УДК 81’33

Оськина К. А.

АЛГОРИТМ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ СПИСКА ЛЕММ И СЛОВОФОРМ ДЛЯ ПРЕДМЕТНОЙ ОБЛАСТИ ФУНДАМЕНТАЛЬНОГО И ПРИКЛАДНОГО РЕЧЕВЕДЕНИЯ[1]

В статье рассматривается проблема обработки недетерминированных слов и составление алгоритма, в ходе которого генерируется совокупность словоформ для слов, отсутствующих в словаре. Описываемый алгоритм рассматривается применительно к русскому языку. Анализируются современные подходы к морфологическому анализу, а также достоинства и недостатки современных инструментов морфологической разметки текстов на русском языке.
Ключевые слова: *предредактирование, лемматизация, морфологический анализ, настройка на предметную область.*

Оськіна К. А. Алгоритм автоматичної генерації списку лем і словоформ для предметної галузі фундаментального та прикладного мовленнєзнавства. – Стаття.
У статті розглядається проблема оброблення недетермінованих слів і складання алгоритму, під час якого генерується сукупність словоформ для слів, відсутніх у словнику. Описаний алгоритм розглядається стосовно російської мови. Аналізуються сучасні підходи до морфологічного аналізу, а також переваги й недоліки сучасних інструментів морфологічної розмітки текстів російською мовою.
Ключові слова: *передредагування, лематизація, морфологічний аналіз, налаштування на предметну галузь.*

Oskina K. A. Algorithm of automated lemma and wordform list generation for subject domain of fundamental and applied speech studies. – Article.
This article deals with the problem non-deterministic words processing for Russian. An algorithm of generating a list of lemmas and wordforms for such words is considered. Modern approaches to morphological analysis are investigated. The advantages and disadvantages of modern tools of Russian texts POS-tagging are reviewed as well.
Key words: *automatic pre-editing, lemmatization, morphological analysis, domain adaptation.*

Введение. Существующие решения в области семантического анализа достаточно эффективны при их использовании в хорошо формализованных предметных областях, то есть в случае наличия исчерпывающих словарей. Однако в случае, когда модуль семантического анализа сталкивается с недетерминированной лексикой, например, неологизмами или заимствованиями, качество результатов заметно снижается. Описанный в работе алгоритм лемматизации положен в основу подсистемы предредактирования в перспективном модуле семантического анализа, разрабатываемого на кафедре прикладной и экспериментальной лингвистики Московского государственного лингвистического университета.

Системы машинного перевода (далее – СМП) начали создаваться в 1960-х гг., однако большинство проблем, с которыми сталкивались их разработчики, так и не были решены. К основным проблемам машинного перевода относится нехватка достаточного количества словарей, обработка идиом, грамматическая и семантическая неоднозначность, проблемы определения интерлингвы. Основной трудностью при разработке релевантной системы является возможность адаптации СМП к особенностям предметной области [14, с. 161].

В настоящее время на кафедре прикладной и экспериментальной лингвистики Московского государственного лингвистического университета под руководством д-ра филол. наук, проф. Р.К. Потаповой проводятся исследования, направленные на разработку алгоритма автоматического определения значений слов с настройкой на предметную область (далее – ПО) фундаментального и прикладного речеведения. Основными проблемами лексического пласта вышеуказанной ПО является наличие омонимов и полисемов вследствие заимствования терминов из других предметных областей, а также большое количество неологизмов, что приводит к трудностям при обработке тематических текстов методами, основанными на словарях.

В целях формального описания предметной области необходимо сформировать список базовых форм и соответствующих им словоформ. Построение лемматизатора для вышеуказанной предметной области позволяет решить ряд проблем, связанных с обработкой и тэгированием объемных текстовых массивов, в частности, проблемы сортировки и систематизации текстовых массивов, сегментации текстов, общелингвистического поверхностного анализа, или аннотирования, текстов, внутренней разметки (расстановка морфологических, синтаксических и семантических обозначений) [11, с. 95].

Современные подходы к морфологическому анализу. В современных модулях морфологического анализа принято выделять два подхода к организации словарей лексики языка: *лемматизация* и *стемминг*. *Лемматизацией* называют подход, при котором в главном словаре анализатора хранятся леммы – основные формы слов с указанием основы. Ему противопоставляется *стем-*

[1] Проект поддержан Российским научным фондом (РНФ), проект №14-18-01059. Науч. рук. – д-р филол. наук, проф., действительный член Международной академии информатизации Потапова Р.К.

