to implement the results of our work in the preprocessing pipeline of the new texts in the corpus.

Keywords: Chinese word segmentation (CWS); grapheme-to-phoneme conversion (G2P); PoS-tagging; out-of-vocabulary problem (OOV); code-switching detection

1 Введение

Лингвистическая разметка является одной из ключевых составляющих современного лингвистического корпуса, так как она позволяет вести поиск и получать статистику выдачи не только по словоформам и символьным строкам, но и по лингвистическим характеристикам – от леммы до семантических ролей. Кроме того, в случае использования корпуса как датасета для алгоритмов автоматической обработки естественного языка (далее – NLP), единицы лингвистической разметки могут повысить качество языковых моделей. Наиболее распространенным уровнем лингвистической разметки является морфологический парсинг, включающий присвоение каждой словоформе ее леммы и морфосинтаксических характеристик (PoS-тэгов).

В области лингвистической аннотации особняком стоит задача лингвистической разметки китайских текстов. Это связано с рядом причин, среди которых ключевыми являются следующее: отсутствие традиционного и общепринятого принципа разделения текста на слова, иероглифическая письменность (и отсутствие взаимно однозначного соответствия между иероглифами и транскрипцией), а также морфосинтаксический строй, который принципиально отличается от европейских языков. Все вышеперечисленные проблемы усугубляются в случае, если в тексте присутствуют заимствованные слова, так как их устройство обычно противоречит стандартным свойствам китайских слов.

С набором этих проблем столкнулась команда разработчиков Русско-китайского параллельного корпуса НКРЯ (далее – ruzhcorp, [6]). Ruzhcorp – корпус русских и китайских текстов с взаимными переводами, разрабатываемый с 2016 года. На сегодняшний день это единственный российский параллельный корпус, одновременно обладающий следующими свойствами: пара языков «русский – путунхуа», доступность онлайн, наличие лингвистической разметки и удобный для широкого круга пользователей интерфейс. Актуальность развития такого корпуса очевидна для гуманитарных исследований, разработки алгоритмов компьютерной лингвистики и методики преподавания китайского языка.

На текущий момент ruzhcorp обладает следующими уровнями разметки китайских текстов: разделением на слова, фонетической нотацией иероглифов в системе пиньинь и толкованием каждого слова на основе англо-китайского словаря. К сожалению, вышеописанные элементы разметки китайских текстов в текущей версии имеют проблемы. Так, алгоритм словоделения (далее CWS – от Chinese word segmentation) представляет собой вариант жадного поиска Backward Maximum Matching, т.е. итеративный поиск самой длинной подстроки с конца предложения, которую можно встретить в загруженном словаре. Это вызывает проблемы с выделением заимствований имен собственных из русского языка и иных слов, не входящих в словарь (проблема out of vocabulary – OOV). Алгоритм присвоения пиньиня построен на том, что каждому иероглифу приписываются все возможные прочтения, что представляется избыточным. Наконец, в корпусе отсутствует уровень морфологического парсинга (здесь и далее термины “морфологический парсинг”, “морфологическая разметка” и “PoS-тэггинг” используются как синонимы).

В рамках настоящей статьи мы опишем ряд исследований и экспериментов по улучшению текущего положения дел. В первой части статьи рассматривается экспериментальная работа по дизайну автоматического словоделения в корпусе. Во второй части работы предложено улучшение алгоритма автоматической пиньиневой аннотации текстов в ruzhcorp на основе сравнительного анализа наиболее популярных архитектур в области фонетической аннотации (далее – G2P, от grapheme-to-phoneme). В третьей части рассмотрено сравнительное лингвистическое исследование стандартов PoS-тэггинга для китайских текстов и проведены эксперименты по оценке качества морфологической разметки текстов, содержащих заимствования. В заключении мы представим перспективы применения наших наработок на практике и обозначим горизонты дальнейших исследований.

2 Исследования в области словоделения

Первичной задачей для большинства аспектов NLP является токенизация текста. В случае с китайскими текстами, аналогом токенизации является словоделение.

Задача CWS осложняется рядом факторов, к важнейшим из которых можно отнести следующие. Во-первых, китайская филологическая и лингвистическая традиция до встречи с западным языкознанием сравнительно редко оперировала термином “слово”, предпочитая в качестве единицы текста рассматривать иероглиф, который соответствует на фонетическом уровне слогу, а на морфологическом – морфеме. Во-вторых, в силу высокого уровня аналитизма, в китайском языке достаточно мало морфологических маркеров границы слова. В силу этих двух причин китайский слог (выражающийся на письме одним иероглифом) может представлять в тексте как отдельное слово, так и часть более длинной (обычно – двусложной) лексемы. Наконец, большая часть китайского лексикона обладает свойством конверсии, поэтому различные морфосинтаксические критерии также слабо применимы для задачи словоделения и PoS-тэггинга.

