УДК 004.056

МЕТОД ПОЛУАВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ СЛОВАРЯ МОРФОЛОГИЧЕСКИХ ОПИСАНИЙ СЛОВ

С.В. Лапшин, И.С. Лебедев

Предложен метод полуавтоматического формирования словаря морфологических описаний слов. Предложенный метод позволяет существенно упростить процедуру пополнения и актуализации словарей новыми словоформами и повысить таким образом показатели точности и полноты морфологических анализаторов DLP- и IPC-систем.

Ключевые слова: полуавтоматическое формирование словаря, морфологические описания слов, DLP- и IPC-системы.

Введение

Усиление влияния информационной сферы деятельности на жизненно важные интересы общества и государства требует ее защиты от информационных воздействий. Это обусловливает возрастание роли и значения обработки и мониторинга текстовой информации в компьютерных сетях и открытых системах документооборота.

Одной из ключевых частей современных систем защиты информации классов DLP (Data loss prevention) и IPC (Information Protection and Control) является морфологический анализатор. В своей работе он использует словарь, который содержит морфологические описания словоформ того или иного языка. При работе с аналитическими языками, например, с французским, задача морфологического анализа не столь актуальна, поскольку основными передатчиками грамматического значения являются порядок слов и другие служебные слова. Но при обработке текстов на синтетическом языке, например, на русском, морфологическое описание слова является основой для построения естественно-языковой конструкции. Таким образом, возникает задача генерации словарей, которые содержат эту информацию.

Текстовые сообщения, циркулирующие в вычислительных сетях, обрабатываемые с целью мониторинга состояния информационной безопасности, имеют ряд особенностей, среди которых необходимо отметить небольшую длину и использование специфических выражений и аббревиатур [1]. Примером могут являться сообщения в интернет-мессенджерах или социальных сетях. Поскольку качество обработки текстов естественного языка напрямую зависит от полноты используемых для этого словарей, очень важно иметь достаточно полные и актуальные словари с необходимыми описаниями для каждой словоформы.

Ручное пополнение словаря новыми словами с их морфологическими описаниями является трудоемким процессом. То, что естественный язык не является статичным, особенно в разговорной речи, еще более усложняет задачу поддержания актуальности таких словарей.

Работы в этом направлении ведутся уже достаточно давно. Большой вклад в рассматриваемый вопрос внес коллектив Санкт-Петербургского экономико-математического института (ЭМИ РАН) [2], а также компании АОТ, Noolab, RCO и др. [3]. В настоящей работе рассматриваются вопросы адаптации морфологического анализа для систем мониторинга, обрабатывающих текстовую информацию. В связи со спецификой анализируемых текстов в таком словаре должны быть не только корректные словоформы, но и словоформы с типичными ошибками, которые допускаются людьми при написании текстов, а также выражения и аббревиатуры, специфичные для конкретной группы людей или области знаний.

Предлагаемый в работе метод позволяет существенно упростить процедуру пополнения и актуализации словарей новыми словоформами и, таким образом, повысить показатели точности и полноты морфологических анализаторов DLP- и IPC-систем.

Постановка задачи

Основой предлагаемого предметно-ориентированного морфологического анализатора, содержащего идентификационные признаки словоформ предметной области, разработанного для русского языка, служит словарь А.А. Зализняка [4].

Формализация морфологии в словарных базах данных программы представлена следующим образом.

Пусть $S = \{ S_i \}$, i=1,..., n – множество исходных форм слов базы данных средства защиты информации (БД СЗИ). Пусть $M = \{M_j\}$, j=1,..., k – множество парадигм, причем каждому элементу множества соответствует морфологический признак $M_j \rightarrow P_j$. Пусть s – словоформа. Пусть $c = \{ c_r \}$, r=1,..., z – множество стандартных окончаний слов.

Тогда необходимо найти такие функции f и g, что

$$S \xrightarrow{f} M; \qquad (1)$$

$$M \xrightarrow{g} S,$$

где f – функция соответствия элементов множества $\boldsymbol{S}$ элементам множества M; g – функция соответствия элементов множества $\boldsymbol{M}$ элементам множества S. Нахождение этих функций позволяет говорить о решении прямой и обратной задач морфологии для СЗИ.

Предлагаемый метод решения основан на том, что любой словоформе сопоставим класс основ B и класс окончаний C, из которого состоит данная словоформа:

$\forall s \longrightarrow \{ B; C \}$.

