АНАЛИЗ СТРУКТУРНЫХ ЗАКОНОМЕРНОСТЕЙ
(Вычислительные системы)
2005 год Выпуск 174

УДК 519.769 : 801.314.4

КОМБИНИРОВАННЫЙ АЛГОРИТМ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ДЛЯ НОРМАЛИЗАЦИИ НЕИЗВЕСТНЫХ СИСТЕМЕ СЛОВ [1]

Н.В. Саломатина

Введение

Процедура морфологического анализа применяется в системах обработки текстов самых разных типов: в корректорах — для обнаружения орфографических ошибок, в поисковиках — для автоматического создания словарей и унификации запросов, в системах машинного перевода — для идентификации иноязычного эквивалента и других, где требуется анализ и синтез текстов.

Нормализация или приведение слова к каноническому виду осуществляется на основе грамматических характеристик, полученных на этапе морфологического анализа. Целью ее является устранение морфологической вариативности слов, которая может служить помехой для дальнейшего исследования текстов. При работе в автоматическом режиме морфологический анализ, а также и нормализацию, требуется проводить для всех слов текста, в том числе не содержащихся в словаре системы. Известные подходы к решению этой проблемы похожи в части применения процедуры поиска аналогов. В случае обнаружения

[1] Работа выполнена при финансовой поддержке РФФИ (проект №03-06-80118).

системой "нового" слова целесообразным считается нормализовать его и назначить ему грамматическую информацию по аналогии, исходя из того, что слова, имеющие сходный буквенный состав концов, аналогичны и по грамматическим характеристикам [1]. По утверждению авторов, этот принцип работает правильно в 90 % случаев.

Однако, в разных подходах аналогами "новому" слову служат элементы разных языковых уровней: например, в [1] — словоформы, в [2, 3] — совокупность квазиморфем. Но основной трудностью нормализации по аналогии для всех подходов остается выбор аналога из нескольких претендентов, различающихся, по крайней мере, типом словоизменения, т.е. правилами восстановления канонической формы. Определение аналога только по максимальному совпадению буквенного состава концов слов иногда уводит от выявления правильной формы, например, в случае, когда анализируются имена собственные и/или имеет место вложенность слов.

В случае анализа текстов без словаря словоформ, как в [4], аналог "новому" слову ищется с учетом максимально совпадающей последовательности суффиксов, список которых (с набором грамматических помет) имеется в системе. Достоверность назначения определенной грамматической информации (например, частеречного значения, а часто и более существенной — числа, падежа и т.п.) зависит от того, насколько верно проведено разбиение слова на окончание и основу, в которой будет проводиться поиск суффиксных последовательностей. Чтобы выяснить без словаря, правильно ли выделена основа, используют информацию о других вхождениях слова в текст. Высокая степень надежности правильного разбиения слова обеспечивается в случае, если в тексте присутствуют хотя бы три разных формы слова.

Обработка "нового" слова в системе морфологического анализа, снабженной словарем, аналогична работе анализатора без словаря. Поэтому разумно было бы дополнить систему анализа по словарю модулем, работающим с "новыми" словами примерно так же, как в [4], когда объектом исследования является текст.

Цель данной работы состоит в том, чтобы дать описание комбинированного алгоритма морфологического анализа, рассчитанного на нормализацию "новых" слов и использующего при назначении морфологической информации по аналогии также и текстовую — о словоупотреблении.

1. Нормализация "новых" слов по аналогии

Грамматические формы, в которых слова употребляются в текстах, будем называть текстовыми формами и обозначать TF. Словарный эквивалент текстовой формы принято называть канонической формой (CF). Для основных частей речи это означает следующее: существительные, прилагательные, причастия, местоимения и числительные представлены в словаре в именительном падеже единственного числа или множественного, если единственного нет, глаголы — в неопределенной форме. В процессе работы алгоритма нормализации текстовые формы преобразуются в канонические.

