УДК 004.912

СРАВНЕНИЕ ИНСТРУМЕНТОВ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ
MORPHOLOGICAL TAGGING TOOLS COMPARISON

Асирян А.К. / Asiryan A.K.

Московский государственный университет им. М.В.Ломоносова, факультет вычислительной математики и кибернетики, Москва, Ленинские Горы 1, стр. 52, 119234

Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Moscow, Leninskie Gory 1, str. 52, 119234

Аннотация. В работе представлено сравнение пяти морфологических анализаторов: TreeTagger, MyStem, TnT, pymorphy2 и FreeLing. Для оценки качества были рассмотрены три корпуса текстов: СинТагРус, Открытый корпус и ГИКРЯ. Перед началом работы инструментов выбранный корпус был обработан, а наборы граммем унифицированы в соответствии с ним. Лучший результат достиг 94% меры F1.

Ключевые слова: обработка естественного языка, автоматический морфологический анализ, морфологические анализаторы, корпуса, снятие омонимии.

Вступление.

Морфологический анализатор или теггер – это инструмент, определяющий грамматические характеристики слова (часть речи и соответствующий ей набор граммем, например: число, падеж, лицо, время и т. д.). Одной из главных проблем является омонимия: в предложении «Мама мыла раму.» слово «мыла» – глагол (начальная форма «мыть»), а в «Не было мыла.» – существительное (начальная форма «мыло»). Также при сравнении инструментов возникает не менее важный вопрос: унификация наборов граммем. Уже для трех инструментов задача является сложной, при этом становится неясно как измерять итоговое качество: по всем граммемам вместе или по отдельности.

Поэтому было принято решение сравнивать только часть речи – самую важную характеристику для анализа. Стоит отметить, что сравнение теггеров уже производилось в некоторых работах [1, 2]. В обоих случаях в сравнении не участвовал инструмент MyStem [3], так как были взяты корпуса, размеченные этим инструментом, что не позволяло правильно оценить его работу. Также остается без ответа вопрос предобработки корпуса. Обе работы выделяют TreeTagger [4] в качестве лучшего.

Основной текст

В данной работе был рассмотрен ряд морфологических анализаторов, каждый из которых предлагает свой подход для решения задачи морфологической разметки. Часть из них основывается на словарном подходе, часть – на статистическом. Все инструменты распространяются либо с открытым исходным кодом, либо под лицензией, разрешающей использование в исследовательских целях.

TreeTagger – независимый от языка теггер, основанный на скрытых марковских моделях, отличающийся от традиционных методов оценкой вероятности перехода. Для этого TreeTagger использует бинарное дерево решений, построенное рекурсивно из обучающего множества триграмм (модифицированный алгоритм ID3). Также анализатор может распознавать неизвестные слова, используя лексикон суффиксов, представленный в виде дерева. Инструмент должен быть предварительно обучен на некоторых данных.

MyStem – морфологический анализатор русского языка с поддержкой снятия морфологической неоднозначности, разработанный в компании «Яндекс». Алгоритм базируется на словаре Зализняка [5]. Инструмент способен формировать морфологические гипотезы о незнакомых словах, используя словарь как множество префиксных деревьев, пытаясь найти наиболее близкое. Для разрешения омонимии MyStem использует алгоритмы машинного обучения с учетом и без учета контекста.

Третьим инструментом был выбран TnT [6]. Это морфологический анализатор, являющийся реализацией алгоритма Витерби для марковских

моделей второго порядка. Для сглаживания используется линейная интерполяция униграмм, биграмм и триграмм, веса которой определяются удаленной интерполяцией. Неизвестные слова обрабатываются с помощью анализа суффиксов. Как и в случае с TreeTagger необходимо обучение. Обученные модели для этих инструментов были взяты на странице Сергея Шарова [7].