Тогда для каждого слова БД СЗИ можно выделить морфологический класс k его парадигм M_k, такой, что словоформа данного морфологического класса S_k является подмножеством парадигм этого класса, и выражается суммой основ и окончаний слова данного морфологического класса.

f_k: $s_k \in M_k = B_k + C_k$.

Это означает, что частная задача нахождения функции f прямой задачи морфологии решена. Совокупность решения частных задач даст решение прямой задачи в общем виде.

Соответствие $M_j \rightarrow P_j$, позволяет получить морфологический и идентификационный признак, содержащий информацию, используемую для обнаружения угроз информационной безопасности (морфологический шаблон).

Решение обратной задачи предполагает существование некоторой словоформы БД СЗИ.

Пусть $s_k \subset M_k$. Определим ее основу как разность между словоформой и ее окончанием:

$\{ B_k \} = s_k - \{ c_r \}_k$, k=1,..., n, r=1,..., z.

Сопоставим эту основу с множеством допустимых для нее исходных словоформ:

$B_k \longrightarrow S_k$, k=1,..., n.

Используя прямую задачу, вычислим множество парадигм этих основ:

$\{ S_k \} \xrightarrow{f} \{ M_k \}$.

Сравним исходную словоформу с этими парадигмами. В случае совпадения определяется исходная форма слова для данной парадигмы:

g_k: $M_k \longrightarrow S_k$.

Совокупность решения частных задач даст решение обратной задачи в общем виде для СЗИ, обрабатывающих текстовую информацию.

Метод формирования словаря

Рассмотрим словарь $Z = \{ z_i \}$, полученный путем «чтения» предметно-ориентированных текстов новостных агенств, блогов и комментариев, каждая запись z_i в котором имеет структуру

$z_i = \{ s_i; S_i; P_i \}$,

т.е. состоит из словоформы s_i, исходной формы слова S_i и морфологического описания P_i.

Словарь Z на сегодняшний день содержит более 2,5 млн словоформ. Особенностью словаря является то, что в нем содержатся последовательности символов, употребляемые внутри групп пользователей, и словоформы, имеющие специфические опечатки. Задача состоит в том, чтобы, учитывая регулярность русского языка [5], поддерживать актуальность и полноту словарной базы данных в условиях его постоянного пополнения новыми словоформами с наименьшими трудозатратами. Решение поставленной задачи основывается на словаре, содержащем морфологические описания словоформ А.А. Зализняка, куда входят только базовые словоформы русского языка и множество соответствующих им окончаний.

Рассмотрим, как образованы словоформы слов «ПРЕОБРАЗОВАТЕЛЬ» и «РОЯЛЬ» (таблица).

Словоформы M_j		Морфологические описания P_j
ПРЕОБРАЗОВАТЕЛЬ	РОЯЛЬ	Сущв Муж Неодуш Им, Вин
ПРЕОБРАЗОВАТЕЛЯ	РОЯЛЯ	Сущв Муж Неодуш Род
ПРЕОБРАЗОВАТЕЛЮ	РОЯЛЮ	Сущв Муж Неодуш Дат
ПРЕОБРАЗОВАТЕЛЕМ	РОЯЛЕМ	Сущв Муж Неодуш Тв
ПРЕОБРАЗОВАТЕЛЕ	РОЯЛЕ	Сущв Муж Неодуш Пред
ПРЕОБРАЗОВАТЕЛИ	РОЯЛИ	Сущв Муж Неодуш Им, Вин
ПРЕОБРАЗОВАТЕЛЕЙ	РОЯЛЕЙ	Сущв Муж Неодуш Род
ПРЕОБРАЗОВАТЕЛЯМ	РОЯЛЯМ	Сущв Муж Неодуш Дат
ПРЕОБРАЗОВАТЕЛЯМИ	РОЯЛЯМИ	Сущв Муж Неодуш Тв
ПРЕОБРАЗОВАТЕЛЯХ	РОЯЛЯХ	Сущв Муж Неодуш Пред

Таблица. Словоформы и их морфологические описания

Из таблицы видно, что словоформы получены из базовой формы S одинаковым образом, путем добавления соответствующих окончаний C. Следовательно, достаточно иметь морфологические описания P словоформ слова «ПРЕОБРАЗОВАТЕЛЬ», чтобы построить аналогичные описания для словоформ слова «РОЯЛЬ».