Проблема CWS усложняется при наличии фонетических заимствований, записанных иероглифами. Дело в том, что в китайском языке отсутствует кодифицированный набор иероглифов, предназначенных для транслитерации заимствований (в отличие от японского языка, где подобную роль выполняет катакана). Из-за этого 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. В примере (1) представлен такой набор иероглифов. Более того, в китайской орфографии практически не существует маркеров, указывающих на начало или конец последовательности символов, которые необходимо читать “фонетически”, игнорируя семантику иероглифов. Хотя носитель китайского языка в большинстве случаев может безошибочно определить границы “фонетического” прочтения иероглифов, для автоматического алгоритма CWS это не всегда тривиальная задача.

(1) “Семантическое” (слева) и “фонетическое” (справа) прочтение китайских иероглифов.

马	里	马	里
mǎ	lǐ	mǎ	lǐ
лошадь	LOC	“Мали [страна в Африке]” или	
“в/внутри лошади”		“Мары [город в Туркменистане]”	

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, классическая китайская литература), 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. Мы могли бы подойти к этой проблеме путем создания двух независимых модулей для каждого блока. В пользу такого решения говорит тот факт, что в китайском NLP традиционно выделяется задача CUWE (Chinese Unknown Word Extraction), целью которой является в частности распознавание заимствований в китайском тексте – см. [21], [18]. Кроме того, наши предыдущие исследования показывают, что даже простые статистические метрики показывают существенное отличие иероглифических заимствований от обычных слов – например, ранг иероглифических биграмм внутри заимствований в большинстве случаев значительно выше ранга окружающих слов, в то время как иероглифические триграммы, частотные внутри заимствований, практически не встречаются за их пределами [7, с. 58].

Однако решение встроить отдельный модуль CUWE, на наш взгляд, имеет и ряд недостатков. Во-первых, большинство исследований в области CUWE было направлено скорее не на обработку непосредственного текста, а на более фундаментальные задачи вроде пополнения словаря и т.д. Таким образом, отдельных усилий требовало бы совмещение результатов алгоритма CUWE и дальнейшего CWS на всем остальном предложении. Кроме того, несмотря на до сих пор встречающиеся исследования в этой области (напр., [13]), в последние годы задача CUWE теряет свою популярность. На наш взгляд, это может быть связано с развитием нейросетевых алгоритмов словоделения и других задач китайского NLP, которые, с одной стороны, нередко

решают задачу определения несловарных слов лучше отдельных модулей, с другой – на высоком уровне обрабатывают непосредственно китайские слова.
Руководствуясь этими рассуждениями, мы приняли решение не разрабатывать на текущем этапе исследования отдельный модуль CUWE, а исследовать качество ведущих алгоритмов CWS для задачи выделения заимствований, и затем, при необходимости, доработать их при помощи дообучения или простых правиловых методов. Это, на наш взгляд, даст нам возможность добиться хорошего качества выделения заимствований при сохранении качества выделения “стандартных” китайских слов.
Чтобы сравнить алгоритмы CWS и с теоретической, и с практической стороны, мы рассмотрели как различные стандарты CWS, так и их наиболее популярные реализации. В настоящий момент существует несколько популярных стандартов CWS, при этом ни один из них не является общепринятым абсолютным большинством исследователей и специалистов в NLP. Стандарты словоделения могут принципиально различаться, так как они берут за основу различные аспекты и уровни языковой структуры – семантику, морфосинтаксис или лексику. В нашем исследовании мы использовали 5 основных стандартов и их вариаций – все они представлены в Таблице №1.
Для большинства из этих стандартов существует более одного алгоритма, реализующего этот стандарт. В нашем исследовании было использовано 17 алгоритмов и их вариантов (когда одна архитектура алгоритма CWS была обучена на разных датасетах), соотнесенных с тем или иным стандартом (см. Таблицу №1).

Стандарт	**Краткое описание**	**Алгоритмы, соответствующие стандарту**
Стандарт Пекинского университета (PKU) [24]	Самый старый стандарт словоделения (разрабатывается с 1990-х годов). Критерии определения слова связаны с лексической семантикой и лексической сочетаемостью.	pkuseg, fastHan // {pku, sxu}, LTP
Государственный стандарт от Academia Sinica (CNS) [27]	Наиболее старый стандарт, разработанный на Тайване. Критерии определения слова – морфосинтаксические.	ckiptagger, fastHan // {as, cnc}
Стандарт Microsoft Research Asia [10]	Критерии определения слова – морфосинтаксические; при этом стандарт не самодостаточен и ориентирован на совместимость с другими.	fastHan // msr
Стандарт Penn Chinese Treebank (CTB) [22]; вариация – Universal Dependencies	Критерии определения слова – синтаксические.	stanza, spacy, fastHan // {ctb, udc, wtb, zx}, Spacy-UDPipe
Словарные стандарты	Главный критерий — наличие слова в загруженном в алгоритм словаре.	NLPIR, jiagu

Таблица №1: Сравнительное описание стандартов CWS

Обратимся более подробно к техническим параметрам каждого из использованных алгоритмов[1]:

1 Необходимо отметить, что в нашей выборке отсутствует пакет jieba (URL: https://github.com/fxsjy/jieba). Несмотря на его популярность, у алгоритма есть ряд недостатков. Во-первых, в этом алгоритме не специфицирован стандарт словоделения, из-за чего нельзя быть уверенным в последовательности самого алгиртма.Во-вторых, качество рассмотренных нами алгоритмов в большинстве случаев значительно превышает качество jieba на стандартных китайских текстах. Наконец, наши предварительные исследования показали, что качество выделения русских заимствований у jieba значительно меньше, чем у других алгоритмов (обычно - не выше 50%) - см. [8, с. 124].