© К. А. Оськина, 2017

минг – подход без использования словаря основ. В стемминге есть только правила обрабатывания суффиксов и небольшие словари исключений [9, с. 21]. Стемминг используется в тех случаях, когда морфология не важна.

Оба метода имеют свои достоинства и недостатки. Для стеммингового метода характерна высокая скорость анализа за счет упрощения алгоритма и уменьшения объема выдаваемой информации; при отсутствии словаря основ по факту становится доступной морфологическая база неограниченного объема, настраиваемая непосредственно на имеющийся текст, что является практичным при создании информационно-поисковых систем с нефиксированной лексикой. Однако стеммингновый метод характеризуется невысокой точностью, невозможностью морфологического синтеза на базе без основ, возможностью порождения одинаковых стемов для различных слов («люб» для «люб-овь» и «люб-ить»), смешением различных понятий (к «люб-ить» будет отнесен глагол «люб-оваться»). Также стеммингновый метод не справляется с обработкой чередования гласных (для слова «идти» стемы «ид», «шл», «ше») [4, с. 119]

Лемматизаторы, напротив, характеризуются высокой точностью выдаваемых результатов. Наряду с этим они справляются с супплетивизмом и чередованием. К минусам лемматизаторов относится то, что совокупность словоформ, полученных в результате их работы, занимает больше памяти относительно данных, хранимых при обработке стеммингновыми методами. В то же время с учетом современного уровня развития вычислительных мощностей этой разницей можно пренебречь. Помимо этого, лемматизаторы не справляются с такими проблемами, как омонимия и полисемия, однако в пределах единой предметной области данный недостаток не является критичным. Так, из-за распространения в русском языке таких явлений, как супплетивизм и чередования [9, с. 21], наиболее оптимальным для его обработки будет использование лемматизатора.

Предморфологический анализ. Первым шагом для разработки лемматизатора является сбор корпуса по фундаментальному и прикладному речеведению. На этом этапе был сформирован неаннотированный корпус лингвистических текстов общим объемом 144 тыс. слов. В корпус вошли статьи из журналов «Вестник Московского государственного лингвистического университета» [5], журнал «Речевые технологии» [7], сборник «Компьютерная лингвистика и интеллектуальные технологии» [6], а также книга Р.К. Потаповой «Речь: коммуникация, информация, кибернетика» [12]. Статьи были собраны произвольно с 2010 по 2016 гг.

Далее необходимо было отобрать словник для специализированного терминологического словаря. Критерием для включения специальной лексики в собираемый словник являлся факт отсутствия слова в списке лемм и словоформ для русского языка. Был написан сценарий на языке программирования Perl [18], который в автоматическом режиме отсеивал из собранного корпуса термины, уже имеющиеся в вышеуказанном списке (рис. 1). Каждая словоформа анализируемого текста последовательно сравнивалась со словоформой из перечня словоформ русского языка. Совпавшие словоформы исключались из текста, остальные выводились в файл, который обрабатывался специалистом вручную. Так был сформирован базис для словника терминологических единиц в вышеуказанной области, который впоследствии может быть дополнен терминами из других текстов и статей в зависимости от анализируемой предметной области.

Общий список выделенных специализированных терминов составил 739 слов. Список лемм и словоформ русского языка был взят из источника [2].

Рис. 1. Отрывок из книги Р.К. Потаповой «Речь: коммуникация, информация, кибернетика» после обработки

Частеречная разметка. На следующем этапе необходимо было определить, к каким частям речи относятся слова из полученного списка словоформ. Перед проведением второго этапа были рассмотрены основные проблемы, связанные с морфологическим анализом, а также были апробированы некоторые инструменты, позволяющие производить частеречную разметку в автоматическом режиме:

1) phpmorphy – представляет собой морфологический анализатор для русского языка, написанный на скриптовом языке PHP, на данном этапе поддерживает и перечень других языков. Существует возможность анализа неизвестных языков со словарями ispell (рис. 2) и АОТ (рис. 3). Анализатор позволяет решать задачи лемматизации, получения

