ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ И ПРОИЗВОДСТВЕННЫЙ ЖУРНАЛ

ВЕСТНИК КОМПЬЮТЕРНЫХ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

3(81)
2011

Издается с 2004 года

Журнал зарегистрирован в Министерстве РФ по делам печати, телерадиовещания и средств массовых коммуникаций. Свидетельство о регистрации ПИ № ФС77-36553 от 5 июня 2009 г.

СОДЕРЖАНИЕ

Журнал входит в Перечень изданий, утвержденных ВАК РФ для публикации трудов соискателей ученых степеней.

Журнал распространяется по подписке, которую можно оформить в любом почтовом отделении (индексы: по каталогу агентства "Роспечать"– **84197**, по Объединенному каталогу "Пресса России"– **39244**, по каталогу "Почта России"– **60263**) или непосредственно в редакции.
Тел.: 8(499) 268-69-19; 269-54-96; (495)514-76-50. Факс: 8(499) 269-48-97.
http://www.mashin.ru; www.vkit.ru. E-mail: vkit@mashin.ru, vkitpost@rambler.ru

УДК 004.85:81.32

Д.М. Черненьков, Э.С. Клышинский, *канд. техн. наук (Московский институт электроники и математики);*
e-mail: klyshinsky@itas.miem.edu.ru

Формальный метод пополнения словарей морфологического анализа с использованием несловарной лексики*

Разработан математический аппарат описания словаря системы морфологического анализа и синтеза. Предложен на его основе алгоритм кластеризации слов, отсутствующих в словаре. Построены гипотезы о парадигме изменения таких слов в целях пополнения словаря на основе полученных кластеров.

The description mathematical apparatus of the morphological analysis and synthesis system dictionary is developed. On its basis the clustering algorithm of the words missing in the dictionary is offered. Hypothesis about a paradigm of such words change for the dictionary addition on the basis of the received clusters are constructed.

Ключевые слова: *морфологический анализ и синтез; кластеризация; пополнение словарей.*

Keywords: *Morphological analysis and synthesis; Clustering; Dictionary replenishment.*

Введение

Важная часть большинства систем обработки текстов – подсистема морфологического анализа. Некоторые из подобных подсистем сочетают в себе сразу несколько функций [1–3]. Одной из них является порождение гипотез относительно нормальной формы и набора параметров незнакомого слова.

Однако системы словарного анализа требуют предварительного заполнения. Ручной ввод слов на данный момент представляет собой хорошо отработанную, но несколько долгую процедуру. Увеличение вычислительной мощности компьютеров позволило перейти к решению проблемы автоматического или автоматизированного заполнения словарей [4–6]. Подобная задача уже успешно решалась для систем *стемминга* (процесс нахождения основы слова для заданного исходного слова) или без учета полного набора параметров. Однако для систем *лемматизации* (процесс привода словоформы к лемме – ее нормальной (словарной) форме), работающих с флективными языками, например с русским, возникают серьезные проблемы, связанные с омонимией [7].

В статье рассматривается метод автоматизированного выделения несловарной лексики (лексики, отсутствующей в морфологическом словаре) и формирования на ее основе гипотез о парадигме ее изменения. Метод учитывает большое количество словоформ, принадлежащих слову, их омонимичность, в том числе в связи с разветвленной системой лексических параметров, и использует статистическую информацию о встречаемости словоформ для выбора корректной гипотезы. Сам вопрос применения метода снятия омонимии при работе с неизвестными словами уже исследовался, например в [6, 8, 9], однако для задач автоматизации процессов лемматизации несловарной лексики еще не применялся. Метод максимально независим от особенностей реализации и хранения данных в морфологическом словаре.

Математическая модель морфологического анализа

При проведении лексического анализа для слов естественного языка обычно выделяют нормальную форму и образованные от нее словоформы. При этом каждой словоформе приписывают часть речи и набор лексических параметров.

Для удобства введем множество $R=\{r\}$ частей речи, употребляемых в данном языке. Определим лексический параметр как пару $p=\langle n, \mathrm{v}\rangle$, где n –

* Работа поддержана Государственным контрактом в рамках Федеральной целевой программы "Научные и научно-педагогические кадры инновационной России" на 2009–2013 гг.