1. Ckiptagger [14] — нейронная сеть, включающая слои bidirectional LSTM (двунаправленная долгая краткосрочная память) и multi-head attention (множественное внимание);
2. pkuseg [1] — алгоритм adaptive online gradient descent based on feature frequency information (ADF, адаптивный градиентный спуск, основанный на частоте признаков), описанный в статье [20];
3. fastHan [4] — трансформер BERT, с целью уменьшения количества занятой памяти используется 4 или 8 (в зависимости от выбора модели: base или large) из 12 слоёв с механизмом внимания[2];
4. NLPIR [25] — метод, основанный на словаре с последующим применением графового метода поиска кратчайших путей (n-shortest path);
5. jiagu [16] – метод, основанный на словаре с последующим применением направленного ацикличного графа (DAG – Directed acyclic graph);
6. Stanza [17] — нейронная сеть с несколькими слоями bidirectional LSTM;
7. SpaCy [11] — нейронная сеть с несколькими слоями bidirectional LSTM;
8. LTP — [5] предобученная модель ELECTRA [9];
9. UDPipe [19] – нейронная сеть на основе bidirectional GRU (двунаправленный управляемый рекуррентный блок).

Цель нашего исследования – определить наилучший алгоритм словоделения для наших данных. Для этого мы провели эксперимент, описанный в подробностях ниже.

Чтобы определить оптимальный алгоритм CWS для нашего корпуса, необходимо уделить внимание следующим параметрам: во-первых, теоретическому (какие алгоритмы словоделения используют наиболее последовательные и прозрачные стандарты), во-вторых, ряду практических. К последим относится качество работы алгоритмов с учетом двух особенностей наших данных: наличия большого количества несловарных заимствований из русского языка, не маркированных в потоке иероглифов; и наличия текстов разных доменов (художественного и новостного). Первая особенность является частью проблемы out-of-vocabulary (OOV), вторая – частью проблемы out-of-domain (OOD).

Для того, чтобы учесть особенности нашего корпуса, для проверки качества алгоритмов мы создали два датасета: первый состоял из 408 предложений из текстов художественной литературы, содержащих русские имена собственные, второй – из 87 предложений из средств массовой информации с таким же свойством. К каждому предложению были приписаны индексы границ заимствований.

Для сравнения алгоритмов CWS как с теоретической, так и с практической точек зрения были взяты 17 алгоритмов и вариантов, описанных в предыдущем разделе. Каждый алгоритм принимал на вход предложение и делил его; после этого мы сверяли индексы границ слов на предмет того, совпадают ли реальные индексы границ заимствований с теми, что предложены алгоритмом.

Для оценки качества использовались следующие метрики:

- Полнота (recall) — отношение количества верно выделенных заимствований к реальному количеству заимствований.
- F-мера (F-score) — гармоническое среднее точности (precision – отношения количества верно выделенных границ заимствований к общему количеству разделенных токенов на месте заимствований) и полноты.
- Собственная метрика (our):

$$\frac{1}{n}\sum_{i=0}^{n}\left(1-\frac{r_i}{length_i}\right)\times\frac{1}{tokens_i}$$

Где n — общее количество заимствований, r_i — количество повторяющихся символов в i-том заимствовании, $length_i$ — длина i-того заимствования, $tokens_i$ — количество реаль-

2 В приведенной выше таблице можно увидеть, что алгоритм fastHan предобучен на разных датасетах. Информация о датасетах того или иного стандарта CWS, на которых был обучен fastHan, приведена в каждой ячейке через двойной слэш (если датасетов больше одного, они перечислены в фигурных скобках).

но разделенных токенов в сегментации i-того заимствования. Данная метрика была введена для того, чтобы оценить степень, с которой выделенная алгоритмом подстрока отличается от правильной подстроки, являющейся заимствованием. В метрике учитывается как недостаточное выделение заимствования (когда в заимствование “включаются” граничащие с ним полнозначные иероглифы), так и чрезмерную сегментацию (когда одно заимствование делится на 2 и более слов).

