П.В. Паничева, О.А. Митрофанова

ИНТЕГРАЦИЯ МОРФОАНАЛИЗАТОРОВ ДЛЯ АННОТАЦИИ РУССКОЯЗЫЧНЫХ КОРПУСОВ ТЕКСТОВ

Морфологическая аннотация русских корпусов и разрешение морфологической неоднозначности – задачи, имеющие множество возможных решений, которые различаются по качеству и по трудоемкости [1]. Существующие открытые инструменты морфологического анализа для русского языка AOT (http://www.aot.ru) [2], mystem (https://tech.yandex.ru/mystem) [3], TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) [4] и некоторые другие показывают, что словарные и статистические модели компьютерной морфологии [5], реализуемые порознь, не позволяют производить морфологический анализ в полном объеме: в частности, не все инструменты предусматривают разрешение морфологической неоднозначности, либо ее точность недостаточно высока. Тем не менее, существуют закрытые процессоры, обеспечивающие точность морфологического анализа более 95% [6–8]. В нашем исследовании предлагается метод улучшения качества морфологического анализа русскоязычных текстов, предполагающий интеграцию нескольких открытых инструментов, а именно, морфоанализатора PyMorphy2 (http://pymorphy2.readthedocs.org/en/latest/) и морфологических теггеров в составе NLTK (http://www.nltk.org/) [9].

PyMorphy2 – морфологический анализатор русских текстов, который работает с морфологическим словарем OpenCorpora (http://opencorpora.org/) [10], создаваемым на основе базы данных «Грамматического словаря русского языка» А.А. Зализняка [11]. При разборе несловарных словоформ используется предсказатель, объединяющий два алгоритма: предсказание по префиксу и по концу слова. PyMorphy2 может предлагать несколько вариантов разбора. Всем вариантам приписывается параметр *score* ($P(tag|word)$), который определяется по данным OpenCorpora. При необходимости выбора одного разбора из множества выбирается наиболее вероятный разбор. В

PyMorphy2 при морфологическом анализе контекстная информация в явном виде не используется.

NLTK (Natural Language Toolkit) [9] – специализированная среда для автоматической обработки текстов, созданная для работы с Python и оснащенная библиотеками и лингвистическими данными (корпусами и словарями), NLTK позволяет осуществлять полный цикл автоматической обработки текста (графематический анализ, токенизация, стемминг, лемматизация, морфологический анализ, фрагментационный анализ, построение синтаксических структур и логических форм для предложений во входном тексте), а также ряд процедур, связанных с классификацией и кластеризацией единиц корпуса. Предусмотрены средства для автоматического извлечения фактов и оценки тональности текстов, а также ряд других операций.

Наш интерес к NLTK определяется богатством его ресурсов для морфологического анализа. Так, в NLTK имеются различные морфологические теггеры – специальные средства для морфологического анализа, опирающиеся на словари и / или контекстные данные: RegexpTagger, NgramTagger, AffixTagger, HMMTagger, BrillTagger. Эти теггеры можно использовать по отдельности или же в комбинации «основной теггер – бэк-офф теггер(ы)». Большим недостатком NLTK является то, что до сих пор отсутствует его адаптация для работы с русским языком. Наше исследование частично восполняет этот пробел.

Нами был разработан инструмент, интегрирующий морфологический анализатор PyMorphy2 и теггеры NLTK. Первый этап обработки производится с помощью PyMorphy2. Если PyMorphy2 дает несколько вариантов разбора, то происходит вызов биграммного теггера NLTK, который выбирает наилучший вариант разбора на основе тегов предшествующей словоформы: выбирается тот тег, который наиболее часто встречается последовательно в биграмме с предыдущим тегом в обучающей выборке. На данный момент эксперименты проводились с биграммным теггером, бэк-офф теггеры не были задействованы (хотя инструмент их предусматривает).

В тестах использовался подкорпус OpenCorpora [10] со снятой омонимией (объем примерно 5,5 тыс. предложений, 30 тыс. токенов). На основе данного подкорпуса формировались две выборки – обучающая и тестовая. Была проведена серия экспериментов по морфологическому анализу с изменением объема обучающей выборки: от 10 до 100 предложений с шагом 10 (10, 20, 30…100), от 100 до 1000 предложений с шагом 100 (100, 200, 300, … 1000), от 1000 до 3000 предложений с шагом 1000 (1000, 2000, 3000). Наилучший результат по точности был получен при объеме обучающей выборки 400 предложений (т.е. примерно 10% анализируемого подкорпуса). Эксперименты проводились на тестовой выборке объемом 25,5 тыс. с/у. Точность разбора в этом случае составляет 84,4%. Для сравнения укажем, что разбор, который выдает PyMorphy2 на основе статистической оценки, верен примерно в 79% случаев. Мы рассмотрели контексты с разметкой, не соответствующей разметке в подкорпусе OpenCorpora (15,6% тестовой выборки) со снятой омонимией и выделили пять типов расхождений (табл. 1).