Следующим морфологическим анализатором является pymorphy2 [8], использующий словарь OpenCorpora [9]. Алгоритм основывается на лингвистических правилах и может обрабатывать неизвестные слова. Последнее возможно за счет набора правил, например: отсечение известных префиксов, предсказание по концу слова, анализ составных слов, и т. д.

FreeLing [10] – проект создания свободного пакета инструментов обработки естественного языка. Состоит из таких модулей, поддерживающих русский язык, как:

а) Морфологический анализатор со снятием омонимии, основанный на скрытых марковских моделях.
б) Разбиение текста на предложения и токены.
в) Распознавание именованных сущностей.

Проверка работы инструментов требует тестовых данных. Были изучены существующие корпуса текстов русского языка, которые могли бы для этого использоваться. На текущий момент имеются три достаточно больших корпуса русского языка с морфологической разметкой. Это:

а) СинТагРус [11]: примерно 1 млн. слов. Корпус размечается в полуавтоматическом режиме. Сначала обработка текста производится морфологическим и синтаксическим анализаторами проекта ЭТАП-3 [12]. Полученный результат проверяется и при необходимости корректируется лингвистом. Корпус состоит из структур со снятой морфологической и синтаксической омонимией.
б) Открытый корпус русского языка (OpenCorpora): примерно 1.5 млн. слов, из них только 50 тыс. слов со снятой омонимией. Обработка текста

происходит с использованием морфологического анализатора АОТ [13]. Неоднозначности снимаются за счет привлечения к проекту пользователей сети Интернет.

в) Генеральный Интернет-корпус Русского Языка (ГИКРЯ) [14]: примерно 2 млн. слов. Текст берется из различных популярных сайтов и форумов. Производится автоматическая морфологическая разметка со снятием омонимии инструментом ABBYY Compreno [15].

Из рассмотренных корпусов был выбран СинТагРус, так как омонимия в нем снята с помощью лингвистов, и оценка морфологических анализаторов не зависела от неточностей в выборке. В подкорпусе OpenCorpora со снятой омонимией пока еще слишком мало слов для качественного анализа. Перед сравнением работы инструментов корпус был обработан. Были удалены предложения, в которые входили элементы, состоящие из нескольких токенов, например, производный предлог «в связи с». Это связано с тем, что инструменты плохо обрабатывают такие случаи, так как считают пробел разделителем токенов. Не рассматривались предложения, в которых элементы содержали точку (например, сокращения «т. д.», «мм.»), потому что они ухудшают качество оценки. Удалялись именно предложения, а не слова, так как большинство инструментов анализируют слова в контексте, а не отдельно. Таким образом, удаление одного слова может отразиться на всем предложении. Также были удалены предложения, содержащие части речи COM и NID, отвечающие композитам (вице, квази, экс, ультра и т. д.) и иноязычным вкраплениям или несловесным формулам (например, Berliner Zeitung, Щ243) соответственно. Это обосновано тем, что почти все инструменты не выделяют ту или иную часть речи. Были исключены знаки препинания: они не влияют на анализ, но затрудняют последующую проверку качества. Итоговый тестовый набор состоял из 57 036 предложений и 768 957 слов.

Как говорилось выше, при сравнении анализаторов всплывает проблема унификации набора граммем. Например, в корпусе СинТагРус причастия не выделяются в отдельную часть речи, а указываются в виде граммемы глагола. В

инструменте pymorphy2 оно наряду с деепричастием является отдельной частью речи. Данный случай легко обрабатывается, но есть и более сложный: в pymorphy2 выделяется часть речи предикатив. В русском языке предикативом может выступать как наречие, так и другая часть речи. То есть невозможно однозначно отобразить часть речи слова ни в ту, ни в другую сторону. Так как наречие чаще всего выступает предикативом, то именно оно и было взято в качестве отображения. В связи с этим наборы всех инструментов были проанализированы и отображены на множество, использующееся в корпусе. Полный список отображений можно увидеть в таблице 1.