Основным ресурсом для проведения процедуры нормализации служит словарь словоформ, в котором для каждого слова в CF есть все его грамматические формы. Назовем их словарными и обозначим DF. Словарь содержит примерно 3,2 млн. DF, из них более 100 тыс. являются каноническими. В оперативной памяти компьютера словарь представлен в виде бинарного дерева [5], построенного по инвертированным (с обратным порядком букв) DF, посимвольно записанным в вершинах. Поступающая на вход дерева (в корневую вершину) TF также предварительно инвертируется. При поиске TF в дереве выполняется проход по тем вершинам, в которых символы DF совпадают с символами анализируемой TF. Если TF имеется в словаре, то проход по дереву заканчивается в листе ссылкой на каноническую форму, соответствующую TF. Иначе, алгоритм прекращает поиск после исчерпания TF или после проверки всевозможных разветвлений из последней совпавшей вершины.

Возникает ситуация, когда в тексте обнаружено "новое" слово, которое нужно нормализовать. Обход поддеревьев, исходящих из вершины, где было последнее совпадение символа DF с символом TF, позволяет выявить всех претендентов-аналогов

для восстановления канонической формы из искомой текстовой. Число исходящих поддеревьев может оказаться равным единице, тогда обход поддерева заканчивается ссылкой на один лист (одну CF). Например, такая ситуация возникла при нормализации фамилии героя сказки "Винни-Пух" Кристофера Робина. В качестве наилучшего аналога для TF "Робина" было выбрано максимально совпадающее с ним слово "дробина". Здесь имеет место вложенность анализируемого слова в его аналог. Восстановление канонической формы для "Робина" по образцу "дробина" приводит, очевидно, к ошибочной нормализации. Чтобы избежать возможных ошибок такого типа, можно расширить поиск, т.е. в качестве аналогов рассматривать слова, совпадающие с "новым" по меньшему числу символов. Для этого необходимо из вершины, в которой было зафиксировано последнее совпадение символов DF и TF, подниматься вверх по дереву на все более высокие уровни до тех пор, пока число поддеревьев в вершине этого уровня не окажется больше единицы. При этом число общих букв в DF и TF сократится на количество пройденных вверх уровней. Для слова "Робина" достаточно оказалось подъема на один уровень, чтобы число поддеревьев стало больше единицы. При их обходе нашелся подходящий аналог — "глобина" с CF "глобин", склоняющийся по той же словоизменительной схеме, что и "Робин". Ниже приведены примеры словарных аналогов "новым" словам из текстов разных жанров и результаты восстановления CF из текстовых по аналогии.

1. Алан А. Милн "Винни-Пух" (перевод Б. Заходера)

TF = шумелка

аналоги из словаря:

1) DF=выбелка, $CF(DF)$=выбелка, $CF(TF)$=шумелка;
2) DF=брелка, $CF(DF)$=брелок, $CF(TF)$=шумелок;

2. Текст, составленный по ленте компьютерных новостей

TF = слот

аналоги из словаря:

1) DF=бейшлот, $CF(DF)$=бейшлот, $CF(TF)$=слот;
2) DF=болот, $CF(DF)$=болото, $CF(TF)$=слото;
3) DF=позолот, $CF(DF)$=позолота, $CF(TF)$=слота;
4) DF=исколот, $CF(DF)$=исколотый, $CF(TF)$=слотый;

3. Труды конференции "Диалог-2002"

TF = дисплея

аналоги из словаря:

1) DF=аллея, $CF(DF)$=аллея, $CF(TF)$=дисплея;
2) DF=троллея, $CF(DF)$=троллей, $CF(TF)$=дисплей;
3) DF=дряхлея, $CF(DF)$=дряхлеть, $CF(TF)$=дисплеть;
4) DF=клея, $CF(DF)$=клеить, $CF(TF)$=дисплеить.

