Паничева П.В., Протопопова Е.В.,
Мирзагитова А.Р., Митрофанова О.А.
Panicheva P.V., Protopopova E.V.,
Mirzagitova A.R, Mitrofanova O.A.

РАЗРАБОТКА ЛИНГВИСТИЧЕСКОГО КОМПЛЕКСА ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА РУССКОЯЗЫЧНЫХ КОРПУСОВ ТЕКСТОВ НА ОСНОВЕ PYMORPHY И NLTK[1]

DEVELOPMENT OF AN NLP TOOLKIT FOR MORPHOLOGICAL ANALYSIS OF RUSSIAN TEXT CORPORA BASED ON PYMORPHY AND NLTK[1]

Аннотация. В статье описан метод улучшения качества морфологического анализа русскоязычных текстов, предполагающий интеграцию морфоанализатора PyMorphy2 и морфологических теггеров NLTK. Рассмотрены два варианта проведения морфологического анализа: (1) первичная обработка с помощью PyMorphy2, доразметка с помощью теггеров NLTK; (2) первичная обработка с помощью теггеров NLTK, доразметка с использованием предсказателя PyMorphy2. Эксперименты с обучением и тестированием комплекса проводились на основе подкорпусов OpenCorpora и НКРЯ.

Ключевые слова. морфологическая разметка, разрешение морфологической неоднозначности, PyMorphy2, NLTK, русскоязычные корпусы текстов

Abstract. The paper describes a technique which allows to improve the quality of Russian morphological tagging and implies integration of morphological analyzer PyMorphy2 and NLTK taggers. We consider two approaches to morphological analysis: (1) primary processing by PyMorphy2, secondary processing by NLTK taggers; (2) primary processing by NLTK taggers, secondary processing by PyMorphy2. Experiments on training and testing the toolkit were based on OpenCorpora and RNC subcorpora.

Keywords. morphological tagging, morphological disambiguation, PyMorphy2, NLTK, Russian text corpora

---

1 Исследование выполнено при частичной финансовой поддержке гранта СПбГУ 30.38.305.2014 «Квантитативные лингвистические параметры определения стилевых характеристик и предметной области текстов».

1. Введение

Морфологическая аннотация и разрешение морфологической неоднозначности в корпусах текстов – задачи, имеющие множество решений, которые различаются по качеству и по трудоемкости. Особенно сложно решать данные задачи при автоматической обработке текстов флективных языков с развитой морфологией, каковыми, в частности, являются славянские языки, применительно к которым используются словарные, статистические и гибридные модели компьютерной морфологии (ср. [Garabík 2005, Hajič 2004; Hajič et al. 2001, Коваль 2005] и т.д.). Для русского языка широко используются разноплановые открытые инструменты морфологического анализа: основанные на компьютерной морфологической базе системы АОТ (http://www.aot.ru) [Сокирко 2004], mystem (https://tech.yandex.ru/mystem) [Segalovich 2003], использующие статистические алгоритмы парсеры TnT (http://www.coli.uni-saarland.de/~thorsten/tnt/), TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) [Sharoff 2005], обеспечивающие точность разбора в пределах «baseline». Вместе с тем, существуют закрытые парсеры, обеспечивающие точность морфологического анализа более 95% [Protopopova, Bocharov 2013; Sokirko, Toldova 2005].

В отечественном лингвистическом сообществе действует форум, посвященный разработке стандартов морфологического анализа и оценке качества разрабатываемых морфологических парсеров [Ляшевская и др. 2010] – модулей, направленных на обработку всех цепочек слов в тексте и при этом учитывающих контекстную информацию (в частности, результаты разбора контекстных соседей). Принято различать два вида парсеров – сильные и слабые (предусматривающие / не предусматривающие разрешение морфологической неоднозначности). *Наша задача – разработать сильный гибридный парсер для морфологического анализа русских текстов, совмещающий словарные данные и не один, а несколько статистических алгоритмов.*

В настоящем исследовании предлагается метод улучшения качества морфологического анализа русскоязычных текстов, предполагающий интеграцию нескольких открытых инструментов, а именно, морфоанализатора PyMorphy2 (http://pymorphy2.readthedocs.org/en/latest/) [Korobov 2015] и различных морфологических теггеров в составе библиотек NLTK (http://www.nltk.org/) [Bird et al. 2009]. *Это первый опыт такого рода в компьютерной обработке русскоязычных корпусов текстов.*

2. Используемые инструменты морфологического анализа

PyMorphy2 – морфологический анализатор русских текстов, который работает с морфологическим словарем OpenCorpora (http://opencorpora.org/), создаваемым на основе базы данных «Грамматического словаря русского языка» А.А. Зализняка [Зализняк 2003]. В PyMorphy2 также разрешено подключение пользовательских словарей. Алгоритм производит морфологический анализ с определением грамматических характеристик целевого слова и лемматизацией. 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 частотный разбор для данной словоформы в подкорпусе с разрешенной неоднозначностью . Таким образом, в PyMorphy2 при морфологическом анализе контекстная информация в явном виде не используется.

