УДК 004.021

Е.А. Тихомирова

МИНИМИЗАЦИЯ ОШИБОК ИДЕНТИФИКАЦИИ ЛЕКСЕМ В ТЕКСТАХ, НАПИСАННЫХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Представлен метод автоматизированной минимизации ошибок идентификации лексем в текстах, написанных на естественном языке. Метод основан на совместном использовании словарной морфологии и аналитического метода. Применение метода позволит улучшить качество автороведческой экспертизы, что повысит вероятность идентификации автора по написанным им текстам.

E-mail: elizarti@mail.ru

Ключевые слова: *частотный метод, автороведческая экспертиза, словоформа, лексема, морфологический анализ.*

Введение. Широко используемый в задачах идентификации текстов метод частотных словарей предполагает подсчет частоты употребления того или иного слова в тексте. При попытке автоматизации этого процесса исследователи сталкиваются с трудностями идентификации лексем, связанных с множеством словоформ одного и того же слова. При подобном подсчете текст необходимо подвергнуть первоначальной обработке: привести весь текст к нижнему регистру и заменить буквы «ё» на «е», так как часто в печатных (и не только) текстах буква «ё» заменяется «е» иногда и самим автором. Затем требуется определить частоту употребления слов, например, с помощью:

1) методов подсчета частоты употребления словоформ;

2) методов словарной морфологии;

3) аналитических методов анализа словоформ.

Несмотря на то, что перечисленные методы используются исследователями при составлении частотных словарей, они обладают недостаточно высокой способностью идентификации лексем. Так, алгоритм метода подсчета частоты употребления словоформ (самый простой, наименее эффективный и малоиспользуемый алгоритм) не предполагает какого-либо морфологического анализа текста, а просто определяет число появления одних и тех же словоформ, идентичных по написанию. Главный недостаток такого метода заключается в том, что слова, являющиеся двумя словоформами одной лексемы, представляют собой различные лексемы для этого вида анализа (например, «дом» и «дома»). Для иллюстрации работы метода подсчета был взят отрывок из повести Н.В. Гоголя «Старосветские помещики». Частота употребления некоторых встречающихся в этом отрывке слов, подсчитанная вручную, приведена в табл. 1.

Таблица 1

Частота употребления некоторых слов, встречающихся в отрывке повести Н.В. Гоголя «Старосветские помещики», подсчитанная вручную

Слово	Частота употребления	Слово	Частота употребления
Говорить	5	Он	11
Обед	4	Пирожок	3
Обедать	1	Попробовать	2
Обыкновение	1	Рыжик	3
Обыкновенный	3	Что	6

Результат работы метода подсчета частоты употребления словоформ представлен в табл. 2 (метод плохо справляется с идентификацией лексем).

Таблица 2

Результат работы метода подсчета частоты употребления словоформ

Слово	Частота употребления	Слово	Частота употребления
Говорил	4	Он	3
Говорила	1	Пирожками	1
Обеда	2	Пирожков	2
Обедать	1	Попробуем	1
Обедом	1	Попробуйте	1
Обеду	1	Рыжиками	1
Обыкновению	1	Рыжиков	2
Обыкновенно	3	Что	3

Методы словарной морфологии основаны на применении словарей для определения лексемы по словоформе, встречающейся в тексте. Неоспоримое достоинство метода — безошибочность, если словоформа содержится в словаре. Однако если словоформа отсутствует по какой-либо причине, то алгоритм метода не может определить лексему. В связи с этим для уменьшения ошибок идентификации слов была сделана попытка работы с несколькими словарями, что потребовало создания баз данных этих словарей. На основе словаря А.А. Зализняка создана база данных, пример таблицы которой представлен на рис. 1 [1].

id	word	base
6235	его	он
6236	ее	он
6237	его	он
6238	ему	он
6239	ей	он
6240	ему	он
6241	его	он
6242	ее	он
6243	его	он
6244	его	он
6245	ее	он
6246	его	он
6247	ей	он
6248	ею	он

Рис. 1. Часть таблицы «Е» базы данных на основе словаря Зализняка:
id — уникальный идентификатор поля; word — словоформа; base — лексема данной словоформы

База данных состоит из 29 таблиц. Таблицы названы по буквам алфавита русского языка с учетом того, что слов, начинающихся на буквы «ь», «ъ» и «ы» в русском языке нет, а буквы «е» и «ё» были приведены к «е» и находятся в одной таблице «Е». Слова располагаются в таблицах по условию совпадения названия таблицы и первой буквы словоформы (см. рис. 1).

Для расширения возможностей идентификации был взят словарь ПроЛинг, который в отличие от словаря Зализняка содержит много аббревиатур, собственных имен, составных слов, причастий [2]. Таким образом, используя указанный словарь совместно со словарем Зализняка, можно получить большее число распознанных словоформ.