Из примеров видно, что аналогов с одинаковым количеством совпавших букв может быть несколько, и все они имеют разные модели словоизменения (соответственно, нормализации). Но среди них, как правило, есть подходящий для верного восстановления CF из TF. Интегральные характеристики неоднозначности восстановления канонических форм, полученные при анализе разных текстов, представлены в табл.2 раздела 3.

При расширении поиска следует ввести ограничение снизу на число совпавших в DF и TF букв, так как иначе в качестве аналога может быть выбрано мало подходящее слово. Кроме того, при переходе в дереве даже на один уровень вверх число поддеревьев может сильно увеличиться, а обход большого количества поддеревьев, иногда содержащих десятки и сотни тысяч словоформ, сильно замедляет работу алгоритма морфологического анализа (на порядок и больше). Например, для словоформы "Хемингуэя" с нетипичным для русского языка сочетанием конечных букв аналогами являются все слова, оканчивающиеся на "я", а их в словаре около 1,8 млн. Чтобы ограничить число аналогов, полагаем, что, по крайней мере, два конечных символа основы TF и DF должны совпадать (основа TF определяется путем отсечения окончания, равного окончанию DF). Для регулирования быстродействия алгоритма это ограничение можно варьировать, пока не будет достигнут компромисс между

требуемым быстродействием и удовлетворительным качеством восстановления CF. Надо заметить, что найденным аналогам, особенно при расширенном поиске, часто сопутствует одна и та же грамматическая информация, т.е. способ нормализации у этих аналогов один и тот же. Поэтому, при запоминании слов-претендентов, участвующих в дальнейшем анализе, необходимо также проводить проверку на уникальность сопутствующей им грамматической информации.

Когда аналог и грамматическая информация найдены, становится известен тип словоизменения TF. Схемы или модели восстановления канонической формы по типу словоизменения являются обращением правил развертывания форм слова в парадигму [6]. Они дифференцируются по частям речи, а внутри частеречных зон - по словоизменительным индексам, отражающим тип склонения или спряжения. В соответствии с этими индексами и реализуется схема нормализации "нового" слова.

2. Использование текстовой информации для выбора аналога

Задача повышения надежности восстановления CF по аналогии может быть решена разными способами. Например, в [7] предлагается создать в системе морфологического анализа дополнительный список слов, которые были нормализованы неправильно. Экспериментальная проверка этого подхода показала, что надежность нормализации, равная 99 %, достигается, когда объем дополнительного словаря составляет примерно 11 тыс. слов. Такой подход явно "утяжеляет" алгоритм по памяти и быстродействию, требует непрерывной ручной доработки и непредсказуем в плане дальнейшего роста дополнительного словаря. Если анализируется текст, возможен другой подход, не требующий дополнительного ресурса. В этом случае для повышения надежности нормализации "нового" слова используются содержащиеся в тексте сведения о разнообразии форм этого слова.

Поиск каждой текстовой формы "нового" слова в словаре системы позволяет получить определенный набор слов-аналогов.

После обработки всего текста будем иметь столько наборов аналогов, сколько форм "нового" слова в нем встретилось. Каждый аналог нормализуется по своей уникальной схеме восстановления канонической формы. По этой схеме предполагается нормализовать и анализируемую текстовую форму. И если какая-то схема восстановления чаще других встречается во всех наборах аналогов, то она и заслуживает доверия. Другими словами, чем из большего числа форм текста восстановлена каноническая, реализующая одну и ту же схему нормализации, тем больший вес она имеет среди претендентов. В то же время, частота встречаемости формы в тексте не оказывает влияния на вес.

