Оценка эффективности гибридного морфологического анализатора NLTK4RUSSIAN в работе с текстами социальных сетей и художественных произведений

А.О. Кириллова, А.Г. Мельник, А.Д. Плетнева,
Е.В. Еникеева, О.А. Митрофанова

Санкт-Петербургский государственный университет

alyonkakirillova@gmail.com, i-ve-got-a-mail@inbox.ru,
adpletneva@perm.ru, protoev@yandex.ru, oa-mitrofanova@yandex.ru

Аннотация

В статье описан эксперимент по тестированию гибридного морфоанализатора NLTK4RUSSIAN на материалах соревнований «Dialogue Evaluation» 2017 г. Эксперименты с обучением морфоанализатора выполнены на основе подкорпусов НКРЯ и OpenCorpora. Для тестирования использовались выборки из художественных текстов и текстов социальных сетей. В ходе исследования решен ряд задач, в том числе задача конвертации морфологической разметки из формата Universal Dependencies в формат PyMorphy2, используемый в OpenCorpora. Результаты тестирования гибридного морфоанализатора NLTK4RUSSIAN соответствуют «золотому стандарту» для русского языка.

Ключевые слова: морфологический анализ, разрешение морфологической неоднозначности, NLTK4RUSSIAN, корпусы русскоязычных текстов

1. Введение

Морфологическая аннотация и разрешение морфологической неоднозначности при автоматической разметке текстов — задачи, которые имеют множество решений, отличающихся трудозатратами и стандартами качества. Среди систем, обеспечивающих достаточно точный морфологический анализ текста на русском языке, стоит выделить систему АОТ (http://www.aot.ru), анализатор mystem (https://tech.yandex.ru/mystem/) и парсер TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger). Существует распространенное мнение, что неоднозначность на одном языковом уровне должна сниматься за счет следующего уровня (в частности, неоднозначность в морфологии должна решаться с использованием данных о структуре синтаксических групп и синтаксических связей между словами в предложении) [1]. Такой метод может быть эффективным, но при этом он является дорогостоящим и требующим больших усилий.

В настоящее время развиваются специализированные модули для снятия омонимии, представляющие три основные группы:

- системы, основанные на правилах (созданных вручную);
- системы, основанные на вероятностных моделях и обученные на размеченных корпусах;
- гибридные системы, использующие как подход на правилах, так и вероятностные модели.

Системы, основанные на правилах, как раз похожи по своей реализации на предлагаемый подход к снятию омонимии с помощью следующего языкового уровня. Такие системы работают линейно. Системы второго типа, т.е. основанные на вероятностных моделях, работают медленнее, но их проще реализовать. Немаловажную

роль здесь играет размер обучающего корпуса. Некоторые системы должны обучаться на корпусах небольших размеров, так как большие корпуса могут повредить работе алгоритма.

Ежегодно в России проводится международная научная конференция по компьютерной лингвистике «Диалог» (http://www.dialog-21.ru/). Цель конференции — обмен идеями между специалистами в области лингвистики, а также представление накопленных знаний и результатов разработанных систем для автоматической обработки текста.

В рамках конференции проводится соревнование «Dialogue Evaluation» (http://www.dialog-21.ru/evaluation/), которое включает в себя тестирование по различным темам, таким как анализ тональности, анализ семантического сходства, машинный перевод и др. В 2017 г. одной из таких тем был морфологический анализ (http://www.dialog-21.ru/evaluation/2017/morphology/). Соревнование морфологических анализаторов проводилось в двух форматах: на закрытой дорожке оценивались алгоритмы, обученные на ограниченном количестве заранее предоставленных данных, а на открытой дорожке участники могли обучать анализаторы на любом материале. На обеих дорожках оценивалось качество частеречной разметки, полной разметки (т.е. определения части речи и значений основных лексико-грамматических категорий) и лемматизации. Для обучения алгоритмов на закрытой дорожке участникам были предоставлены размеченные корпусы со снятой вручную омонимией, а именно подкорпус Национального корпуса русского языка (НКРЯ, http://www.ruscorpora.ru/), подкорпус Генерального интернет-корпуса русского языка (ГИКРЯ, http://www.webcorpora.ru/), данные проекта OpenCorpora (http://opencorpora.org/) и материалы корпуса СинТагРус (http://www.ruscorpora.ru/instruction-syntax.html) [2].