NLTK – 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 процедуры классификации и кластеризации. 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, опирающиеся на словари и/или контекстные данные: это теггер на регулярных выражениях, *n*-граммные теггеры, аффиксный теггер, теггер на скрытых марковских моделях, теггер Брилла. Эти теггеры можно использовать по отдельности или же в комбинации «основной теггер – бэк-офф теггер(ы)». *До сих пор отсутствует широкодоступная адаптация NLTK для работы с русским языком. Наше исследование частично восполняет этот пробел.*

3. Интеграция алгоритмов морфологического анализа в комплексе PyMorphy2+NLTK

Нами был разработан лингвистический комплекс, интегрирующий морфологический анализатор PyMorphy2 и теггеры NLTK в двух вариантах (далее PyMorphy2+NLTK).

1. Первый вариант обработки тестового корпуса 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.

2. Второй вариант обработки производится с помощью теггеров NLTK: униграммного, биграммного и триграммного (http://www.nltk.org/api/nltk.tag.html#module-nltk.tag.sequential).

Далее проводится доразметка тестового корпуса с учетом

предсказаний PyMorphy2 тогда, когда отсутствует разбор в первой модели.

До настоящего времени ни один морфологический парсер для русского языка не предусматривал такую комбинацию алгоритмов.

4. Лингвистические данные

В тестировании комплекса PyMorphy2+NLTK использовался подкорпус OpenCorpora со снятой омонимией (объем примерно 6 тыс. предложений, 30 тыс. с/у) и две выборки из Национального корпуса русского языка (http://ruscorpora.ru/corpora-usage.htm) со снятой омонимией из подкорпуса художественных текстов (RNC-fiction) и публицистики (RNC-media). Подробные сведения об этих данных приведены в табл. 1.

Таблица 1. Используемые корпусные данные

Корпус	**Объём: предложения**	**Объём: с/у**
OpenCorpora	3000	17666
RNC-fiction	3000	37715
RNC-media	4303	73026

Надо отметить, что принципы морфологической разметки используемых корпусов значительно отличаются. Кроме того, несмотря на то, что PyMorphy2 основывается на словаре OpenCorpora, некоторые теги в результатах работы морфоанализатора не совпадают со словарными. В связи с этим мы сопоставили разметку из трёх источников и разработали инструмент для приведения разметки обоих корпусов к формату PyMorphy2.

Большая часть помет имеет однозначное соответствие. Граммемы из следующих категорий – род, число, падеж, одушевлённость, вид, время, переходность, лицо, наклонение – совпадают. Проанализированные нами различия в основном касаются описания специфических категорий (тип местоимения, описание существительных singularia и pluralia tantum), а также несловарных

форм. Для приведения разметки к одному формату были оставлены только грамматические и лексико-грамматические категории.

5. Эксперименты по морфологическому анализу и оценка результатов тестирования комплекса PyMorphy2+NLTK

Для тестирования комплекса PyMorphy2+NLTK использовались комбинации обучающих и тестовых корпусов, указанные в табл. 2. В табл. 3–4 представлены результаты оценки двух вариантов интеграции морфоанализаторов. Нами оценивались следующие параметры: *точность работы комплекса* 1) *на всех морфологических категориях* и 2) *на частеречной разметке.* В табл. 3 приводится точность работы комплекса на всех морфологических категориях, в табл. 4 – точность работы комплекса на частеречной разметке. Для сравнения приводятся, во-первых, результаты тестирования модуля PyMorphy2 с внутренним алгоритмом снятия неоднозначности, основанном на статистике корпуса OpenCorpora. Далее в столбце PM-disamb мы указали результаты первого комбинированного подхода: использование PyMorphy2 с алгоритмом снятия неоднозначности на основе предыдущего тега. Затем в столбцах P1, P2, P3 приведены результаты второго комбинированного подхода, основанного на теггерах NLTK с доразметкой PyMorphy2.

Показатели точности в серии экспериментов по частеречной разметке располагаются в промежутке **P = 90,0…95,8%.** Наилучшие результаты были получены в экспериментах с подкорпусом публицистики (RNC-media). Значения точности в серии экспериментов со всеми морфологическими категориями находятся в промежутке **P = 79,7…86,0%**. Здесь наилучшие результаты были получены в экспериментах с подкорпусом художественных текстов (RNC-fiction).