Our	*F-score*	*Recall*	**Метрика**
0,762	0,504	0,564	**ckiptagger**
0,855	0,679	0,749	**pkuseg**
0,937	0,865	0,888	**fastHan//as**
0,952	0,906	0,920	**fastHan//cnc**
0,955	0,908	0,916	**fastHan//ctb**
0,940	0,896	0,905	**fastHan//msr**
0,945	0,884	0,901	**fastHan//pku**
0,948	0,898	0,906	**fastHan//sxu**
0,834	0,599	0,753	**fastHan//udc**
0,950	0,901	0,910	**fastHan//wtb**
0,944	0,883	0,901	**fastHan//zx**
0,845	0,607	0,746	**NLPIR**
0,413	0,075	0,179	**jiagu**
0,593	0,345	0,433	**stanza**
0,827	0,609	0,696	**spacy**
0,926	0,843	0,866	**LTP**
0,807	0,574	0,672	**UDPipe**

Таблица №2: Сравнение результатов словоделения на данных художественной литературы

Our	***F-score***	***Recall***	**Метрика**
0,626	0,395	0,433	**ckiptagger**
0,932	0,833	0,858	**pkuseg**
0,899	0,809	0,836	**fastHan//as**
0,8616	0,8550	0,8582	**fastHan//cnc**
0,927	0,878	0,888	**fastHan//ctb**
0,743	0,812	0,821	**fastHan//msr**
0,928	0,878	0,888	**fastHan//pku**
0,928	0,882	0,888	**fastHan//sxu**
0,758	0,516	0,664	**fastHan//udc**
0,932	0,889	0,896	**fastHan//wtb**
0,927	0,878	0,888	**fastHan//zx**
0,844	0,600	0,739	**NLPIR**
0,548	0,178	0,388	**jiagu**
0,551	0,318	0,410	**stanza**
0,830	0,603	0,687	**spacy**
0,932	0,879	0,896	**LTP**
0,625	0,301	0,440	UDPipe

Таблица №3: Сравнение результатов словоделения на данных новостных текстов

Сравнение метрик качества разных алгоритмов приведено в таблицах №2 (на данных художественной литературы) и №3 (на данных новостных текстов). Запись “fastHan // <dataset>” озна-

чает, что использовался вариант алгоритма fastHan, обученный разработчиками на указанном после "//" датасете. Каждая модель запускалась без дообучения на наших данных.

Заметим, что наилучшие результаты на обоих датасетах показывают версии алгоритма fastHan, основанные на синтаксическом стандарте Penn Chinese Treebank: fastHan // wtb и fastHan // ctb. Другие версии fastHan также демонстрируют высокие значения метрик. К числу высокоэффективных алгоритмов также можно отнести нейросетевые методы LTP и PKUSeg. Эти наблюдения подтверждают высокую эффективность архитектуры BERT и ее альтернативы ELECTRA, которые сейчас являются одними из доминирующих в ряде задач NLP. Значительно меньшую эффективность показали другие нейросетевые модели, построенные на рекуррентных и LSTM-сетях. Это соответствует распространенному мнению (выраженному, например, в хрестоматичной работе Attention is All You Need 2017 года) о превосходстве моделей с механизмом внимания над LSTM — за счет, в частности, увеличения скорости обучения и глубины сети.

Примечательно, что модели Stanza и UDPipe, ориентированные на выполнение задач NLP в широком спектре языков, показывают лишь среднюю эффективность на наших данных. Нейронные алгоритмы, предназначенные для задач китайского NLP, справляются с нашими данными лучше.

Также важно отметить, что алгоритмы NLPIR и jiagu, в основе которых лежат графовые методы, показали свою сравнительно низкую эффективность. Учитывая, что графовые методы были одними из наиболее распространенных в задаче CWS в предыдущие годы (таковым был, например, наиболее популярный jieba), мы можем сказать, что нейросетевой подход к поиску заимствований не в ущерб словоделению "стандартных" слов оказывается в целом более продуктивным, вне зависимости от конкретных реализаций.

В заключение необходимо отметить, что, помимо зависимости качества выделения заимствований от архитектуры алгоритма CWS, можно наблюдать корреляцию между стандартом CWS и качеством выделения заимствований. Так, алгоритмы, основанные на стандартах CTB (и, в меньшей степени, на иных синтаксических стандартах) и PKU, показывают большую эффективность, чем алгоритмы на основе CNS. Впрочем, это наблюдение стоит признать именно корреляцией, а не каузацией, так как fastHan, обученный на датасетах в формате CNS, справляется с поставленной задачей хорошо.

3 Исследования в области автоматической фонетической аннотации

Для начала отметим, что задача автоматической фонетической аннотации — вторичная задача по отношению к словоделению. Так как одному иероглифу в зависимости от контекста могут соответствовать разные варианты фонетической транскрипции, модели для G2P применяются поверх CWS.

На данный момент в ruzhcorp работает только поверхностная фонетическая разметка: к каждому иероглифу предложены все возможные варианты транскрипции в системе пиньинь. Так как в китайском языке большое количество иероглифов имеют больше одного прочтения, основную проблему составляет разрешение фонетической омонимии.

Существует несколько подходов к транскрибированию иероглифов. Алгоритмы, основанные на правилах, предполагают, что на первом этапе из словаря извлекаются все возможные варианты прочтения одного иероглифа, а затем выбирается нужный, исходя из некоторых условий. Трудность такого подхода состоит в том, что приходится применить достаточно большой набор правил, чтобы охватить весь контекст.