имя параметра, v – его значение, например <род, мужской>.

Для каждой части речи определяется множество возможных приписываемых ей имен параметров $N(r)$. Так, N (существительное) = {род, одушевленность, число, падеж}. Домен параметра, т.е. множество принимаемых им значений при фиксированном имени, будем обозначать как $V(n)$. Например, V (род) = {мужской, женский, средний, ∅}. Нулевое значение ∅ вводится для параметров, которые в определенных случаях не имеют грамматически осмысленного значения. Например, в русском языке у прилагательного во множественном числе с точки зрения грамматики нет рода. В этом случае будем считать, что параметр "род" принимает значение, равное ∅.

Для фиксированной части речи параметры можно разбить на *неизменяемые* и *изменяемые*, которые не изменяют или изменяют свои значения при смене формы слова соответственно. Тогда для каждой части речи можно ввести два множества имен параметров:

$N_{\text{const}}(r)$ – множество имен неизменяемых параметров;

$N_{\text{var}}(r)$ – множество имен изменяемых параметров.

Например, N_{const} (существительное) = {род, одушевленность}, а N_{var} (существительное) = {число, падеж}. Множество всех имен параметров для части речи обозначим как $N(r)$, причем $N(r) = N_{\text{const}}(r) \cup N_{\text{var}}(r)$.

Кроме того, введем множества собственно параметров части речи $P_{\text{const}}(r)$ – множество неизменяемых параметров для части речи r, и $P_{\text{var}}(r)$ – множество ее изменяемых параметров. При этом $p = <n, \text{v}> \in P_{\text{const}}(r)$, если $n \in N_{\text{const}}(r)$. Аналогично $p = <n, \text{v}> \in P_{\text{var}}(r)$, если $n \in N_{\text{var}}(r)$. Например, <род, мужской> ∈ P_{const} (существительное), а <падеж, именительный> ∈ P_{var} (существительное). Следует заметить, что параметр <род, мужской> принадлежит P_{var} (прилагательное) или P_{var} (причастие), т.е. отнесение параметров к тому или иному множеству зависит от части речи.

Для унификации строковую запись слова назовем *токеном*. Под словоформой будем понимать кортеж $f = <s, r, P_{\text{var}}(r, s)>$, где s – токен (написание словоформы), r – часть речи словоформы, $P_{\text{var}}(r, s) = \{p\}$ – множество изменяемых параметров, приписанных данной словоформе, причем $p \in P_{\text{var}}(r)$, при этом $|\{p\}| = |N_{\text{var}}(r)|$ и имена параметров различны. Иными словами в множестве заданы значения для всех изменяемых параметров данной части речи.

Все словоформы данного слова образуют лексему. На практике бывает удобно объединить в одну лексему словоформы, принадлежащие различным частям речи. Так, причастие и деепричастие удобно считать формами глагола. При этом словоформы, принадлежащие лексеме, разобьются на несколько подмножеств, объединенных одной частью речи. Исходя из этого, лексема задается кортежем

$$l = <f_{\text{nf}}, \{<P_{\text{const}}(r, s_{\text{nf}}), L(r)>\}>,$$

где $f_{\text{nf}} \in L(r)$ – словоформа, принимаемая за нормальную форму;

s_{nf} – токен, приписанный нормальной форме;

$P_{\text{const}}(r, s_{\text{nf}})$ – такое множество неизменяемых параметров данной лексемы при фиксированной части речи, что $P_{\text{const}}(r, s_{\text{nf}}) = \{p\}$, $p \in P_{\text{const}}(r)$, при этом $|\{p\}| = |N_{\text{const}}(r)|$ и имена параметров различны, т.е. в множестве представлены все неизменяемые параметры для данной части речи;

$L(r)$ – подмножество словоформ с одной частью речи:

$$L(r) = \{f_i\},\ \forall f_i = <s_i, r_i, P_{\text{var}}(r, s_i)>\ r_i = r.$$

Пример. Лексема "яблоко" задается следующим кортежем (для нормальной формы вместо словоформы указан токен):

```
<"яблоко", {< { <одушевленность, неодушевленное>, <род, средний>},
        {<"яблоко", существительное,
        {< падеж, именительный >,
< число, единственное >} >,
        <"яблока", существительное,
        {< падеж, родительный >,
< число, единственное >} >,
        <"яблоки", существительное,
{<падеж, именительный>, <число, множественное >}>,
   ...
   }>  }>
```

Псевдоосновой лексемы называется цепочка наибольшей длины, с которой начинаются все ее словоформы (т.е. наибольший общий префикс). *Псевдоокончанием словоформы* называется ее часть, не входящая в псевдооснову. Например, псевдоосновой лексемы "яблоко" будет "яблок", а псевдоокончанием для формы "яблоками" – "ами".