Таблица 2. Конфигурации тестирования комплекса

№ экспер.	Обучающий корпус	Объем обуч. корпуса	Тестовый корпус	Объём тест. корпуса
1	OpenCorpora	1500	RNC-fiction	1500
2	OpenCorpora	1500	RNC-fiction	1500
3	RNC-fiction	1500	RNC-fiction	1500
4	RNC-fiction	1500	RNC-fiction	1500
5	RNC-media	2129	RNC-media	2174
6	RNC-media	2174	RNC-media	2129

Таблица 3. Точность работы комплекса на всех морфологических категориях

№ экспер.	PyMorphy2	PM-disamb	P1	P2	P3
1	84,0	81,3	82,8	83,6	83,4
2	84,0	81,1	74,5	83,0	83,1
3	84,0	84,4	82,7	83,3	83,3
4	84,0	84,2	86,1	83,5	83,4
5	82,6	84,1	**86,0**	83,5	83,5
6	82,7	85,0	82,9	79,7	79,7

Таблица 4. Точность работы комплекса на частеречной разметке

№ экспер.	PyMorphy2	PM-disamb	P1	P2	P3
1	93,0	90,2	93,1	92,9	92,9
2	92,7	90,0	93,1	92,8	92,8
3	93,0	91,6	92,8	92,8	92,8
4	92,7	91,2	95,0	93,0	92,9
5	94,1	92,1	95,0	93,0	93,0
6	94,2	92,2	**95,8**	92,6	92,6

Подробный анализ ошибок разбора представлен в статье [Паничева, Митрофанова 2015]. В настоящем эксперименте некоторые виды ошибок удалось исключить за счет приведения различных видов разметки к единому формату. Кроме того, можно отметить, что лучшие значения получаются при обучении и тестировании анализатора на корпусах одного жанра. Однако обучение на подкорпусе OpenCorpora также дает неплохие результаты, что, вероятно, подтверждает его сбалансированность, несмотря

на небольшое количество предложений с полностью снятой морфологической неоднозначностью.

6. Оценка полученных результатов

Сравнение морфологических парсеров [Ляшевская и др. 2010] предполагает решение нескольких типов задач. Для сильных парсеров предусмотрены две дорожки: «Дизамбигуация: леммы» и «Дизамбигуация: частеречные теги». В нашем случае решается существенно *более сложная задача*, а именно, проводятся эксперименты по морфологическому анализу *всего текста* и разрешению неоднозначности *по всей морфологической аннотации*, включающей и теги лемм, и частеречные теги, и теги грамматических категорий.

В ходе соревнования морфологических парсеров был выработан стандарт морфологической разметки, при этом экспертная оценка точности ручной разметки эталонного корпуса такова: леммы – 94,4%, частеречные теги – 95,4%, теги грамматических категорий – 89,0%, вся морфологическая аннотация – 85,5%. В нашем случае (с изначально более жесткими условиями) точность разрешения неоднозначности по всей морфологической аннотации (**86%**) превышает точность разметки эталонного корпуса (**85,5%**). В варианте экспериментов по частеречной аннотации полученная нами точность достигает **95,8%,** тогда как эталонные результаты составляют **95,4%.** Эти наблюдения позволяют сделать заключение о том, что *результаты работы комбинированного анализатора являются состоятельными и удовлетворяют высоким требованиям соревнования морфологических парсеров для русского языка.* Более детальное сравнение нашего парсера и других имеющихся открытых ресурсов затруднено из-за различий во внутреннем устройстве этих систем, а также из-за различий в выборе обучающих и тестовых корпусов.

6. Заключение

В статье (1) дана характеристика основных инструментов морфологического анализа русских текстов, (2) представлена архитектура разрабатываемого лингвистического комплекса PyMorphy2+NLTK и входящие в него основные открытые алгоритмы, (3) описаны анализируемые подкорпусы текстов (OpenCorpora, RNC-fiction, RNC-media), (4) изложен ход экспериментов по обучению и тестированию комплекса, (5) предложена интерпретация полученных данных и дана оценка результатов. Можно утверждать, что *использование комбинаций алгоритмов, реализованных в нашем комплексе, обеспечивает высокое качество разметки* в том числе для таких типов текстов, которые значительно отличаются от корпуса OpenCorpora, используемого для внутренней настройки алгоритма PyMorphy2. *Комбинирование алгоритмов* позволяет также приблизить значение точности частеречной разметки к верхней границе для русскоязычных ресурсов (свыше 95%).

Литература

1. *Зализняк А.А.* (2003), Грамматический словарь русского языка. М.

2. *Коваль С.А.* (2005), Лингвистические проблемы компьютерной морфологии. СПб.

3. *Ляшевская О.Н. и др.* (2010), Оценка методов автоматического анализа текста: морфологические парсеры русского языка. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог–2010». Вып. 9(16). М.