Другие подходы основаны на данных и включают в себя статистические методы и машинное обучение. Такие модели работают со снятием омонимии как с задачей классификации, где классы – варианты транскрипции, а объект – векторное представление иероглифа в контексте. Чтобы учитывать синтаксическую структуру предложения, пользуются популярностью модели Bi-LSTM [3] и RNN [2].

Мы рассмотрели несколько готовых инструментов для пиньиневой анноации, которые доступны в виде библиотек для Python: G2pC [2], G2pM [3], xpinyin [15] и pypinyin [12].

G2pC — контекстно-зависимая модель для перевода иероглифов в транскрипцию. G2pM представляет собой нейросеть, обученную на специально составленном тренировочном датасете, содержащим много полифонов. Xpinyin предлагает модель стохастического решения, основанную

на частотности пиньиневых транскрипций. Pypinyin учитывает информацию об n-граммах и использует собственный словарь коллокаций.
Среди рассмотренных библиотек G2pM и G2pC используют словоделитель, поверх которого применяются алгоритмы для фонетической аннотации. G2PM использует собственную встроенную модель для CWS, тогда как в G2pC можно встроить внешний словоделитель. G2pC работает с сырым предложением. На первом этапе предложение делится на слова с помощью стороннего парсера (по умолчанию — PKUSeg), дальше для полученных слов из словаря извлекаются все возможные транскрипции. Затем модель CRF решает задачу классификации. На последнем этапе применяются правила изменения тона (например, тоновая пара 3-3 в китайском переходит в пару 2-3).
Мы предположили, что качество G2P зависит от стандарта и качества CWS. Поэтому модель G2pC мы настраивали, подключая к ней различные слововоделители: PKUSeg, fastHan и UDPipe.
Для оценки качества работы моделей на основе разных словоделителей мы составили пилотный набор данных из 20 размеченных вручную предложений из ruzhcorp, где каждому иероглифу была сопоставлена пиньиневая транскрипция с тоном. В таблице №4 представлены результаты сравнения на датасете. В качестве метрики мы использовали процент верно определенных пиньиней для каждой отдельной пары пиньинь-иероглиф.

Модель	**Доля верных пар “иероглиф-транскрипция”**
G2pC (pkuseg)	0,903
G2pC (FastHan)	0,899
G2pC (UDPipe)	0,880
Xpinyin	0,861
Pypinyin	0,831
G2pM	0,824

Таблица №4: Результаты алгоритмов фонетической аннотации

Лучший результат показала модель G2pC на слоделении PKUSeg. Мы предполагаем, что причина заключается в следующем: словоделитель PKUSeg натренирован на нескольких датасетах из разных доменов (медицина, искусство и тд.), что позволяет ему работать лучше на новых данных, чем другим моделям, которые обучаются преимущественно на новостных текстах.

4 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 [26, с. 37–126]). Это также обусловлено тем, что PoS-тэггинг является вторичной задачей по отношению к задаче CWS, которая сама по себе неоднозначна. Исходя из этого, существует большое количество стандартов морфопарсинга для китайского языка, основная часть которых будет упомянута ниже. Более того, в ряде систем PoS-тэггинга учитываются не только морфологические различия, но и семантические. Так, имена людей и названия географических объектов могут размечаться различными PoS-тэгами.

Чтобы оценить качество морфопарсинга, нужно либо принять одну из точек зрения и придерживаться ее, оценивая качество размеченных PoS-тэгов, либо смотреть, на какой стандарт опирается каждый из алгоритмов, и оценивать его, исходя из того, какого принципа придерживаются сами создатели инструмента. Но наша задача состояла в другом: так как наш корпус обладает спецификой – содержит заимствования – мы проверяли алгоритмы на то, как они будут размечать именно заимствованные слова, к которым относятся имена людей и названия географических объектов.

В качестве данных были взяты те же предложения, что и для оценки качества алгоритмов по словоделению (Раздел 1.). Все данные были разделены на две группы: те, которые содержат имена людей, и те, которые содержат названия географических объектов. Так как в одном предложении могли оказаться и имена людей, и географические названия, такие предложения попали в обе группы. Данные были разделены на две группы для того, чтобы точнее оценить качество разметки, потому что для личных имен и географических названий присваиваются различные PoS-тэги.