Для реализации данного подхода в процедуре морфологического анализа необходимо ввести дополнительную структуру данных для учета разнообразия текстовых форм всех "новых" слов. В процессе работы с этой структурой требуется отождествлять разные текстовые формы "нового" слова. Впрямую это сделать сложно, т.к. на этапе образования форм часто варьируются не только окончания, но и сама основа. Поэтому был использован прием косвенного учета форм "новых" слов: все TF считаются принадлежащими к одной парадигме, если восстановленные из них CF совпадают. В ситуации, когда из TF может быть восстановлено несколько разных CF, считаем форму текста принадлежащей к каждой из этих парадигм. Проблема выбора CF часто может быть разрешена с помощью учета всех TF, встретившихся в тексте.

Таким образом, для каждой TF_i, $i = 1, ..., \mathrm{M}$ (M — суммарное число различных форм всех "новых" слов в тексте), в системном словаре морфологического анализатора ищутся аналоги CF_i^l ($l = 1, ..., n_i$, — номер аналога, а n_i — число аналогов i-й формы). Согласно каждому l-му словоизменительному индексу однозначно восстанавливается каноническая форма "нового" слова CF_i^l. Слова, являющиеся омоформами, т.е. одинаковыми по написанию формами, образованными из разных CF, к рассмотрению в качестве аналогов не принимаются. Поскольку проводится расширенный поиск аналогов, такое отбрасывание претендентов никак не сказывается на полноте их списка для дальнейшего анализа.

В результате работы модифицированной процедуры морфологического анализа получаем два списка: первый — A(i) — содержит различные текстовые формы TF_i, $i = 1, ..., M$. Второй — B(j) — различные канонические формы $\overline{CF}(j)$, $j = 1, ..., N$, восстановленные из всех TF_i текста. Совокупность $\overline{CF}(j)$ отличается от $\bigcup_{i,l}(CF_i^l)$ тем, что в ней все канонические формы представлены однократно: $\overline{CF}(j) = \bigcup_{i \neq j}((CF_i \bigcup CF_j) \backslash (CF_i \bigcap CF_j))$, где $CF_i = \bigcup_l CF_i^l$, $CF_j = \bigcup_l CF_j^l$, соответственно $N = \mid \overline{CF}(j) \mid$.

Чтобы сохранить связь между текстовой и канонической формами при формировании списков A и B создается третий — $C(i,l)$, $l = 1, ..., \max_i(n_i)$, в котором для каждой TF_i запоминаются номера j всех ее CF_i^l в списке B. На табл. 1 представлен результат анализа текстовой формы "Робина" (а вместе с ней и всех других производных от "Робин", встретившихся в тексте) модифицированной процедурой. Словарными аналогами TF = "Робина" послужили словоформы DF^1 = "дробина" и DF^2 = "глобина" с CF^1 = "дробина" и CF^2 = "глобин" соответственно. Омонимия конечных символов других форм этих слов с другими формами "Робина", встретившимися в тексте, высока: Робине — дробине — глобине, Робину-- дробину — глобину, Робин — дробин — глобин. Если бы в тексте не встретилась форма "Робином", для которой есть аналог у слова "глобин", а у слова "дробина" такого аналога нет, то выбор CF был бы затруднителен.

Из табл. 1 видно, что элементы списка B состоят из двух компонентов. Первый компонент — это j-ая каноническая форма, как уже говорилось выше; второй компонент — число, характеризующее вес ($W(j)$) этой CF, равный количеству ссылок на нее по всем i в списке C, или количеству форм TF_i, из которых она восстановлена.

Вычисление весов $W(j)$ канонических форм проводится следующим образом:

1. Положим $i := 0$, $k := 0$ и $W(j) := 0$ для всех j.

Таблица 1

Результат обработки текстовых форм, производных от TF = "Робин", комбинированной процедурой морфологического анализа

A:		C:					B:		
			l						
i	TF	i	1	2	...	n_i	j	$\overline{CF}$	W
1	Робин	1	1	2	...		1	Робин	5
2	Робина	2	1	2	...		2	Робина	4
3	Робину	3	1	2	...				
4	Робине	4	1	2					
5	Робином	5	1						
:	:	:	:	:	:	:	:	:	
M		M					N		

2. Если анализируемая TF является "новой", просматриваем список A, чтобы установить, встречалась такая TF или нет. Если она уже имеется в списке A, то переходим к рассмотрению следующей текстовой формы (частота встречаемости TF не влияет на W), т.е. повторяем п.2 до исчерпания текста. Если TF в A нет, переходим к п. 3.