4. *Паничева П.В., Митрофанова О.А.* (2015), Интеграция морфоанализаторов для аннотации русскоязычных корпусов текстов. Сборник материалов по итогам XLIII Международной филологической конференции. Секция прикладной и математической лингвистики. СПб.

5. *Сокирко А.В, Толдова С.Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. М., 2005.

6. *Сокирко А.В.* Морфологические модули на сайте www.aot.ru // http://www.aot.ru/docs/sokirko/Dialog2004.htm

7. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing, 2009.

8. *Garabík R.* (2005), Levenshtein Edit Operations as a Base for a Morphology Analyzer. Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2005. Ed. R. Garabík. Bratislava.

9. *Hajič J.* (2004), Disambiguation of Rich Inflection (Computational Morphology of Czech). Prague.

10. *Hajič J., Krbec P., Květoň P., Oliva K., Petkevič V.* (2001), Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. ACL.

11. *Korobov M.* (2015), 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 [in press].

12. *Protopopova E.V., Bocharov V.V.* (2013), Unsupervised Learning of Part-of-Speech Disambiguation Rules. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог–2013». Вып. 12(19). М.

13. *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/

14. *Sharoff S.* (2005), Methods and Tools for Development of the Russian Reference Corpus // D. Archer, A. Wilson, P. Rayson (eds.) Corpus Linguistics Around the World. Amsterdam.

References

1. *Bird S., Klein E., Loper E.* (2009), Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. Beijing, 2009.

2. *Garabík R.* (2005), Levenshtein Edit Operations as a Base for a Morphology Analyzer. Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2005. Ed. R. Garabík. Bratislava.

3. *Hajič J.* (2004), Disambiguation of Rich Inflection (Computational Morphology of Czech). Prague.

4. *Hajič J., Krbec P., Květoň P., Oliva K., Petkevič V.* (2001), Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. ACL.

5. *Korobov M.* (2015), 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 [in press].

6. *Koval' S.A.* (2005), Lingvisticheskije problemy kompjuternoj morfologii [Linguistic Problems of Computational Morphology]. St.-Petersburg.

7. Lyashevskaya O.N. et al. (2010), Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka [Evaluation of NLP Methods: Morphological Parsers of Russian]. Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii «Dialog–2010». Vyp. 9(16) [Computational Linguistics and Intellectual Technologies: Proceedings of International Conference «Dialog–2010». Vol. 9(16)]. Moscow.

8. *Panicheva P.V., Mitrofanova O.A.* (2015). Integracija morfoanalizatorov dlya annotacii russkojazychnyh korpusov tekstov [Integration of Morphological Analyzers for Annotation of Russian Text Corpora]. Sbornik materialov po itogam XLIII Mezhdunarodnoj filologicheskoj konferencii. Sekcija prikladnoj i matematicheskoj lingvistiki [Proceedings of XLIII International Philological Conference. Applied and Mathematical Linguistics Workshop.]. St.-Petersburg.

9. *Protopopova E.V., Bocharov V.V.* (2013), Unsupervised Learning of Part-of-Speech Disambiguation Rules. Komp'juternaja lingvistika i intellektual'nye tehnologii: Po materialam mezhdunarodnoj konferentsii «Dialog–2013». Vyp. 12(19) [Computational Linguistics and Intellectual Technologies: Proceedings of International Conference «Dialog–2013». Vol. 12(19)]. Moscow.

10. *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/

11. *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.

12. *Sokirko A.V., Toldova S.Ju.* (2005), Sravnenije effektivnosti dvuh metodik snjatija leksicheskoj i morfologicheskoj neodnoznachnosti dlya russkogo jazyka [Comparison of effectiveness of two techniques of lexical and morphological disambiguation in Russian]. Internet-matematika 2005 [Internet-mathematics 2005]. Moscow.

13. *Sokirko A.V.* (2004), Morfologicheskije moduli na sajte www.aot.ru [Morphological Modules on www.aot.ru Website]. http://www.aot.ru/docs/sokirko/Dialog2004.htm

14. *Zaliznyak A.A.* (2003), Grammaticheskij slovar' russkogo jazyka [Grammatical Dictionary of the Russian Language]. Moscow.

Паничева Полина Вадимовна, Протопопова Екатерина Владимировна, Мирзагитова Алия Ришатовна, Митрофанова Ольга Александровна

С.-Петербургский государственный университет (Россия).

Panicheva Polina, Protopopova Ekaterina, Mirzagitova Aliya, Mitrofanova Olga

St.-Petersburg State University (Russia).

E-mail: ppolin86@gmail.com, protoev@gmail.com, amirzagitova@gmail.com, oa-mitrofanova@yandex.ru