Существует несколько инструментов, которые позволяют размечать PoS-тэги в китайских текстах автоматически. Каждый из инструментов основан на своем стандарте и имеет свой набор PoS-тэгов. Мы рассматривали некоторые из них, а именно: Ckiptagger [14], PKUSeg [1], fastHan (тэги совпадают с системой Penn Chinese Treebank – [23]), NLPIR [25], stanza [17], spacy [11], LTP [5]. В Таблице №5 приведены тэги для заимствований, соответствующие каждому стандарту и инструменту. Поскольку китайская система PoS-тэггинга содержит не только морфологическую информацию, но и семантическую дифференциацию, мы выделили по 3 класса PoS-тэгов для каждого инструмента: для имен людей; для названий географических объектов; в третий класс попали PoS-тэги, которые в целом подходят для заимствований, но недостаточно конкретны (например, поскольку все имена собственные относятся к классу существительных, к третьей группе были отнесены PoS-тэги, соответствующие классу имен существительных). При оценке качества алгоритма мы считали, что инструмент работает удовлетворительно, если он размечает слово одним из тэгов, представленных в таблице, но предпочтительно, чтобы он корректно различал и семантическую дифференциацию (в случае если в данном стандарте она есть).

Инструмент	**Стандарт PoS-тэггинга**	**Имена людей**	**Названия географических объектов**	**Более общие и смежные классы**
ckiptagger	Chinese national standard (CNS)	Nb	Nc	Na
pkuseg	Peking university (PKU)	nr	ns	n, nz
fastHan	Penn Chinese Treebank (CTB)	NR	NR	NN
NLPIR	PKU (модифицированный)	nrf	nsf	n, nr, ns, nt, nz
stanza	Universal Dependencies + Penn Chinese Treebank (UPOS)	PROPN	PROPN	NOUN
spacy	UPOS	PROPN	PROPN	NOUN
LTP	PKU (модифицированный)	nh	ns	n, ni, nz

Таблица №5: Сравнение китайских PoS-тэгов, которые могут быть отнесены к заимствованиям

При оценке мы также разделили корректность разметки PoS-тэгов на три категории. К категории I относятся PoS-тэги из первых двух колонок Таблицы №5 (алгоритм верно поставил PoS-тэг с учетом семантического класса слова), к категории II – PoS-тэги из третьей колонки (алгоритм поставил PoS-тэг корректно с точки зрения морфосинтаксиса, однако не учел семантическую дифференциацию), к категории III – все остальные (алгоритм поставил PoS-тэг неверно с точки зрения морфосинтаксиса). В таблице №6 отображены результаты оценки каждого инструмента по трем категориям: для имен людей и географических названий отдельно.

Категория корректности	**I**		**II**		**III**	
Класс	**Имена людей**	**Названия географических объектов**	**Имена людей**	**Названия географических объектов**	**Имена людей**	**Названия географических объектов**
ckiptagger	0,72	0,84	0,11	0,08	0,17	0,08
pkuseg	0,76	0,71	0,09	0,11	0,15	0,18
fastHan	0,995	0,98	0,002	0,02	0,003	0
NLPIR	0,6	0,25	0,12	0,14	0,28	0,61
stanza	0,8	0,88	0,02	0,04	0,18	0,08
spacy	0,66	0,73	0,17	0,12	0,17	0,15
LTP	0,87	0,87	0,03	0,04	0,1	0,09

Таблица №6: Результаты PoS-тэггеров на нашем датасете

Из таблицы №6 следует, что явным лидером является fastHan. Он размечает верно практически в 100% случаев, при этом правильно выделяя токены с заимствованиями. Помимо fastHan, высокое качество морфопарсинга обеспечивают PKUSeg и LTP. Кроме того, примечательно, что с задачей PoS-тэггинга алгоритм stanza справляется значительно лучше, чем с CWS: возможно, это объясняется тем, что задача разделения текста на слова все же достаточно нестандартна для “обычных” европейских языков, в то время как задача PoS-тэггинга с точки зрения алгоритмической реализации достаточно универсальна.

5 Заключение

В результате нашего исследования мы получили следующие выводы. С точки зрения корректности и последовательности в области словоделения и морфосинтаксической аннотации наиболее подходящими для нашего корпуса являются стандарты PKU (вместе со стандартом морфопарсинга) и Penn Chinese Treebank. С точки зрения качества алгоритма, наиболее подходящим алгоритмом является fastHan и его варианты, обученные на датасетах, сделанных в стандарте Penn Chinese Treebank. Наконец, наилучшим алгоритмом фонетической аннотации является модель G2pC на базе pkuseg-словоделения.

Учитывая, что финальной целью исследований является улучшение алгоритма лингвистической аннотации ruzhcorp, мы планируем встроить вышеописанные доработанные алгоритмы в систему препроцессинга китайских текстов. Проблемным здесь остается конфликт между предпочтительными алгоритмами CWS (они должны быть скорее основаны на стандарте CTB) и алгоритмами G2P PoS-тэггинга (для них оптимально словоделение на стандарте PKU). Нам предстоит сделать окончательный выбор между стандартами словоделения.

Наша рабочая группа планирует продолжать работу и эксперименты в области аннотации китайских текстов. С одной стороны, алгоритм fastHan (который скорее всего будет взят в качестве CWS) имеет возможности для дообучения (fine-tuning). Сейчас мы проводим эксперимент, заключающийся в дообучении нескольких вариантов fastHan (работающих на основе разных стандартов словоделения) с целью сравнить степень, с которой улучшается один и тот же алгоритм, в зависимости от изначально загруженного датасета.