На основе этой идеи разработан предлагаемый метод полуавтоматического формирования словаря. Он состоит из следующих частей:

1. разбор словаря Зализняка, генерация всех словоформ на основе исходных форм слова;
2. разбор словаря с некоторыми морфологическими описаниями описанного выше вида;
3. сопоставление словоформ из словарей, полученных на первых двух шагах, с целью выделения характерных морфологических описаний для каждого окончания;
4. на основе множества соответствий вида окончание–морфологическое описание, полученных на предыдущем шаге, словоформам из словаря Зализняка дается морфологическое описание.

Задачи на первых двух шагах являются чисто техническими, и их описание не представляет какого-либо интереса.

Выделение характерных морфологических описаний для каждого окончания, описанное на третьем шаге, осуществляется следующим образом. Каждое окончание входит в свой «класс» окончаний. Для слова «ПРЕОБРАЗОВАТЕЛЬ» это окончания «Я», «Ю», «ЕМ», «Е», «И», «ЕЙ», «ЯМ», «ЯМИ» и «ЯХ». Окончание «Ю» слова «ЗЕМЛЮ», хотя и совпадает с окончанием «Ю» слова «ПРЕОБРАЗОВАТЕЛЮ», но входит в совершенно другой «класс», и поэтому будет иметь другой набор морфологических описаний. Кроме класса, также учитывается часть речи слова и одушевленность/неодушевленность для имен существительных.

Таким образом, полный ключ в ассоциативном массиве с морфологическими описаниями состоит из «класса» окончания, части речи и признака одушевленности/неодушевленности для имен существительных. Таким образом, в случае, когда одна исходная форма относится к разным частям речи, для каждой части будет храниться свой набор морфологических описателей.

Полученные на третьем шаге соответствия применяются на четвертом шаге для составления словаря с морфологическими описаниями. Из сгенерированного на основе словаря Зализняка списка словоформ берется словоформа, а затем по ключу, описанному выше, в ассоциативном массиве находится морфологическое описание для этой словоформы. Таким образом, реализуется функция f, приведенная в формуле (1).

За счет того, что русский язык достаточно регулярен и многие слова формируются похожим способом, появляется возможность автоматически получить для исходной формы слова морфологические описание всех его словоформ.

Важной особенностью представленной системы является то, что она не определяет автоматически класс добавляемого слова, а предполагает, что он уже задан. Задача классификации нового слова отведена специалисту-лингвисту либо может быть решена вспомогательными средствами, например, описанными в [6].

Таким образом, процедура пополнения словаря новыми словами с их морфологическими описаниями сводится к определению «морфологического класса» исходной формы слова.

Эксперимент

Для оценки качества метода введем следующие обозначения: количество правильных извлечений системы анализа DLP-фильтра – h, количество требуемых извлечений – d, общее количество извлечений – n. Тогда для полноты R и точности P справедливы следующие соотношения:

$$R_i = \frac{h_i}{d_i} \text{ и } P_i = \frac{h_i}{n_i}.$$

Эксперимент по поиску с использованием словарей проводился на основе случайной выборки предложений из национального корпуса русского языка [7]. Объем выборки – 180 000 словоупотреблений, из которых 90 000 взяты из прессы и по 30 000 – из научных текстов, художественных текстов и законодательства.

Для проведения эксперимента была разработана простая поисковая система на основе булевской модели поиска [8]. Разработанная система позволяла автоматически формировать поисковые запросы и обрабатывать результаты поиска. Таким образом, значение d числа требуемых извлечений было известно при формировании поисковых запросов, что обеспечивало правильность полученного результата. Общее количество извлечений p и количество правильных извлечений h вычислялись в ходе эксперимента после обработки каждого поискового запроса.

В первом случае поисковая система использовала словарь Зализняка и словарь с полными морфологическими описаниями для только одного слова каждого класса. Во втором случае использовался словарь, сгенерированный с помощью описанного выше метода.

В ходе эксперимента измерялись полнота R и точность P поиска на случайной выборке из национального корпуса русского языка. Результаты измерения приведены на рис. 1, 2.