С помощью такого метода осуществляется запрос к базам данных и определяются лексемы по найденной в тексте словоформе. Результат работы метода приведен в табл. 3.

Таблица 3

Результат работы метода, основанного на использовании словарей Зализняка и ПроЛинг

Слово	Частота употребления	Слово	Частота употребления
Говорил	4	Он	11
Говорила	1	Пирожковый	2
Обед	4	Пирожок	1
Обедать	1	Попробовать	2
Обыкновение	1	Рыжик	3
Обыкновенный	3	Что	6

В результате определяются лексемы, которые на самом деле отсутствуют в тексте, например, слово «пирожковый» (см. табл. 3). Это является следствием омонимии. Для устранения указанного эффекта использован частотный словарь русского языка С.А. Шарова [3]. Таким образом, при появлении омонимов выбирается та лексема, которая наиболее часто употребляется в русском языке. В рассматриваемом случае лексема «пирожок» после использования словаря Шарова встречается в тексте три раза, а лексема «пирожковый» отсутствует.

Как было отмечено выше, существенным недостатком метода является то, что словарная морфология работает корректно только тогда, когда словоформа находится в словаре. В противном случае лексема не распознается. В связи с этим необходимо упомянуть методы,

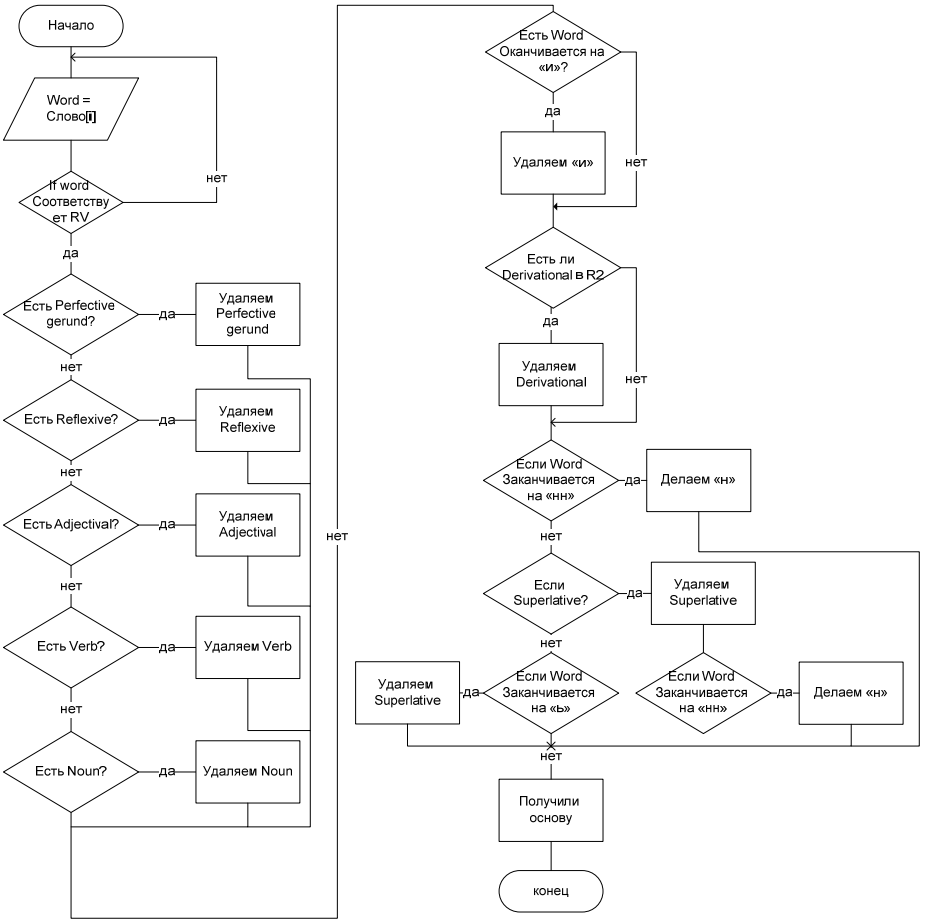

Рис. 2. Алгоритм Snowball:

RV — область, выделяемая в слове после первой гласной или являющаяся окончанием слова, если в нем отсутствуют гласные; Perfective gerund, Adjectival, Participle, Reflexive, Verb, Noun, Superlative, Derivational — группы окончаний, соответствующие определенным частям речи; Word — слово, основу которого необходимо определить

не зависящие от словарей — аналитические методы анализа словоформ. Суть таких методов — применение алгоритма выделения основы слова. В качестве примера был взят алгоритм Snowball (рис. 2) [4]. Результат работы алгоритма представлен в табл. 4.