С другой стороны, мы рассматриваем возможность встраивания в разметку модуля с детекцией смены кодов (code-switching), под которой мы понимаем всякое присутствие транслитерирован-

ных слов. Распознавание смены кодов является задачей разметки последовательностей (sequence labeling). В качестве базовой модели мы используем сети с долгой краткосрочной памятью (LSTM), которые на вход получают скрытые представления предложений в модели словоделения, в нашем случае, fastHan, и на выходе мы получаем для каждого токена из разделения на слова предложения соответственно свой класс.

Благодарности

Работа проведена при поддержке Комиссии по поддержке образовательных инициатив ФГН НИУ ВШЭ в рамках Конкурса проектных групп для обучающихся НИУ ВШЭ ФГН (название проекта – «Лингвоспецифическая разметка китайских текстов в Русско-китайском параллельном корпусе НКРЯ»).

Литература

[1] Luo R. [и др.]. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation // arXiv:1906.11455 [cs]. 2019.

[2] Cai Z. [и др.]. Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features // arXiv:1907.01749 [cs, eess, stat]. 2019.

[3] Park K., Lee S. g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset // arXiv:2004.03136 [cs]. 2020.

[4] Geng Z. [и др.]. fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP // arXiv:2009.08633 [cs]. 2020.

[5] Che W. [и др.]. N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models // arXiv:2009.11616 [cs]. 2021.

[6] Дурнева С. П., Кузнецова Ю. Н., Семенов К. И. Русско-китайский параллельный корпус НКРЯ: проблемы и перспективы. X Международная научно-практическая конференция «Россия и Китай: история и перспективы сотрудничества». Благовещенск: БГПУ, 2020. С.633-640.

[7] Семенов К. И. Стратегии преобразования русских фонетических заимствований в китайском языке: фонетические и графические аспекты // Вестник РГГУ. Серия «Литературоведение. Языкознание. Культурология». 2020. № 7. С. 30–63.

[8] Семенов К. И. Заимствования из русского языка в стандартном китайском: проблемы фонетической и морфологической адаптации и методы автоматического распознавания в китайских текстах. Выпускная квалификационная работа студента 4 курса бакалавриата ОП "Фундаментальная и компьютерная лингвистика". М.: НИУ ВШЭ, 2020.

[9] Clark K. [и др.]. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators 2020.

[10] Gao J. [и др.]. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach // Computational Linguistics. 2005. № 4 (31). С. 531–574.

[11] Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing [Электронный ресурс]. URL: https://spacy.io/.

[12] Huang H. pypinyin [Электронный ресурс]. URL: https://github.com/mozillazg/python-pinyin.

[13] Jia Y. [и др.]. A Chinese unknown word recognition method for micro-blog short text based on improved FP-growth // Pattern Analysis and Applications. 2020. № 2 (23). С. 1011–1020.

[14] Li P.-H., Ma W.-Y. CkipTagger [Электронный ресурс]. URL: https://github.com/ckiplab/ckiptagger.

[15] Luo E. xpinyin [Электронный ресурс]. URL: https://github.com/lxneng/xpinyin.

[16] Ownthink Jiagu [Электронный ресурс]. URL: https://github.com/ownthink/Jiagu.

[17] Qi P. [и др.]. Universal Dependency Parsing from Scratch Brussels, Belgium: Association for Computational Linguistics, 2018.С. 160–170.

[18] Shen M., Kawahara D., Kurohashi S. Chinese Word Segmentation and Unknown Word Extraction by Mining Maximized Substring // Journal of Natural Language Processing. 2016. № 3 (23). С. 235–266.

[19] Straka M. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task Brussels, Belgium: Association for Computational Linguistics, 2018.С. 197–207.

[20] Sun X., Wang H., Li W. Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection Jeju Island, Korea: Association for Computational Linguistics, 2012. С. 253–262.

[21] Wang L.-J., Li W.-C., Chang C.-H. Recognizing unregistered names for Mandarin word identification Nantes, France: Association for Computational Linguistics, 1992. С. 1239.

[22] Xia F. The Segmentation Guidelines for the Penn Chinese Treebank (3.0) // IRCS Technical Reports Series. 2000. № 37. С. 33.

[23] Xia F. The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) // IRCS Technical Reports Series. 2000. № 38. С. 46.

[24] 俞[Yu] 士汶[Shiwen] 现代汉语语料库加工规范——词语切分与词性标注 [Standards for Processing Modern Chinese Corpus - Word segmentation and part-of-speech tagging] / 士汶[Shiwen] 俞[Yu], Beijing: 北京大学计算语言学研究所[Institute of Computational Linguistics, Peking University], 1999. 19 с.

[25] 张[Zhang] 华平[Huaping], 商[Shang] 建云[Jianyun] 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]. 2019. № 6(1). С. 87–104.

[26] М. В. Софронов (ред.). Новое в зарубежной лингвистике. Выпуск 22. Языкознание в Китае. М.: Прогресс, 1989. 472 с.