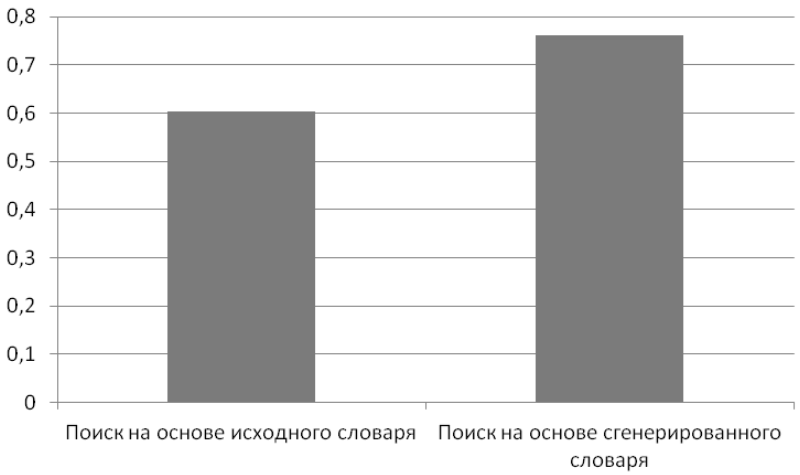

Рис. 1. Результаты измерения точности поиска *P*

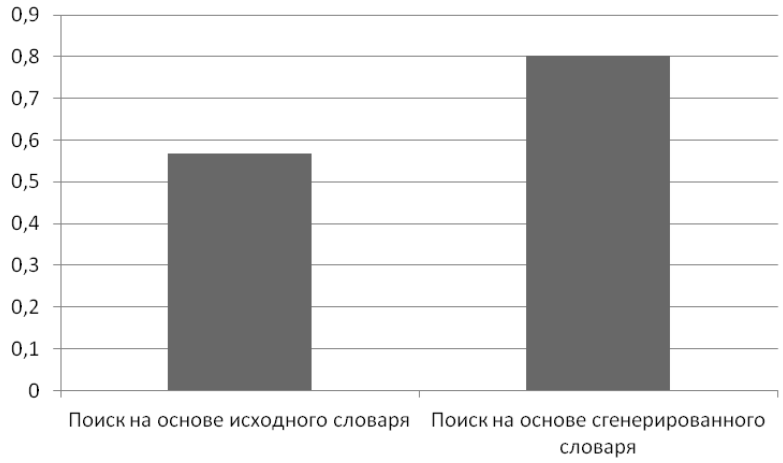

Рис. 2. Результаты измерения полноты поиска *R*

Эксперимент показал, что при использовании сгенерированного описанным выше методом словаря точность поиска возросла на 26%, а полнота – на 42%.

Заключение

Предложенный метод полуавтоматического формирования словаря морфологических описаний слов позволяет существенно упростить и ускорить процедуру получения таких словарей. Человеку достаточно задать класс добавляемого слова, после этого все словоформы автоматически получат морфологические описания. Это позволяет быстро адаптировать систему анализа естественного языка и, таким образом, приводит к повышению показателей точности и полноты морфологических анализаторов DLP- и IPC-систем.

Работа выполнена в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы» по государственному контракту № 07.524.12.4009.

Литература

1. Лебедев И.С., Борисов Ю.Б. Анализ текстовых сообщений в системах мониторинга информационной безопасности // Информационно-управляющие системы. – 2011. – № 2. – С. 37–43.
2. Каневский Е.А. Некоторые вопросы пополнения морфологического словаря терминами предметной области // Труды Международного семинара «Диалог'2001» по компьютерной лингвистике и ее приложениям. – М.: РосНИИ искусственного интеллекта, 2001. – Т. 2. – С. 156–160.
3. Большаков И.А., Большакова Е.И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии. По материалам конференции «Диалог» (2012). – Т. 1. – Вып. 11. – С. 81–92.

4. Зализняк А.А. Грамматический словарь русского языка. – М.: Русский язык, 1987. – Изд. 4-е, испр. и доп. – 880 с.
5. Тузов В.А. Компьютерная семантика русского языка. – СПб: Изд-во СПбГУ, 2004. – 400 с.
6. Боярский К.К., Каневский Е.А. Проблемы пополнения семантического словаря // Научно-технический вестник СПбГУ ИТМО. – 2011. – № 2 (72). – С. 132–137.
7. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа: http://ruscorpora.ru/corpora-usage.html, свободный. Яз. рус. (дата обращения 30.05.2012).
8. Manning C.D., Raghavan P., Schutze H. An Introduction to Information Retrieval. – Cambridge University Press, Cambridge, England. – 2009. – 504 p.

Лапшин Сергей Владимирович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, аспирант, sv.lapshin@gmail.com

Лебедев Илья Сергеевич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кандидат технических наук, доцент, lebedev@cit.ifmo.ru