Таблица 4

Результат работы алгоритма Snowball

Основа	Частота употребления	Основа	Частота употребления
Говор	5	Пирожкам	1
Обед	4	Попроб	1
Обеда	1	Попробу	1
Обыкновен	4	Рыжик	2
Он	6	Рыжикам	1
Пирожк	2	Что	3

После анализа работы данного алгоритма предложено учесть следующее.

Особенность работы регулярных выражений в базовом функционале C#: машина определяет совпадение с регулярным выражением по минимальному условию. Поэтому необходимо разделить группы окончаний не только по частям речи, но и по числу букв, составляющих данное окончание. Проверку следует начинать с максимально длинных окончаний. В противном случае, словоформы «генералам» и «генералами» будут иметь разные основы.

Особенность русского языка заключается в том, что словообразование осуществляется с помощью не одной морфемы, а нескольких. Тогда требуется несколько проходов по алгоритму одной и той же словоформы.

Результаты работы модифицированного алгоритма Snowball представлены в табл. 5.

Таблица 5

Результат работы модифицированного алгоритма Snowball

Основа	Частота употребления	Основа	Частота употребления
Говор	5	Пирожк	3
Обед	5	Попроб	2
Обыкнов	4	Рыжик	3
Он	6	Что	3

Достоинство аналитического метода — независимость анализа от словаря и его объема. Однако существует и недостаток — при работе наблюдается процент ошибок при определении принадлежности словоформ к одной и той же лексема (ошибки первого и второго рода).

Комплексный метод. С учетом изложенного выше, можно сделать вывод о том, что описанные методы имеют свои достоинства и недостатки, которые смогут в значительной степени компенсировать друг друга. Поэтому в статье предложен комплексный метод, обобщающий достоинства перечисленных методов. Алгоритм работы комплексного метода представлен на рис. 3.

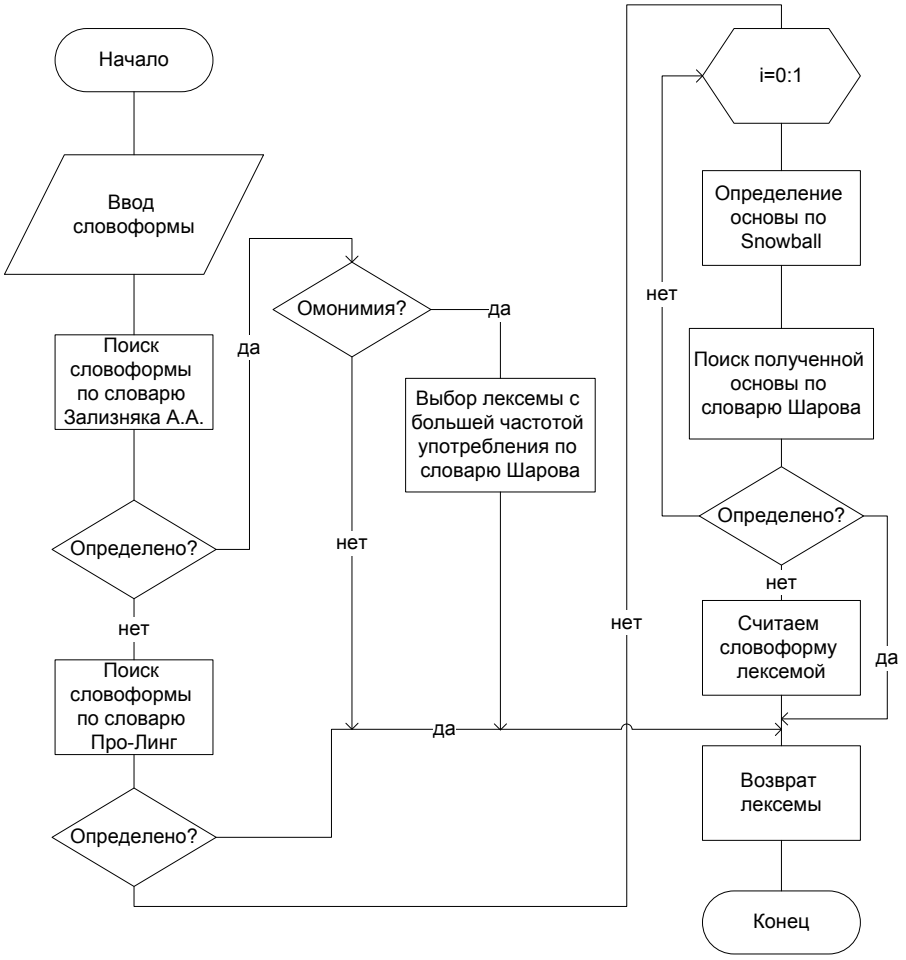

Рис. 3. Алгоритм работы комплексного метода

Суть алгоритма — последовательный поиск словоформы в словарях Зализняка и ПроЛинг. Если в этих словарях словоформа отсутствует, то к ней применяется аналитический метод выделения основы с помощью алгоритма Snowball:

1) Выделяется основа словоформы алгоритмом Snowball.