Таким образом, токен словоформы можно представить как $s = be$, где b – псевдооснова лексемы, а e – псевдоокончание словоформы.

При объединении деепричастий и причастий в одну лексему с глаголами слово сохраняет свою псевдооснову. Так, для глагола "передавать" псевдооснова будет "переда", псевдоокончанием для глагола "передаю" будет "ю", для деепричастия "передавая" псевдоокончание "вая", а для причастия "переданный" псевдоокончанием будет "нный".

Парадигмой лексемы $l = < f_{\text{nf}}, \{< P_{\text{const}}(r, s_{\text{nf}}), L(r)>\}>$ назовем множество $G = \{g\}$, где $g = < e, r, P_{\text{var}}(r, e)>: \exists f = < s = be, r, P_{\text{var}}(r, s)> \in L(r)$, т.е. парадигма содержит информацию об изменении слова. В случае омонимии в множестве G один и тот же префикс e может встречаться для одного слова несколько раз с различным сочетанием параметров $P_{\text{var}}(r, e)$. Более того, один и тот же постфикс может встречаться в различных парадигмах.

Приведем пример парадигмы для лексемы "яблоко":

```
{<"о", существительное, {<падеж, именительный>, <число, единственное>} >,
    <"а", существительное, {<падеж, родительный>, <число, единственное>} >,
    <"и", существительное, {<падеж, именительный>, <число, множественное>} >, …
}
```

Таким образом, лексема может быть представлена как псевдооснова и парадигма лексемы: $l = < g_{\text{nf}}, b, \{< P_{\text{const}}(r, s_{\text{nf}}), G(r)>\}>$. Здесь под g_{nf} понимается элемент парадигмы, соответствующий нормальной форме, а под $G(r)$ – парадигма для данной лексемы при фиксированной части речи, причем $G \bigcup_{n=1}^{k} G(r)$, где k – число частей речи для данной лексемы, а G_n – очередная лексема.

Морфологический словарь представляет собой набор известных лексем $M = \{l\}$, при этом лексемы в словаре удобнее определять именно через псевдоосновы и парадигмы. Для морфологического словаря можно определить множество всех парадигм $M_G = \{G\}$, при этом $|M_G| << |M|$. Задачей морфологического анализа является определение по токену множества всех словоформ словаря, в которые входит этот токен. Задача морфологического синтеза состоит в нахождении токена словоформы с определенными значениями параметров из заданной лексемы.

Математически такая задача может быть описана следующим образом. При анализе слова w необходимо найти все лексемы

$$l = < g_{\text{nf}}, b, \{< P_{\text{const}}(r, s_{\text{nf}}), G(r)>\}>$$

такие, что

$$\exists g_i \in G(r): g_i = < e, r, P_{\text{var}}(r, s)>; \; w = be.$$

Результатом будет множество $\{< s_{\text{nf}}, r, P_{\text{const}}(r, s_{\text{nf}}) \cup P_{\text{var}}(r, s)>\}$. При синтезе по кортежу $< s_{\text{nf}}, r, P_{\text{const}}(r, s_{\text{nf}}) \cup P_{\text{var}}(r, s)>$ необходимо найти элементы парадигмы $g = < e, r, P_{\text{var}}(r, s)>$ такие, что $\exists l = < g_{\text{nf}}, b, \{< P_{\text{const}}(r, s_{\text{nf}}), G(r)>\}>, g \in G(r)$. При этом результатом будет являться $w = be$.