Целью данного исследования является тестирование гибридного морфоанализатора NLTK4RUSSIAN на разных типах текстов. Ранее была проведена оценка качества работы морфоанализатора с привлечением тестовых выборок из НКРЯ и OpenCorpora [3]. На данном этапе представляется необходимым обновить эти результаты и протестировать морфоанализатор на корпусных данных второго соревнования по морфологическому анализу в рамках «Dialogue Evaluation». Эти тестовые материалы включают художественные и новостные тексты, а также текстовые материалы из социальной сети ВКонтакте.

Экспериментальный дизайн данного исследования включает в себя следующие этапы:

- исследование основных режимов работы морфоанализатора NLTK4RUSSIAN;
- создание конвертера, преобразующего теги, используемые для морфологической разметки в Universal Dependencies, в теги формата PyMorphy2;
- обучение морфоанализатора и его тестирование на выбранных данных;
- обработка результатов и оценка качества общей морфологической и отдельно частеречной разметок с помощью метрики «точность средства измерений» (accuracy);
- сравнение полученных данных для использованных корпусов.

2. Морфоаналогический анализатор NLTK4RUSSIAN

2.1. Принцип работы морфоанализатора

Анализатор NLTK4RUSSIAN — лингвистический комплекс для исследования русскоязычных корпусов текстов, разрабатывающийся на кафедре математической лингвистики СПбГУ (https://github.com/named-entity/nltk4russian) [3, 4, 5]. Главная задача — улучшение морфологического анализа с использованием уже существующих инструментов. NLTK4RUSSIAN называется гибридным морфоанализатором, так как объединяет в себе алгоритмы NLTK (http://www.nltk.org/) [6] и PyMorphy2 (http://pymorphy2.readthedocs.org/en/latest/) [7].

NLTK — Natural Language Toolkit — библиотека Python, которая предназначена для работы с текстом на всех уровнях языка. Средства библиотеки позволяют проводить как базовые операции (лемматизация, токенизация, стемминг), так и многоуровневый анализ с применением алгоритмов машинного обучения. NLTK также позволяет проводить морфологический анализ и разметку. В числе инструментов теггеры, использующие словарные данные, теггер на основе скрытых марковских моделей, анализаторы с применением контекстной информации и др. [6]. Важная особенность NLTK заключается в том, что в большинстве случаев используется размеченный корпус, что делает лингвистический комплекс независимым от языка текста для анализа. Одновременно с этим среда NLTK не адаптирована для конкретных языков, она использует универсальную метаинформацию. NLTK4RUSSIAN, таким образом, является адаптацией NLTK к русскоязычным текстам, создаваемой путем интеграции лингвонезависимых алгоритмов NLTK и морфологического анализатора для русского языка.

PyMorphy2 — анализатор русских текстов, который использует морфологический словарь OpenCorpora. Для этой системы не нужен обучающий корпус, все необходимые словоформы описаны в словаре. Для редких или несуществующих форм работает предсказатель, который выводит предполагаемый тег по префиксу и по концу слова. Для совпадающих словоформ или форм с неоднозначностью выводятся все возможные интерпретации, вероятность того или иного варианта рассчитывается не на основе контекста, а по данным словаря. Можно сказать, что NLTK4RUSSIAN предлагает альтернативное решение неоднозначности: если PyMorphy2 выводит несколько вариантов разбора, то теггер пользуется возможностями анализа соседних словоформ от NLTK. Таким образом, NLTK4RUSSIAN объединяет достоинства NLTK и PyMorphy2: текст анализируется с помощью анализатора PyMorphy2, для дизамбигуации вызывается теггер NLTK. При этом система может работать как в гибридном режиме, так и с использованием NLTK и PyMorphy2 по отдельности.