3. Положим $i := i + 1$, внесем TF_i в список A: $A(i) := TF_i$. Ищем слова-аналоги для TF_i в словаре системы, а восстановленные согласно их морфологическому индексу CF_i^l $(l = 1, ..., n_i)$ ищем в списке B. Если CF_i^l нет в B, то переходим к п. 4, если есть, то к п. 5.

4. Записываем CF_i^l в B: $j := j + 1$; $\overline{CF}(j) = B(j, 1) := CF_i^l$, а $W(j) = B(j, 2) := 1$. Заполняем соответствующий TF элемент C (вносим номер j CF_i^l в списке B): устанавливаем k равным номеру первой незаполненной ячейки $C(i, *)$ (символ "$*$" означает, что просматриваются все ссылки i-ой формы), $C(i, k) := j$. Переходим к п. 2.

5. Если CF_i^l уже есть в B, увеличиваем ее вес на единицу: $W(j) = B(j, 2) := B(j, 2) + 1$, заносим в соответствующий TF элемент C порядковый номер j канонической формы в B:

устанавливаем k равным номеру первой незаполненной ячейки $C(i,*)$, $C(i,k) := j$. Переходим к п. 2.

Последовательный просмотр списков A и B при выяснении, есть ли в них исследуемые TF и CF (п.2 и п.3), может быть оптимизирован множеством различных способов.

По окончании работы процедуры всем TF "новых" слов сопоставляется каноническая форма с максимальным весом W. Для этого предварительно каждая строка массива ссылок $C(i,*)$ упорядочивается согласно весам канонических форм так, чтобы в начале списка ссылок стояла ссылка на CF с наибольшим весом.

Хотя при подсчете весов не производится непосредственное отождествление текстовых форм TF_i, принадлежащих одной парадигме, опасность неправильного подсчета веса CF за счет того, что TF_i принадлежат разным парадигмам, возникает лишь в случае, когда парадигмы совпадают по всем формам, встречающимся в тексте. Такая ситуация является заведомо редкой. При апробации алгоритма на текстах довольно большого объема (до полумиллиона словоупотреблений) и различных по тематике она ни разу не была зафиксирована.

Если несколько восстановленных канонических форм имеют одинаковый вес, то нормализацию можно проводить по аналогии (выбирать форму с максимальным числом совпадающих конечных символов) или не производить вовсе. Тогда в дальнейшем анализе будет участвовать форма текста. Какой вариант предпочесть — определяется пользователями.

3. Результаты апробации алгоритма

Комбинированный алгоритм нормализации был апробирован на трех типах текстов: 1) художественных — книга Алана А. Милна "Винни-Пух" в переводе Б. Заходера ("З"); 2) новостных (лента компьютерных новостей — "Н"); 3) научных — труды конференции "Диалог-2002" ("Д"). В табл. 1 указаны количественные характеристики анализируемых текстов: объемы (V) в словоупотреблениях; сведения о количестве "новых" слов (M) без учета повторяемости в каждом из текстов и с учетом повторяемости (K) (в скобках — доля "новых" слов от объема

текста). Самый объемный текст — труды конференции — содержит 146 докладов на тему "Компьютерная лингвистика и интеллектуальные технологии"; в ленте компьютерных новостей содержатся короткие сообщения рекламного характера, состоящие из 200 — 300 слов.