Таблица 1

Отображение наборов частей речи инструментов на множество частей речи СинТагРус

Часть речи	**СинТагРус**	**TreeTagger**	**MyStem**	**TnT**	**pymorphy2**	**FreeLing**
гл.	V	V	V	V	VERB, PRTF, PRTS, GRND, INFN	V, Q
сущ.	S	N, P, Y	S, SPRO	N, P, Y	NOUN, NPRO LATN	N, E, R
прил.	A	A	A, A-NUM, A-PRO	A	ADJF, ADJS, COMP	A
нар.	ADV	R	ADV, ADVPRO	R	ADVB, PRED	D, P
числ.	NUM	M	NUM	M	NUMB, NUMR, ROMN	Z, Y
предл.	PR	S	PR	S	PREP	B
союз	CONJ	C	CONJ	C	CONJ	C
част.	PART	Q	PART	Q	PRCL	T
межд.	INTJ	I	INTJ	I	INTJ	J

Обработка результатов MyStem повлияла на оценку инструментов. Это связано с тем, что для слов, разбитых на несколько токенов, части речи нельзя конкатенировать, поэтому соответствующие слова не рассматривались. После сокращения набор для проверки состоял из 57 025 предложений и 767 058 слов. Статистику по частям речи можно увидеть в таблице 2.

Таблица 2

Статистика корпуса после обработки

Часть речи	Количество слов	Процент от всех слов
V	116 011	15.12
S	305 849	39.87
A	113 925	14.85
ADV	44 928	5.86
NUM	12 776	1.67
PR	87 104	11.36
CONJ	51 227	6.68
PART	35 142	4.58
INTJ	96	0.01

Основные результаты представлены в таблице 3. При подсчете усреднений в расчет не были взяты результаты для междометий (INTJ), так как они сильно выделяются не в лучшую сторону у всех инструментов, а процент этой части речи слишком мал. Взвешенное усреднение не дает много информации, так как соотношение частей речи может измениться в зависимости от предметной области или корпуса. Из оставшихся восьми частей речи MyStem продемонстрировал лучшие показатели в шести. Хоть и из-за числительных (NUM) макро-усреднение получилось меньше некоторых остальных, все же его можно поставить на первое место. Это подтверждают и другие усреднения.

Для компенсации плохих результатов на числительных было принято решение выбрать второй анализатор. Этим инструментом оказался pymorphy2 по следующим причинам:

а) Возможность применения к отдельным словам, так как алгоритм не зависит от контекста. Таким образом, если MyStem не может распознать часть речи, то достаточно вызвать один метод и получить ответ.

б) Лучшие показали по числительным и близкие к лучшим по основным частям речи: глаголы, существительные, прилагательные и наречия.

в) Лучшее макро-усреднение и второе место по другим.

Таблица 3

F1 мера инструментов морфологической разметки

Часть речи	**TreeTagger**	**MyStem**	**TnT**	**pymorphy2**	**FreeLing**	**MyStem + pymorphy2**
V	0.97	0.98	0.98	0.96	0.92	0.98 (=)
S	0.93	0.97	0.93	0.95	0.92	0.97 (=)
A	0.80	0.91	0.80	0.88	0.78	0.91 (=)
ADV	0.68	0.84	0.70	0.76	0.80	0.84 (=)
NUM	0.80	0.43	0.82	0.90	0.39	0.91 (+0.48)
PR	1.00	1.00	1.00	0.99	0.99	1.00 (=)
CONJ	0.90	0.91	0.90	0.84	0.85	0.91 (=)
PART	0.80	0.80	0.81	0.76	0.79	0.80 (=)
INTJ	0.53	0.53	0.54	0.10	0.65	0.53 (=)
Усреднение						
макро	0.86	0.86	0.87	0.88	0.81	0.92 (+0.06)
взвешенное	0.90	0.94	0.90	0.92	0.88	0.95 (+0.01)
микро	0.90	0.94	0.91	0.92	0.89	0.95 (+0.01)

Результаты применения комбинации MyStem и pymorphy2 можно увидеть в последнем столбце таблицы 3. Междометия снова не рассматривались для сравнения. В последнем столбце приведена разница с использованием только MyStem. Видно, что результаты для числительных сильно возросли, что повлияло на усреднения. Таким образом, при анализе текста комбинация этих двух инструментов дает хороший результат.