Предсказание парадигмы изменения для несловарной лексики

Под *текстом* будем понимать кортеж токенов $T = < w_1, w_2, ..., w_n >$, при этом токены в тексте могут повторяться, под *несловарным токеном* – такой токен, морфологический анализ которого возвращает пустое множество, т.е. токен, не входящий ни в одну словоформу словаря. Обозначим множество несловарных токенов текста через T'. При этом токен может входить в T' несколько раз.

Можно считать, что несловарный токен принадлежит некоторой лексеме, не входящей в словарь. При этом парадигма изменения большинства несловарных лексем уже имеется в M_G. Поэтому гипотезой о словоизменении несловарного токена w (в дальнейшем просто гипотезой) будем называть лексему $h(w) = < g_{\text{nf}}, b, \{< P_{\text{const}}(r, s_{\text{nf}}), G(r)>\}>$ такую, что $G(r) \in M_G$, $\exists g_i \in G(r): g_i = < e, r, P_{\text{var}}(r, s)>$ и $w = be$. При этом сочетание $< P_{\text{const}}(r, s_{\text{nf}}), G(r)>$ выбирается из M, т.е. под гипотезой понимается лексема, содержащая в себе парадигму изменения, в которой один из ее постфиксов является постфиксом несловарного токена, а ее неизменяемые параметры выбираются исходя из того, с каким набором неизменяемых параметров встречается выбранная парадигма.

В связи с явлением омонимии один и тот же постфикс может встречаться в различных парадигмах. Кроме того, может выясниться, что для данного токена существует более одного постфикса, представленного в M_G, и более одной комбинации неизменяемых параметров и фрагментов парадигмы при фиксированной части речи. В связи с этим по одному токену можно сформировать целое множество гипотез $H(w) = \cup h(w)$. Для текста в целом можно сформировать общее множество гипотез

текста $H_т = \bigcup_{n=1}^{k} H_n(w)$, где k – число токенов в тексте, H_n – очередной токен.

Для внесения в словарь несловарной лексики необходимо предварительно выделить корректные гипотезы. Ручные операции требуют большого количества времени, в связи с чем встает вопрос об автоматизации данного процесса. Для корректного предсказания лексем несловарных слов необходимо, чтобы они отвечали следующим предположениям:

1. В тексте достаточно большого объема токены одной лексемы должны встречаться многократно, в противном случае предсказание упирается в малый размер статистики (слово, встретившееся единственный раз, с высокой вероятностью может оказаться ошибкой).

2. В тексте достаточно большого объема должно быть представлено максимальное количество форм данного слова.

3. Слова, относящиеся к одной парадигме, могут иметь несколько одинаковых букв в конце псевдоосновы.

Как показали предыдущие исследования [7], в русском языке слова, уже имеющиеся в фиксированном морфологическом словаре, и слова, отсутствующие в нем, ведут себя по-разному. На рис. 1 показана зависимость относительной встречаемости парадигм от их заполнения (для словарных словоформ), полученная на корпусе текстов объемом почти $2 \cdot 10^9$ словоупотреблений.

На рис. 2 представлена аналогичная зависимость для несловарных слов. Во-первых, сами данные распределены несколько иначе – процент лексем, в которых встретилось от 95 до 100 % словоформ, у несловарных токенов ниже почти в 300 раз, а процент слабо заполненных парадигм (менее 60 %) отличается в 21 раз (4 % у словарных слов против 84 % у несловарных).