грамматической информации для слова, а также изменять форму слова с заданными параметрами. В его основу был положен проект АОТ [1], в частности, алгоритм и базовые словари;

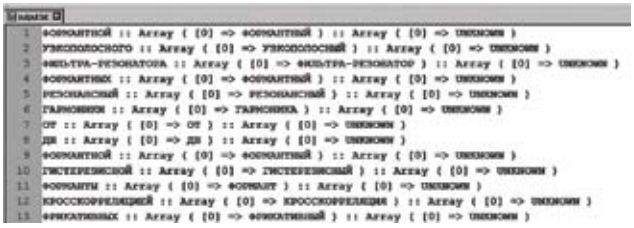

Рис. 2. Результат анализа словоформ со словарем ispell

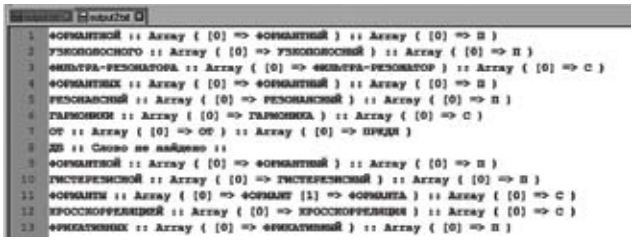

Рис. 3. Результат анализа словоформ со словарем АОТ

2) pymorphy2 – представляет собой морфологический анализатор, написанный на высокоуровневом языке программирования Python. В перечень его возможностей входят: приведение слова к форме леммы, составление грамматической информации о слове и создание нужной формы слова. Для анализа используется словарь OpenCorpora [10], а для слов, не входящих в состав словарей, строятся гипотезы [8]. Результат обработки вышеуказанного фрагмента текста представлен на рис. 4;

Рис. 4. Результат работы анализатора pymorphy2

3) mystem – также является морфологическим анализатором русского языка. Разработчики позиционируют его как парсер, частично снимающий морфологическую неоднозначность. В основе анализатора лежат словари, а также возможность формирования гипотезы о неизвестных словах. Технология имеет закрытый исходный код, но обладает хорошей доступностью для некоторых языков программирования высокого уровня [13].

Рис. 5. Результаты работы морфологического анализатора mystem

Основные проблемы частеречных теггеров:

1) языкозависимость – большинство коммерческих анализаторов работают лишь с английским языком (в частности, NLTK);

2) отсутствие алгоритма анализа недетерминированных слов;

3) отсутствие правил на анализ слов с дефисом;

4) невозможность корректного анализа аббревиатур.

В результате проведенных исследований было установлено, что анализатор mystem выдает результат с наименьшим количеством ошибок и погрешностей, поэтому было принято решение о его использовании.

Таким образом, на данном этапе были выявлены части речи выделенных на предыдущем этапе слов, а также их основная форма.

Генерация списка лемм и словоформ. На следующем этапе был использован стеммер Snowball [17], переписанный под выделение псевдоокончаний анализируемых слов (к примеру, у термина «кросскорреляция» было выделено псевдоокончание «-ция»). Фрагмент доработки отображен на рис. 6.

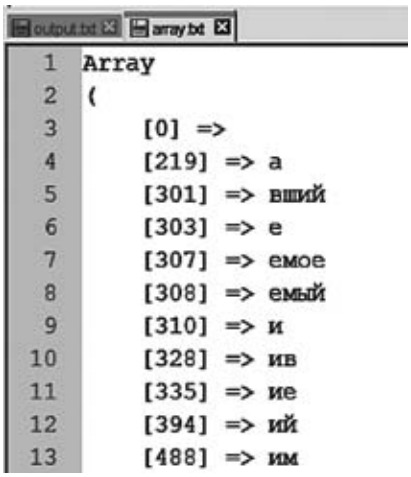

Рис. 6. Фрагмент результата работы стеммера Портера после доработки программы

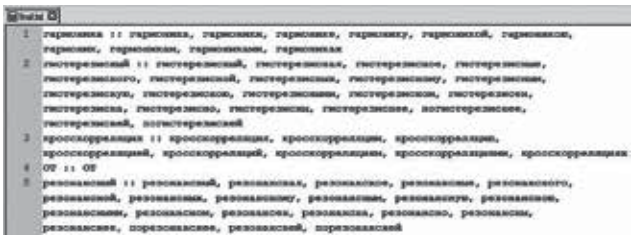

Рис. 7. Результат работы лемматизатора

После этого были сформированы правила для формирования списка лемм и словоформ для русского языка, которые были применены к псевдоосновам слов, выделенным на предыдущем этапе (например, с термином «кросскорреляция»: псевдооснова «кросскорреля-» и псевдоокончание «-ция»). Для терминов этого типа были выделены окончания, которые при добавлении к псевдоосновe образовывали словоформы исходного термина: псевдооснова + -ция, -ции, -цию, -цией и т. д.