2.2. Режимы работы морфоанализатора

NLTK4RUSSIAN работает в 3 режимах, для переключения каждого из них при вызове морфоанализатора необходимо указать специальный параметр-название "-n":

- гибридный (PyMorphy2 + NLTK);
- PyMorphy2 (PyMorphyTagger);
- униграммный, биграммный и триграммный теггеры NLTK.

Гибридный режим — это режим с использованием PyMorphy2 и биграммного теггера NLTK nltk.tag.sequentials. При вызове анализатора указывается файл с текстом, размеченный корпус для обучения и название файла, который будет создан в ходе работы. Файл с неразмеченным текстом читается построчно. Вызывается функция MorphAnalyzer() от PyMorphy2, при обнаружении неоднозначности запускается биграммный теггер NgramTagger с соответствующим параметром. Режим получил название "PyMorphyContext", при запуске анализатора нужно указать параметр "-n" "pmcontext".

В режиме PyMorphy2 вызывается функция MorphAnalyzer() от PyMorphy2, и, как в гибридном режиме, выполняется разметка на основе словарных данных. Но и в отличие от PyMorphyContext, неоднозначность снимается с помощью оценки вероятности. Параметр рассчитывается на основе данных корпуса OpenCorpora: для всех слов со снятой неоднозначностью вычисляется, сколько раз словоформа была отмечена данным тегом. На основе этих частот и выводится условная вероятность тега (параметр "score"). (http://pymorphy2.readthedocs.io/en/latest/user/guide.html)

Последний режим анализирует текст на основании контекста — предшествующих словоформе тегов слов. Униграммный теггер (UnigramTagger) ищет подходящий вариант разбора в обучающем корпусе, биграммный (BigramTagger) анализирует тег предыдущего слова в тексте и ищет соответствие в обучающем корпусе, триграммный (TrigramTagger)

составляет контекст из разбора слова и двух предыдущих слов и ищет подходящие разборы в обучающем корпусе.

Для вызова морфоанализтора в одном из этих режимов при запуске нужно указать соответственно "-n 1gram", "2gram" или "3gram".

3. Тестирование морфоанализатора NLTK4RUSSIAN

3.1. Исходные данные

В качестве обучающих данных были взяты материалы НКРЯ и проекта OpenCorpora, в качестве тестовых — тексты, которые также являлись тестовыми для участников соревнований «Dialogue Evaluation» 2017 г., а именно корпусы текстов новостного сайта Lenta, социальной сети ВКонтакте и текст романа О. Зайончковского «Петрович» (далее — JZ). Стоит отметить, что данный материал представляет собой тексты с различными жанровыми и стилистическими особенностями.

Обучение морфоанализатора NLTK4Russian выполнено на размеченных текстах со снятой омонимией. Обучающие материалы включают выборки из НКРЯ, а также подкорпус OpenCorpora объёмом 30 тыс. словоупотреблений.

3.2. Предподготовка обучающих и тестовых выборок

На предварительном этапе тестирования была сформулирована задача конвертации разметки в формате Universal Dependencies в формат PyMorphy2, используемый в OpenCorpora. За основу для будущего конвертера был взят имеющийся скрипт для конвертации из pymorphy в Universal Dependencies. Отметим, что при конвертации тегов в другой формат часто возникают несоответствия (см. пример на рис. 1). В нашем случае большая часть таких несоответствий связана с тем, что в разметке OpenCorpora предлагается выделять 12 падежей (шесть стандартных, звательный, два дополнительных родительных, дополнительный винительный и два дополнительных предложных). Стандарт Universal Dependencies содержит только шесть тегов для падежей, поэтому при конвертации часть тегов в формате PyMorphy2 не восстанавливается.

```
'Case': {
    'Ins':  'ablt',
    'Acc':  'accs',
    'Dat':  'datv',
    'Gen':  'gen1',
    'Loc':  'loct',
    'Nom':  'nomn',
},
```

Рис. 1. Фрагмент кода для конвертации падежных тегов в формат pymorphy

Для файлов с тестовой выборкой потребовалась дополнительная обработка, связанная с извлечением слов для разметки. Эта задача была реализована при помощи специального скрипта. На вход поступает файл с тестовой выборкой (Lenta, VK или JZ), в котором из каждой строки извлекаются имеющиеся словоформы. На выходе мы получаем отдельный файл с извлечёнными формами, с которым далее работает морфоанализатор.