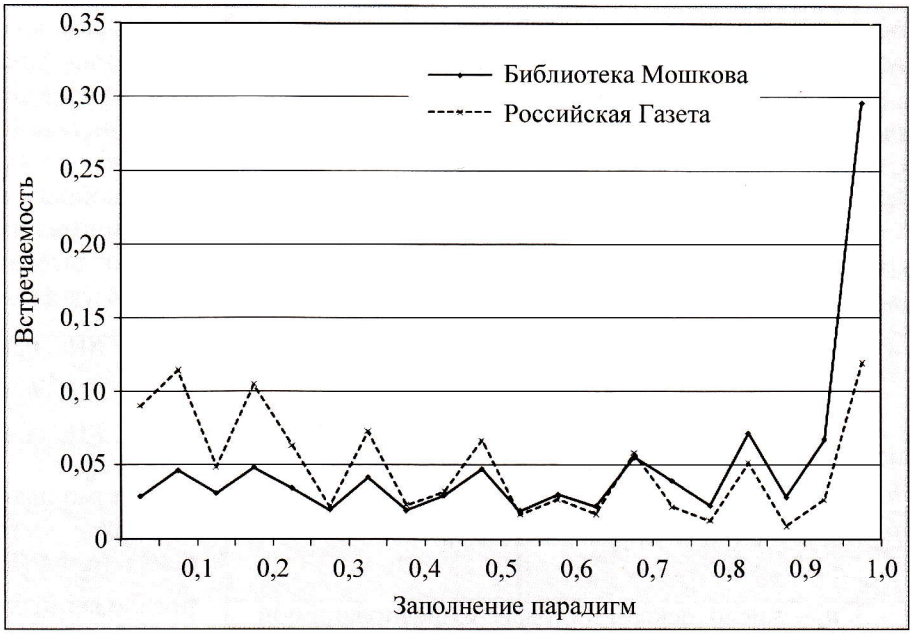

Рис. 1. Зависимость относительной встречаемости парадигм от их заполнения (для словарных словоформ)

Однако общий объем в $2 \cdot 10^7$ несловарных словоупотреблений позволяет рассчитывать на достаточно крупное пополнение морфологического словаря. Тем более, что краткий просмотр полученных результатов показал, что слова, относящиеся к слабо заполненным парадигмам, содержат

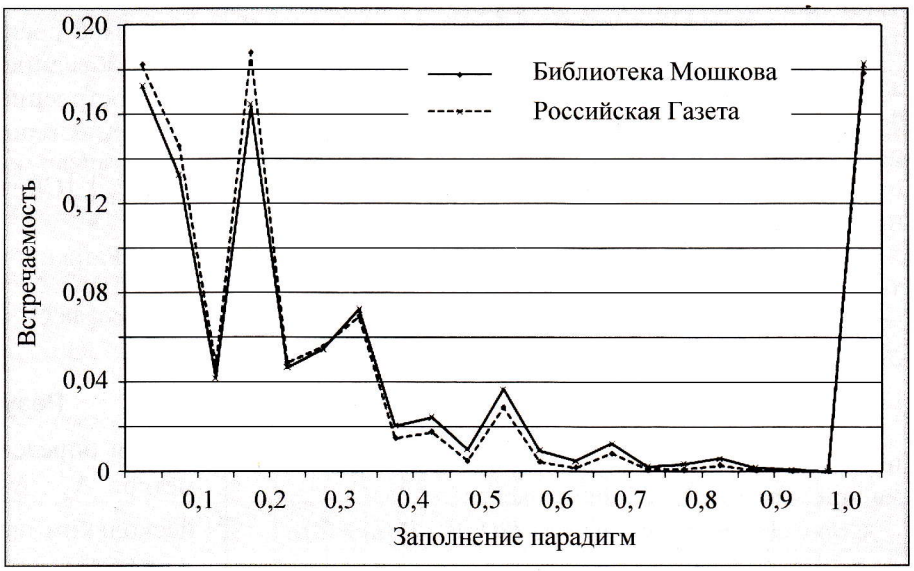

Рис. 2. Зависимость относительной встречаемости парадигм от их заполнения (для несловарных словоформ)

большое количество неточностей отнесения, редко встречающихся имен собственных и ошибок.

Для оценки верности гипотез введем ряд метрик, вычисляемых для каждой гипотезы:

1. n_T – число вхождений в текст токенов, относящихся к данной гипотезе: $n_T(h) = |\{w \in T' : h = < g_{nf}, b, \{< P_{const}(r, s_{nf}), G(r) >\} >, \exists g = < e, r, P_{var}(r, s) > \in G(r), w = b_e\}|$.

2. n_U – число различных словоформ текста, относящихся к данной гипотезе: $n_U(h) = |\{w \in T' : h = < g_{nf}, b, \{< P_{const}(r, s_{nf}), G(r) >\} >, \exists g = < e, r, P_{var}(r, s) > \in G(r), w = b_e, w_i \neq w_j, i \neq j\}|$.