2) Происходит поиск данной основы в словаре Шарова [3]. Если основа найдена в словаре Шарова, то возвращается лексема, которой соответствует данная основа. Если основа не была найдена, то шаг 1

повторяется из соображения, что словообразование в русском языке в среднем осуществляется за счет двух морфем.

3) Если после прохождения словоформы по данному алгоритму два раза лексема не будет определена, то словоформа признается лексемой и возвращается, как результат работы алгоритма.

Словарь Шарова был взят для определения лексемы на шаге 3 работы общего алгоритма, так как этот словарь содержит частоты употребления лексем русского языка. Алгоритм поиска лексем с одинаковыми основами по словарю Зализняка дает ошибки первого и второго рода. При определении основы словоформы необходимо минимизировать процент подобных ошибок, что осуществляет путем поиска совпадения в словаре, отсортированном по частоте употребления лексем в русском языке. Таким образом, вероятность корректного выявления лексемы по словоформе повышается.

Для корректной работы алгоритма была создана база данных, основанная на словаре Шарова, пример таблицы которой представлен на рис. 4.

id	base	quantity
1	еще	2786,93
2	если	1974,42
3	есть	1436,68
4	ехать	216,14
5	единственный	203,95
6	едва	191,22
7	естественно	95,73
8	еда	93,65
9	еврей	93,41
10	ездить	89,74
11	европа	82,27
12	единый	76,15
13	еле	65,25
14	естественный	58,09

Рис. 4. Часть таблицы «Е» базы данных на основе словаря Шарова:
id — уникальный идентификатор; base — лексема; quantity — частота употребления лексемы, измеренная в ipm (instances per million words — количество вхождений на миллион слов)

Результат работы алгоритма комплексного метода представлен в табл. 6, соответствия словоформ, встреченных в тексте, лексемам, определенных на шаге 3 работы алгоритма, т. е. найденных аналитически — в табл. 7.

Таблица 6

Результат работы алгоритма комплексного метода

Слово	Частота употребления	Слово	Частота употреблений
Говорить	5	Он	11
Обед	4	Пирожок	3
Обедать	1	Попробовать	2
Обыкновение	1	Рыжик	3
Обыкновенный	3	Что	6

Таблица 7

Результат работы алгоритма комплексного метода на шаге 3

Слово	Лексема	Слово	Лексема
Напившись	Напиться	Говаривал	Говаривать
Говорил	Говорить	Подставляя	Подставлять
Расспрашивал	Расспрашивать	Приносила	Приносить
Сообщал	Сообщать	Говорила	Говорить
Приблизившись	Приблизиться	Принимая	Принимать
Заедал	Заедать		

Для сравнения в табл. 8 приведен результат работы алгоритма на шаге 3 с использованием словарей Зализняка и ПроЛинг вместо словаря Шарова.

Таблица 8

Результат работы комплексного алгоритма на шаге 3 с использованием словарей Зализняка и ПроЛинг

Слово	Лексема, определенная по словарю	
	Зализняка	ПроЛинг
Напившись	Напаять	Напавший
Говорил	Говор	Говор
Приблизившись	Приблизительность	Приблизивший
Приносила	Принос	Принос
Говорила	Говор	Говор

При детальном рассмотрении результатов, представленных в табл. 7 и 8, можно сделать вывод, что использование словаря Шарова является наиболее целесообразным для минимизации ошибок идентификации лексем аналитическим методом.

Заключение. Рассмотренный в статье алгоритм дает наилучший результат при автоматической идентификации лексем в текстах, написанных на русском языке.

СПИСОК ЛИТЕРАТУРЫ

1. Архив форума «Говорим по-русски» // http://speakrus.ru / URL: http://speakrus.ru/dict/#paradigma (Дата обращения 13.04.2012).
2. Архив форума «Говорим по-русски» // http://speakrus.ru / URL: http://speakrus.ru/dict/#proling Дата обращения 20.04.2012.
3. Шаров С.А. Частотный словарь русского языка [Электронный ресурс]. URL: http://www.artint.ru/projects/frqlist.asp (Дата обращения 8.09.2011).
4. Snowball // http://snowball.tartarus.org/ URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html Дата обращения 11.05.2011.

Статья поступила в редакцию 4.07.2012