Результат работы программы отображен на рис. 7 (слова с нулевым псевдоокончанием дописывались вручную).

Выводы. В данной статье был рассмотрен алгоритм для решения задачи автоматической генерации списка лемм и словоформ для отсутствующих в заранее заданном словаре слов по предметной области фундаментального и прикладного речеведения. Данный алгоритм может быть настроен на любую предметную область с учетом ее специфики и особенностей входящей в нее лексики. В ходе разработки алгоритма был выявлен принцип формирования словника для предметной области, были проанализированы некоторые инструменты морфологического анализа, выявлены проблемы, связанные с частеречным тэгированием, разработан способ автоматической генерации словоформ по лемме. Планируется разработка алгоритма для обработки слов с нулевым псевдоокончанием.

Литература

1. Автоматическая обработка текста [Электронный ресурс]. – Режим доступа : http://www.aot.ru/.
2. Архивы форума «Говорим по-русски» [Электронный ресурс]. – Режим доступа : http://www.speakrus.ru/dict/.
3. Библиотека морфологического анализа phpMorphy [Электронный ресурс]. – Режим доступа : http://phpmorphy.sourceforge.net/dokuwiki/.
4. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : [учеб. пособие] / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова.– М. : МИЭМ, 2011. – 272 с.
5. Вестник МГЛУ [Электронный ресурс]. – Режим доступа : http://www.linguanet.ru/science/informatsionno-bibliotechnyy-tsentr/vestnik-mglu.php.
6. Журнал «Компьютерная лингвистика и информационные технологии» [Электронный ресурс]. – Режим доступа : http://www.dialog-21.ru/digest/.
7. Журнал «Речевые технологии» [Электронный ресурс]. – Режим доступа : http://speechtechnology.ru/.
8. Морфологический анализатор phpmorphy2 [Электронный ресурс]. – Режим доступа : http://pymorphy2.readthedocs.io/.
9. Николаев И.С. Прикладная и компьютерная лингвистика / И.С. Николаев, О.В. Митренина, Т.М. Ландо. – М. : ЛЕНАНД, 2016. – 320 с.
10. Открытый корпус [Электронный ресурс]. – Режим доступа : http://opencorpora.org/.
11. Потапова Р.К. Основные тенденции развития многоязычной корпусной лингвистики (Часть 1). – Речевые технологии, 2 / Р.К. Потапова. – И. : «Народное образование», 2009. – С. 92–114.
12. Потапова Р.К. Речь: коммуникация, информация, кибернетика / Р.К. Потапова. – М. : Либроком, 2010. – 600 с.
13. Технология Mystem [Электронный ресурс]. – Режим доступа : https://tech.yandex.ru/mystem/.
14. Okpor M.D. Machine Translation Approaches: Issues and Challenges. – IJCSI International Journal of Computer Science Issues / M.D. Okpor. – Vol. 11(5), No. 2. – M. : SoftwareFirst Ltd, 2014. – Pp. 159–165.
15. Potapova R. On Individual Polyinformativity of Speech and Voice Regarding Speakers Auditive Attribution (Forensic Phonetic Aspect). In: Ronzhin A., Potapova R., Nemeth G. Lecture Notes in Artificial Intelligence, 9811. – H.: Springer, 2016. – Pp. 507–514.
16. Potapova R. Attribution of Social Network Discource. In: Ronzhin A., Potapova R., Nemeth G. Lecture Notes in Artificial Intelligence, 9811. – H. : Springer, 2016. – Pp. 539–546.
17. Snowball [Электронный ресурс]. – Режим доступа : http://snowball.tartarus.org/.
18. The Perl Programming Language [Электронный ресурс]. – Режим доступа : https://www.perl.org/.