3. n_P – число лексем в словаре с предполагаемой парадигмой G, у которых q символов на конце псевдоосновы b те же, что и в предполагаемой псевдооснове: $n_P^q(h) = |\{1 \in M : 1 = < g_{nf}, b_e, \{< P_{const}(r, s_{nf}), G(r) >\} >, h = < g_{nf}, b_h, \{< P_{const}(r, s_{nf}), G(r) >\} >, \exists a, a_1, c : b_e = ac, b_h = a_1 c, |c| = q\}$, здесь c – общий постфикс псевдооснов b_e и b_h; a, a_1 – оставшиеся префиксы этих псевдооснов.

В силу сформулированных ранее предположений 1–3 высокие значения метрик $n_T(h)$, $n_U(h)$ и $n_P^q(h)$ являются показателями верных гипотез. Таким образом, при анализе текста можно проводить фильтрацию гипотез по этим показателям, введя пороговые значения n'_T, n'_U и $n_P^{q'}$ для $n_T(h)$, $n_U(h)$ и $n_P^q(h)$ соответственно, т.е. строится множество $H'_T = \{h \in H_T : n_T(h) > n'_T, n_U(h) > n'_U, n_P^q(h) > n_P^{q'}\}$. Также стоит учитывать, что для каждой несловарной формы в большинстве случаев существует только одна верная гипотеза. Поэтому после фильтрации имеет смысл группировать гипотезы и выбирать из них наиболее вероятную.

Пусть $B(h) = \{w = be : h = < g_{nf}, b, \{< P_{const}(r, s_{nf}), G(r) >\} >, \exists g = < e, r, P_{var}(r, s) > \in G(r)\}$ – множество всех токенов, составляющих гипотезу, а $B'(h) = \{w \in T' : w \in B(h)\}$ – множество всех встретившихся в тексте токенов данной гипотезы.

Тогда кластер гипотезы – такое множество гипотез, что оно включает в себя все гипотезы начиная с данной, имеющие общие токены:

$C(h) = \{h : \exists h_1 \in C(h), h_1 \neq h, B'(h_1) \cup B'(h) \neq \varnothing\}$.

Гипотеза h служит "отправной точкой" для формирования этого множества. На первом шаге гипотеза h добавляется в множество $C(h)$. Далее присоединяются все гипотезы, имеющие непустое пересечение множеств встретившихся токенов, и ищутся гипотезы, имеющие общие токены с добавленными. Процесс повторяется до тех пор, пока ведется добавление гипотез.

Подобным образом можно разбить все множество гипотез на *компоненты связности*, из которых следует выбрать подмножество лучших гипотез, используя функцию сравнения $R(h_1, h_2)$:

$$\left.\begin{aligned} R(h_1, h_2) &= 1, \text{ если } h_1 \text{ "лучше" } h_2; \\ R(h_1, h_2) &= -1, \text{ если } h_1 \text{ "хуже" } h_2; \\ R(h_1, h_2) &= 0, \text{ если } h_1 \text{ "равноценна" } h_2. \end{aligned}\right\}$$

С помощью этой функции получим множество лучших гипотез каждого кластера $C'(h) = \{h_1 \in C(h) : \forall h_2 \in C(h)\ R(h_1, h_2) \geq 0\}$. На основе предложенных метрик эмпирически была получена функция ранжирования $R(h_1, h_2)$, дающая достаточно хорошие результаты.

Сперва сравниваются значения метрик $n_T(h_1)$ и $n_T(h_2)$, при $n_T(h_1) > n_T(h_2)$ возвращается 1, при $n_T(h_1) < n_T(h_2)$ возвращается –1. При совпадении значений метрик аналогичным образом проверяются метрики $n_U(h_1)$ и $n_U(h_2)$, а при их совпадении – метрики $n_P^q(h)$. При совпадении всех трех метрик возвращается 0.

Таким образом, алгоритм выделения несловарной лексики из текста состоит из следующих шагов:

1. Выделение несловарной лексики (построение множества T').

2. Построение гипотез (множества H_T).

3. Объединение гипотез в расширенные гипотезы (построение множества K_T).

4. Кластеризация гипотез (построение множества $C_T \bigcup_{n=1}^{k} C(h)$).

5. Ранжирование и отбор гипотез (построение множества $C'_T = \bigcup_{n=1}^{k} C'(h)$).