Таблица 1.

Типы несоответствий между разметкой **OpenCorpora и PyMorphy2+NLTK**

Тип	**Пример**	**Open Corpora**	**PyMorphy2 +NLTK**	**Кол-во случаев**	**Доля, %**	**Примечание**
Всего				**4044**	**100**	
1	*kn0pkaaaa*	*UNKN*	*LATN*	**121**	**2,99**	В OpenCorpora словоформа не разобрана, комбинированный теггер ее размечает верно, либо также выдает нулевой результат разбора.
2	*4.5.3.3*	*UNKN*	*None*	**596**	**14,74**	
3	*0306817527*	*NUMB*	*NUMB,**intg***	**1425**	**35,24**	Разбор правильный, однако в OpenCorpora и в результатах работы комбинированного теггера нет полного соответствия в числе признаков, комбинированный теггер выдает более детальный разбор.
4	*Франции*	*NOUN,inan,femn,Sgtm,Geox,sing,**datv***	*NOUN,inan,femn,Sgtm,Geox,sing,**gent***	**806**	**19,93**	Разбор возможно правильный, OpenCorpora и комбинированный теггер по-разному интерпретируют один или несколько признаков в разметке.

	доступ	*NOUN,inan,masc,sing,**accs***	*NOUN,inan,masc,sing,**nomn***			
5	*ширялась* *англ* *антигла-мурщики*	*UNKN* *UNKN* *UNKN*	***NOUN,anim,femn,Name,sing,voct,Infr*** ***NOUN**,inan,masc,sing,nomn* *NOUN,anim,masc,plur,nomn*	**1096**	**27,10**	В OpenCorpora словоформа не разобрана, комбинированный теггер выдает результат, правильность которого может быть оценена только в ходе ручной проверки.

Очевидно, что случаи несоответствий типов 1, 2 и 3 связаны с различными наборами тегов, используемых в корпусе OpenCorpora и в комбинированном теггере. На эти случаи приходится 2142 несоотвествия (53% всех несоответствий). Истинными ошибочными разборами следует считать случаи типа 4; также в данной работе мы причисляем к таким случаям тип 5, оценка правильности которого может быть произведена исключительно вручную. Для простоты учета мы считаем их ошибочными, округляя оценку результатов комбинированного теггера в сторону уменьшения. На истинные ошибочные случаи приходится 1902 несоответствия (47% всех несоответствий). Тем самым, при должном совпадении тегсетов точность морфологического анализа достигает не менее чем 92,7%. Эта цифра может быть уточнена в сторону увеличения в результате ручной проверки несовпадений, в которых OpenCorpora не содержит разбора словоформы.

Наше исследование показало продуктивность интеграции морфологических анализаторов для повышения качества автоматической обработки русскоязычных текстов и целесообразность использования инструмента NLTK для исследований, проводимых на материале русских корпусов. Отличительной особенностью проводимого исследования является то, что в нем используются открытые программные средства и лингвистические ресурсы. Перспективы исследования связаны с оптимизацией системы тегов для морфологической разметки и с проведением экспериментов с подключением различных бэк-офф теггеров в составе комплекса PyMorphy2+NLTK.

Литература

1. *Леонтьева Н.Н.* Автоматическое понимание текстов. Системы, модели, ресурсы. М., 2006.
2. *Сокирко А.В.* Морфологические модули на сайте www.aot.ru // http://www.aot.ru/docs/sokirko/Dialog2004.htm
3. *Segalovich I.* A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine // MLMTA-2003 // http://company.yandex.ru/technologies/mystem/
4. *Sharoff S.* Methods and Tools for Development of the Russian Reference Corpus // D. Archer, A. Wilson, P. Rayson (eds.) Corpus Linguistics Around the World. Amsterdam, 2005.
5. *Коваль С.А.* Лингвистические проблемы компьютерной морфологии. СПб., 2005.
6. *Protopopova E.V., Bocharov V.V.* Unsupervised Learning of Part-of-Speech Disambiguation Rules // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). М., 2013.
7. *Сокирко А.В, Толдова С.Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. Автоматическая обработка веб-данных. М., 2005.
8. *Толдова С., Савчук С., Коваль С.* Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М., 2010.
9. *Bird S., Klein E., Loper E.* Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing, 2009.
10. *Бочарв В.В., Грановский Д.В.* Как и зачем мы делаем Открытый корпус // Семинар по автоматической обработке текста/ URL: http://opencorpora.org/doc/presentations/2011_NLPSeminar.pdf
11. *Зализняк А.А.* Грамматический словарь русского языка, М., 1977, …, 2003.