Обучающая выборка (материалы НКРЯ и OpenCorpora) уже содержит разметку, однако эта разметка представлена в формате Universal Dependencies. Поскольку на выходе мы должны получить размеченные тестовые материалы с тегами pymorphy, то и обучающая выборка должна содержать теги нужного нам формата. Поэтому обучающая выборка также потребовала предварительной обработки. С этой целью был разработан

дополнительный скрипт, который обрабатывает файлы с обучающей выборкой, последовательно заменяя в них теги Universal Dependencies на теги PyMorphy2.

Таким образом, на предварительном этапе мы подготовили обучающие и тестовые корпусы к тестированию морфоанализатора.

3.3. Тестирование морфоанализатора и результаты

Обучение и разметка тестовых выборок осуществлялась в соответствии со стандартной процедурой, описанной для компонента train-tagger гибридного морфоанализатора. Конфигурации для тестирования представлены в табл. 1. Морфоанализатор может размечать тестовые корпусы двумя способами: приписывая только частеречные теги или полную грамматическую характеристику данной словоформы. Для полной разметки указывается параметр -full . Соответственно, оценка результатов разметки выполняется по этим двум основаниям при помощи отдельного скрипта eval в составе морфоанализатора. Данные о точности разметки содержатся в табл. 2.

Таблица 1. Конфигурации для тестирования

№ эксперимента	Обучающий корпус	Тестовый корпус
1	НКРЯ (выборка)	JZ
2	НКРЯ (выборка)	Вконтакте
3	НКРЯ (выборка)	Lenta.ru
4	OpenCorpora	JZ
5	OpenCorpora	Вконтакте
6	OpenCorpora	Lenta.ru

Таблица 2. Точность разметки

№ эксперимента	1gram		2gram		3gram		pymorphy		pmcontext	
	Full	POS	Full	POS	Full	POS	Full	POS	Full	POS
1	88,1%	90,9%	**76,8%**	91,7%	**76,1%**	91,5%	**75,5%**	91,3%	**74,8%**	90,5%
2	89%	91,3%	75,5%	91,1%	72,6%	90,9%	70,2%	90,5%	69,3%	89,3%
3	**90,2%**	**93,8%**	68,3%	**91,9%**	66,5%	**91,6%**	65,3%	**91,5%**	64,8%	91%
4	85,4%	90,6%	76%	91,4%	75,4%	91,3%	**75,5%**	91,3%	**74,8%**	90,5%
5	86,1%	90,2%	71,8%	90,2%	70,6%	90,3%	70,2%	90,5%	69,4%	89,4%
6	86,6%	91,7%	66,6%	91,4%	66%	91,3%	65,3%	**91,5%**	65%	**91,1%**

Выводы

Таким образом, при оценке результатов было выяснено, что полученные данные практически не зависят от выбранного обучающего корпуса (например, результат частеречной разметки корпуса текстов автора JZ с НКРЯ и с корпусом OpenCorpora в качестве обучающего в режиме pmcontext оказался одинаковым — 74,8).

Точность частеречной разметки в среднем оказалась выше точности полной морфологической разметки; для частеречной разметки средний показатель равен 91%, а для полной морфологической разметки среднее значение заметно меньше — 74,2%.

Были протестированы все доступные режимы работы морфоанализатора NLTK4RUSSIAN, которые показали в основном примерно одинаковую точность разметки. Было выяснено, что результаты полной морфологической разметки варьируются в зависимости от жанра текстов: наибольшей точности удалось добиться на текстах автора JZ (76,8%), следом идёт корпус текстов социальной сети ВКонтакте (75,5%); наименьшая

точность наблюдается для текстов новостного корпуса (68,3%). В этом ряду выделяются результаты, полученные при работе в режиме униграммного теггера: точность разметки здесь оказалось вообще самой высокой, причем наибольшая точность наблюдается при разметке новостного корпуса (полная морфологическая разметка — 90,2%, частеречная — 93,8%).