Результаты экспериментов

Для определения оптимальных значений параметров N'_U, N'_T, q, N_P^q алгоритм был применен к нескольким научно-техническим текстам:

- материалы лингвистической конференции "Диалог" [10];
- книга Д. Гибсона "Искусство сведения"[11];

1. Число несловарных словоформ в текстах

Текст	Число несловарных словоформ
Конференция "Диалог"	767
"Искусство сведения"	398
"Объектно-ориентированный анализ и проектирование"	861

• книга Г. Буча "Объектно-ориентированный анализ и проектирование" [12].

В табл. 1 приведено число несловарных словоформ в текстах.

Результаты анализа текстов представлены в табл. 2. Данные по каждому параметру в первой строке соответствуют конференции "Диалог", во второй – книге "Искусство сведения", в третьей – книге "Объектно-ориентированный анализ и проектирование".

Эксперимент выявил зависимости точности и полноты предсказания словоформ от пороговых значений параметров. При увеличении порога числа словоупотреблений в тексте (N'_T) возрастает точность и падает полнота (хотя в [12] наблюдается падение точности при некоторых значениях. Это связано с тем, что в книге встречается ряд имен собственных, каждое из которых употребляется небольшое число раз). При увеличении порогового значения для числа уникальных словоформ наблюдается значительное увеличение точности и небольшое снижение полноты. При порогах 20 словоупотреблений и 2 уникальные словоформы во всех рассмотренных случаях точность достигает 100 %. Длина анализируемого конца основы также оказывает прямое влияние на точность, а обратное – на полноту, хотя значимость этой зависимости варьируется от текста к тексту.

Таким образом, в зависимости от поставленных задач могут быть выбраны различные минимальные значения метрик. В целях повышения точности рекомендуется увеличивать число словоупотребле-

2. Результаты анализа текстов конференции "Диалог" / книги "Искусство сведения" / книги "Объектно-ориентированный анализ и проектирование"

Параметр	Номер эксперимента							
	1	2	3	4	5	6	7	8
Минимальное число словоформ в тексте, N'_T	5	10	15	20	10	10	10	5
	5	10	15	20	10	10	10	5
	5	10	15	20	10	10	10	5
Минимальное число различных словоформ, N'_U	2	2	2	2	2	1	3	2
	2	2	2	2	2	1	3	2
	2	2	2	2	2	1	3	2
Длина анализируемого конца основы, q	1	1	1	1	2	1	1	2
	1	1	1	1	2	1	1	2
	1	1	1	1	2	1	1	2
Число использованных словоформ	152	91	55	43	91	95	87	152
	152	91	55	43	91	95	87	152
	144	84	58	44	84	87	76	142
Число сгенерированных гипотез	38	19	9	6	20	23	17	40
	108	89	61	50	87	91	85	106
	43	17	10	7	19	20	13	43
Число правильных гипотез	34	18	9	6	18	3	17	37
	26	18	12	9	17	20	16	25
	36	14	8	7	19	18	11	40
Точность предсказания, %	89	95	100	100	90	13	100	93
	92	94	100	100	94	85	94	92
	84	82	80	100	100	90	85	93
Покрытие словоформ, %	18	11	7	6	11	2	11	18
	24	17	12	9	16	17	15	23
	14	8	5	5	10	9	7	15

ний, участвующих в формировании лексемы. Для увеличения числа словоформ, для которых может быть выдвинута корректная гипотеза, данное значение лучше уменьшить, увеличив при этом количество проверяемых букв в конце псевдоосновы. Оптимальными нам кажутся параметры, применявшиеся в экспериментах 2 и 5, однако их выбор будет зависеть от анализируемого текста. Для полного исключения ручной работы рекомендуются значения параметров из эксперимента 4.

Выводы

Предложен формальный математический аппарат для описания системы морфологического анализа и синтеза. На основе предложенного аппарата построена модель кластеризации несловарной лексики. Результатом кластеризации являются гипотезы о изменении найденных слов. Полученные кластеры подвергаются ранжированию, по результатам которого отбираются кандидаты на внесение в морфологический словарь.

Приведенный в статье метод выделения несловарной лексики и предсказания парадигмы ее изменения был реализован программно. Данная программа получает на вход текст на естественном языке, а ее выходом являются наборы гипотез о парадигмах слов, отсутствующих в морфологическом словаре.