В табл. 2 также приведены параметры, характеризующие нормализацию "новых" слов: L — максимальное количество аналогов, приходящихся на "новое" слово; S — среднее число аналогов на "новое" слово; A и AT — отражают долю правильно нормализованных "новых" слов (в процентах от K), соответственно, только по аналогии, и по аналогии с привлечением информации из текста.

Таблица 2

Количественные характеристики восстановления форм текста (TF) комбинированной процедурой морфологического анализа

	V	K	M	L	S	A	AT
"З"	39 806	1 347 (3,4%)	415	5	3,8	77 %	79%
"Д"	442 356	23 022 (5,2%)	9 766	5	2,5	86 %	89%
"Н"	73 029	2 008(2,7%)	1 116	5	3,7	84 %	93%

Из табл.2 видно, что доля "новых" слов невелика: 3-5%. Однако, в "Д" она несколько больше, чем в двух других текстах, а число слов-аналогов на "новое" слово меньше. Это вполне объяснимо, так как "Д" представляет собой совокупность научных докладов, содержащих много специальных терминов, тогда как в "З" и "Н" больше общеупотребительной лексики. Специальные термины часто имеют иноязычное происхождение, поэтому в среднем число слов-аналогов для них в системном словаре невелико. В то же время максимальное число аналогов на "новое" слово примерно одинаково в текстах всех типов в силу того, что их поиск велся по одному и тому же системному словарю. А их число ограничивалось одним и тем же условием: в DF и TF должны совпадать, по крайней мере, два конечных символа основ.

Надо заметить, что "новыми" словами чаще являются существительные. В силу существующей в языке омоформии аналогами для них могут быть как существительные, так и формы других частей речи, в том числе краткие прилагательные и глаголы. Тем не менее, из столбца AT следует, что в среднем по текстам около 87 % "новых" слов может быть правильно нормализовано с помощью предложенного способа учета текстовой информации. Число ошибок восстановления CF (см. столбец AT), полученное в результате работы модифицированной процедуры, несколько выше заявленного в [7]. Но корректным может быть лишь сравнение результатов обработки одного и того же материала. В столбце A таблицы приведены результаты восстановления CF только по аналогии. Из них видно, что на нашем экспериментальном материале этот метод работает с надежностью существенно меньшей 90 %. Сравнение столбцов A и AT дает представление о том, как использование текстовой информации увеличивает долю правильно нормализованных "новых" слов. На текстах "З" и "Д" это увеличение невелико: 2-3%. Самым трудным текстом для нормализации "новых" слов оказалась книга Милна. Ненормализованными в ней остались некоторые имена собственные ("Бразилию", "Африку", "Роберте", ...) и намеренно искаженные слова ("ушол", "победю" и т.п.). В этом случае найти близкий аналог в базовом словаре трудно, а информации о словоизменении в тексте практически нет, так как встречаются они однократно или в одной и той же форме. Это приводит к тому, что все CF, восстановленные по типу изменения слов-аналогов, имеют одинаковый вес, поэтому выбор одного из них затруднен. В текстах трудов конференции "Диалог'2002" трудности возникают также с именами собственными, а кроме того, с сокращениями, аббревиатурами. Приведение их к каноническому виду дает значительный процент ошибок. Например, сокращение "тел." (телефон) ошибочно трактуется как форма слова "тело". Очевидно, требуется специальная программа предобработки текста, которая позволит учесть специфику таких слов при морфологическом анализе. Лучшее качество восстановления CF достигнуто на материалах из ленты компьютерных новостей: использование текстовой информа-

ции дает существенный вклад в долю правильно нормализованных слов — 9 %. Такие отсутствующие в системном словаре слова как "драйвер", "чипсет", "процессор", "слот", "тайминг", "бренд", "бенгмарк" имеют аналоги, совпадающие с ними минимальным числом букв. И как следствие — аналоги имеют разные словоизменительные модели. Но в тексте эти "новые" слова функционируют (склоняются, спрягаются и т.п.) по правилам русского языка и встречаются в нескольких формах, их разнообразие позволяет выбрать аналог, согласно которому нормализация может быть осуществлена правильно. Очевидное преимущество предлагаемого подхода — отсутствие необходимости создавать дополнительный словарь. Однако это преимущество утрачивается, когда вместо текста морфологическому анализатору будут предлагаться отдельные слова.