Проведенные эксперименты основаны на исследовании возможностей морфологического анализатора, опубликованном в 2015 г. [3]. В ранней серии экспериментов для обучения использовались подкорпус OpenCorpora со снятой омонимией и две выборки из НКРЯ со снятой омонимией: подкорпуса художественных текстов и публицистики. Для тестирования были взяты выборки из НКРЯ. Показатели точности в эксперименте по частеречной разметке располагались в промежутке от 90 до 95,8%, по полной морфологической — от 79,7 до 86%. Самые высокие результаты наблюдались при разметке публицистических (частеречная разметка) и художественных (полная морфологическая разметка) текстов. В новом исследовании удалось улучшить качество разметки: наибольшая точность частеречной и полной морфологической разметок была получена при разметке новостного корпуса (93,8 и 90,2% соответственно).

Результаты нынешних экспериментов также сравнивались с показателями, полученными в ходе соревнований морфологических парсеров «Dialogue Evaluation» 2017 г. Базовые алгоритмы разметки показали результаты в промежутке от 68,44 до 76, 54% при полной морфологической разметке и от 72,1 до 79,49% при частеречной разметке. Лучший результат на данных соревнованиях был получен морфоанализатором ABBYY (точность частеречной разметки — 97,11%). Показатели остальных участников соревнований, полученные в этом сегменте соревнований (частеречная разметка), варьировались от 71,48 до 93,08%. Данные наблюдения позволяют подтвердить вывод о том, что результаты работы парсера NLTK4RUSSIAN являются состоятельными и удовлетворяют высоким требованиям соревнований морфологических анализаторов.

Исследование поддержано грантом РФФИ № 16-06-00529 «Разработка лингвистического комплекса для автоматического семантического анализа русскоязычных корпусов текстов с применением статистических методов» (2016–2018 гг.).

Литература

[1] Сокирко А.В, Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. М., 2005. URL: http://www.aot.ru/docs/RusCorporaHMM.htm (дата обращения 04.05.2018).

[2] Sorokin A., Shavrina T., Lyashevskaya O., Bocharov V., Alexeeva S., Droganova K., Fenogenova A., Granovsky D. MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог–2017». М., 2017. Т. 1, № 16. С. 297–313.

[3] Паничева П.В., Протопопова Е.В., Митрофанова О.А., Мирзагитова А.Р. Разработка лингвистического комплекса для морфологического анализа русскоязычных корпусов текстов на основе Pymorphy и NLTK // Труды международной конференции «Корпусная лингвистика – 2015». СПб, 2015. С. 361–373.

[4] Москвина А.Д., Орлова Д., Паничева П.В., Митрофанова О.А. Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии. Труды XIX Международной объединенной научной конференции «Интернет и современное общество», Санкт-Петербург, 22–24 июня 2016 г. СПб, 2016. С. 44–54.

[5] Паничева П.В., Митрофанова О.А. Интеграция морфоанализаторов для аннотации русскоязычных корпусов текстов // Сборник материалов по итогам XLIII Международной филологической конференции. Секция прикладной и математической лингвистики. СПб, 2014. С. 56–60.
[6] Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing, 2009.
[7] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015. Communications in Computer and Information Science, Springer, 2015. P. 320–332.

Testing and Assessment of the NLTK4RUSSIAN Hybrid Morphological Analyzer

A.K. Kirillova, A.G. Melnik, A.D. Pletneva, E.V. Enikeeva, O.A.Mitrofanova

Saint-Petersburg State University

This paper describes a series of experiments aimed at testing the hybrid morphological analyzer NLTK4RUSSIAN on the datasets of the Dialogue Evaluation 2017. We provide a detailed explanation of the principles and main operation modes of the system. Experiments on training the system were based on RNC and OpenCorpora subcorpora. Certain problems dealing with the data format were solved in course of research: we performed conversion of Universal Dependencies markup format to PyMorphy2 format. The analyzer was tested on social network subcorpus, news subcorpus and fiction subcorpus. Test results for the NLTK4RUSSIAN morphological analyzer correspond to the gold standard of morphological annotation for the Russian language.

Keywords: morphological tagging, morphological disambiguation, NLTK4RUSSIAN, text corpora