Заключение и выводы.

В рамках данной работы были исследованы инструменты, решающие задачу морфологического анализа. Для адекватного анализа были изучены морфологически размеченные корпуса со снятой омонимией. Из них был выбран корпус СинТагРус, содержащий после обработки почти 800 тысяч слов. Из пяти выбранных инструментов наилучшие результаты показал MyStem с мерой F1 равной 94%. Также был исследован вопрос улучшения этого показателя с помощью комбинации других анализаторов. Полученные

результаты могут быть полезны всем, кто занимается обработкой текста на естественном языке.

Литература:

1. Alexandra Blazhievskaya, Elizaveta Kuzmenko, Elmira Mustakimova, Timofey Arkhangelskiy et al. Morphological Analysis for Russian: Integration and Comparison of Taggers // Analysis of Images, Social Networks and Texts. – Cham: Springer, 2016. – V. 661. – P. 162-171.
2. Dereza O. V., Kayutenko D. A., Fenogenova A. S. Automatic Morphological Analysis for Russian: a Comparative Study // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – 2016.
3. Ilya Segalovich. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. – Las Vegas, 2003. – P. 273-280.
4. Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. – Manchester, UK, 1994. – P. 44-49.
5. Зализняк А.А. Грамматический словарь русского языка. — М., Русский язык, 1980. – 880 с.
6. Thorsten Brants. TnT: a statistical part-of-speech tagger // Proceedings of the sixth conference on Applied natural language processing. – Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. – P. 224-231.
7. Tools for processing Russian [Электронный ресурс]. URL: http://corpus.leeds.ac.uk/mocky (дата обращения 02.11.2017).
8. Mikhail Korobov. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. – Cham: Springer, 2015. – V. 542. – P. 320-332.

9. Грановский Д.В., Бочаров В.В., Бичинева С.В. Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской объединенной конференции «Интернет и современное общество». – СПб., 2010. – С. 94.

10. Lluis Padro, Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality // Proceedings of the Eight International Conference on Language Resources and Evaluation. – Istanbul, Turkey: European Language Resources Association, 2012. – P. 2473-2479.

11. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003-2005. – Москва: Индрик, 2005. – С. 193-214.

12. Leonid Iomdin, Vadim Petrochenkov, Victor Sizov, Leonid Tsinman. ETAP parser: state of the art // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue». – Moscow: RSUH, 2012. – Issue 11(18). – V. 2. – P. 119-131.

13. Сокирко А.В. Морфологические модули на сайте www.aot.ru // Труды международной конференции «Диалог-2004. Компьютерная лингвистика и интеллектуальные технологии». – Москва: Наука, 2004. – С. 559.

14. Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П. и др. Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». – Москва: РГГУ, 2013. – Вып. 12 (19). – Т. 1. – С. 84-95.

15. Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A. et al. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies // Computational Linguistics and Intellectual Technologies:

Proceedings of the International Conference «Dialogue». – Moscow: RSUH, 2012. – Issue 11(18). – V. 2. – P. 91-103.

Abstract

The article presents a comparison of five morphological analyzers: TreeTagger, MyStem, TnT, pymorphy2 and FreeLing. To evaluate their quality, three text corpora were considered: SynTagRus, OpenCorpora and GICR. The selected corpus was processed, and the grammeme sets were unified accordingly before running the analyzers. The best result reached 94% of the F1 measure.

Key words: natural language processing, automatic morphological analysis, taggers, corpora, disambiguation.

Статья отправлена: 04.11.2017 г.