Неоднозначность восстановления CF из текстовых форм возможна не только при анализе "новых" слов. Существующая в языке омоформия фиксируется в словаре системы: некоторым TF сопоставлены несколько вариантов CF. Доля омоформов в тексте выше, чем доля новых слов, и составляет порядка 5-7 % словоупотреблений. Часто разрешить ее можно с привлечением контекста, в том числе, используя комбинированный алгоритм морфологического анализа. Так, например, для TF = "рядом" в словаре системы в качестве CF предлагаются две формы: 1) $CF(1)$ = "рядом" — наречие и 2) $CF(2)$ = = "ряд" – существительное. Если в ближайшем контексте найдутся другие формы слова "ряд" ("ряду", "ряда", "ряды" и т.п.), вероятно, что и для TF = "рядом" канонической формой будет "ряд". Но возможность использования контекста для разрешения словарной омоформии в данной работе не исследовалась.

Алгоритм реализован на языке Delphi, быстродействие без нормализации новых слов — около 13000 словоформ в секунду, нормализация замедляет работу программы в зависимости от дополнительных ограничений, накладываемых при поиске слов-аналогов. В случае ограничения на число совпавших конечных символов основы, равное двум, происходит замедление примерно в 3,2 раза. Указанное в табл. 2 (столбец АТ) качество восста-

новления достигается со скоростью обработки порядка 4000 слов в секунду на Celeron Pentium 1700.

Заключение

Предложенный алгоритм нормализации "новых" (не содержащихся в системном словаре слов) эффективно работает в случае анализа текстов. Он объединяет в себе два подхода: а) нормализацию по аналогии, когда в качестве аналога выбирается слово с максимальным совпадением конечных символов основы аналога и "нового" слова, и б) использует информацию о разнообразии форм "нового" слова в исследуемом тексте. Это позволяет из нескольких аналогов "новому" слову, имеющихся в словаре системы, отобрать наиболее согласующийся с другими словоупотреблениями этого слова в тексте. Такой способ не требует лишних затрат на создание трудоемких ресурсов в дополнение к системному словарю с целью улучшения качества нормализации. Алгоритм также может быть использован для разрешения омоформии, присутствующей в статьях системного словаря.

Экспериментальная проверка алгоритма на разных по жанру и значительных по объему текстах показала, что предложенный алгоритм с хорошим качеством решает задачу нормализации "новых" слов и демонстрирует при этом приемлемое быстродействие.

Литература

1. Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем // М.: Наука, 1983. — 288 с.

2. Шереметьева С.О., Ниренбург С. М. Эмпирическое моделирование в вычислительной морфологии // НТИ, серия 2. — 1996. — №7. — С.28–33.

3. Перебейнос В.И., Грязнухина Т.А., Дарчук Н.П., Орлова Л.В. Морфологический анализ в автоматической системе исследования структурной организации реферативного текста // НТИ, серия 2. — 1989. — №6. — С.18–27.

4. Дудковский В.И. Автоматический морфологический анализ текстов без словаря // НТИ, серия 2. — 1990. — №2. — С.36–40.

5. Кнут Д. Искусство программирования для ЭВМ // Основные алгоритмы, — Т.1, М.: Мир, 1976. — 735 с.

6. Зализняк А.А Грамматический словарь русского языка — М.: Русский язык, 1977. — 879 с.

7. Белоногов Г.Г., Зеленков Ю.Г. Еще раз о принципе аналогии в морфологии // НТИ, серия 2. — 1995. — №3. — С.29–32.

Поступила в редакцию
21 апреля 2005 года