[27] 中文資訊處理分詞規範[Segmentation Principles for Chinese Language Processing] под ред. Academia Sinica, Taipei: 經濟部中央標準局 [Bureau of Standards, Metrology and Inspection of the Ministry of Economic Affairs of Taiwan], 1999. 28 с.

References

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

[2] Cai Z., Yang Y., Zhang C., Qin X., Li M. (2019), Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features // arXiv:1907.01749 [cs, eess, stat].

[3] Park K., Lee S. (2020), g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset // arXiv:2004.03136 [cs].

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

[5] 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].

[6] Durneva S.P., Kuznetsova Y.N., Semenov K.I. (2020), Russian-Chinese Parallel Corpus of RNC: Problems and Perspectives [Russko-kitajskij parallel'nyj korpus NKRYA: problemy i perspektivy]. *Proceedings of the 10th International Conference "Russia and China: history and perspectives for cooperation" [X Mezhdunarodnaya nauchno-prakticheskaya konferencziya «Rossiya i Kitaj: istoriya i perspektivy' sotrudnichestva»],* pp. 633-640.

[7] Semenov, K. (2020), Adaptation Strategies of Russian Phonetic Loanwords in Chinese: Phonetic and Graphic Aspects [Strategii preobrazovaniya russkix foneticheskix zaimstvovanij v kitajskom yazy'ke: foneticheskie i graficheskie aspekty']. *RSUH/RGGU Bulletin. "Philology. Linguistics. Culturology" Series [Vestnik RGGU. Seriya «Literaturovedenie. Yazy'koznanie. Kul`turologiya».]*. Vol. 7. pp. 30–63.

[8] Semenov K.I. (2020), Russian Loanwords in Standard Chinese: Problems of Phonetic and Morpological Adaptation, and Automatic Recognition in Chinese Texts [Zaimstvovaniya iz russkogo yazy'ka v standartnom kitajskom: problemy' foneticheskoj i morfologicheskoj adaptaczii i metody' avtomaticheskogo raspoznavaniya v kitajskix tekstax]. Bachelor Thesis at Higher School of Economics, Moscow.

[9] Clark K., Luong M., Le Q., Manning C. (2020), ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *International Conference on Learning Representations.*

[10] Gao J., Li M., Huang C., Wu A. (2005), Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach // *Computational Linguistics*. Vol. 4 (31). pp. 531–574.

[11] Honnibal M., Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. URL: https://spacy.io/.

[12] Huang H. pypinyin. URL: https://github.com/mozillazg/python-pinyin.

[13] Jia Y., Liu L., Chen H., Sun Y. (2020), A Chinese unknown word recognition method for microblog short text based on improved FP-growth // *Pattern Analysis and Applications.* Vol. 2 (23). pp. 1011–1020.

[14] Li P.-H., Ma W.-Y. CkipTagger. URL: https://github.com/ckiplab/ckiptagger.

[15] Luo E. xpinyin. URL: https://github.com/lxneng/xpinyin.

[16] Ownthink. Jiagu. URL: https://github.com/ownthink/Jiagu.

[17] Qi P., Dozat T., Zhang Y., Manning C. (2018), Universal Dependency Parsing from Scratch. *Proceedings of Association for Computational Linguistics, Brussels, Belgium*. pp. 160–170.

[18] Shen M., Kawahara D., Kurohashi S. (2016), Chinese Word Segmentation and Unknown Word Extraction by Mining Maximized Substring // Journal of Natural Language Processing. Vol. 3 (23). pp. 235–266.

[19] 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*. pp. 197–207.

[20] Sun X., Wang H., Li W. (2012), Fast Online Training with Frequency-Adaptive Learning Rates for Chinese Word Segmentation and New Word Detection Jeju Island, Korea: Association for Computational Linguistics. pp. 253–262.

[21] Wang L.-J., Li W.-C., Chang C.-H. (1992), Recognizing unregistered names for Mandarin word identification. *Association for Computational Linguistics*. p. 1239.

[22] Xia F. (2000), The Segmentation Guidelines for the Penn Chinese Treebank (3.0) // *IRCS Technical Reports Series*. Vol. 37.

[23] Xia F. (2000), The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0) // *IRCS Technical Reports Series*. Vol. 38.

[24] Yu S. [俞 士汶] (1999), Standards for Processing Modern Chinese Corpus - Word segmentation and part-of-speech tagging[现代汉语语料库加工规范——词语切分与词性标注] Beijing: Institute of Computational Linguistics, Peking University [北京大学计算语言学研究所].

[25] Zhang H. Shang J. (2019), NLPIR-Parser: An intelligent semantic analysis toolkit for big data [NLPIR-Parser：大数据语义智能分析平台] // Corpus Linguistics [语料库语言学]. Vol. 6(1). pp. 87–104.

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

[27] Academia Sinica (1999), Segmentation Principles for Chinese Language Processing[中文資訊處理分詞規範]. Taipei: Bureau of Standards, Metrology and Inspection of the Ministry of Economic Affairs of Taiwan[經濟部中央標準局].