Создание данной программы позволило провести ряд вычислительных экспериментов и провести оценку значений параметров отсечки. Метод позволяет выбрать полностью точные гипотезы за счет существенного сокращения покрытия выделенных несловарных слов. Подобное покрытие может быть существенно увеличено, однако при этом точность предсказания сокращается до значения ~90 %.

Таким образом, предлагаемый метод может служить основой для создания автоматизированной системы пополнения словарей морфологического анализа.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Сидорова Е.А.** Многоцелевая словарная подсистема извлечения предметной лексики // Тр. Междунар. семинара Диалог'2008 "Компьютерная лингвистика и интеллектуальные технологии". М.: Наука, 2008. С. 475–481.

2. **Елкин С.В., Клышинский Э.С., Стекляников С.Е.** Проблемы создания универсального морфосемантического словаря // Сб. тр. Междунар. конф. IEEE AIS'03 и CAD-2003. Дивноморское. 2003. Т. 1. С. 159–163.

3. **Сегалович И., Маслов М.** Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Тр. Междунар. семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань: ООО "Хэтер", 1998. Т. 2. С. 547–552.

4. **Ляшевская О.Н., Кобрицов Б.П., Сичинава Д.В.** Автоматизация построения словаря на материале массива несловарных словоформ // Интернет-математика 2007: сб. работ участников конкурса науч. проектов по информ. поиску. Екатеринбург: Изд-во Уральского ун-та, 2007. С. 118–125.

5. **Андреев А.М., Березкин Д.В., Симаков К.В.** Обучение морфологического анализатора на большой электронной коллекции текстовых документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: тр. VII Всеросс. науч. конф. (RCDL'2005). Ярославль: Ярославский гос. ун-т, 2005. С. 173–181.

6. **Mikheev A.** Automatic Rule Induction for Unknown Word Guessing // Computational Linguistics. 1997. № 23(3). P. 405–423.

7. **Клышинский Э.С.** Некоторые сложности автоматизированной лемматизации несловарных словоформ // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Междунар. конф. "Диалог–2009", Бекасово, 27–31 мая 2009 г. М.: РГГУ, 2009. Вып. 8 (15). С. 165–169.

8. **Сокирко А.В., Толдова С.Ю.** Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика 2005. Автоматическая обработка веб-данных: сб. работ стипендиатов Yandex 2005. М., 2005. С. 80–94.

9. **Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.** Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет-математика 2005. Автоматическая обработка веб-данных: сб. работ стипендиатов Yandex 2005. М., 2005. С. 38–57.

10. **Международная** конф. по компьютерной лингвистике "Диалог" [Электронный ресурс] // Диалог: [сайт]. [2005–2009]. URL: http://www .dialog-21.ru (дата обращения: 20.12.2010).

11. **Gibson D.** The Art of Mixing, 2nd Ed., ArtistPro, 2005. 344 p.

12. **Буч Г.** Объектно-ориентированный анализ и проектирование, 3-е изд. М.: Вильямс, 2008. 720 с.

Статья поступила в редакцию 27.11.2009 г.

Дебют космического маршрутизатора Cisco "звонок из космоса"

Типичная проблема служб спасения – на месте катастрофы, где нет телефонной связи и возможности спасти себя и других. Любой ураган, землетрясение или наводнение разрушают мобильные базовые станции на многие километры вокруг.

Технология Cisco *IRIS* (*Internet Routing in Space*) – *интернет-маршрутизация в космосе* позволит установить связь по технологии VoIP (голос поверх IP) с помощью спутниковых каналов, причем без использования наземной инфраструктуры для маршрутизации вызовов. В этом состоит радикальное отличие Cisco IRIS от существующей технологии спутниковой связи, которая передает голос и видео между спутниками и конечными пользователями через наземные сетевые узлы.

Cisco неофициально называет свою технологию "звонок из космоса", хотя это название не вполне корректно. Это последнее из приложений, которое компания испытывает с помощью спутника связи Intelsat-14, запущенного на орбиту в прошлом году с установленным на борту впервые в отрасли маршрутизатором Cisco. Кроме того, испытываются решения для IP-мультикастинга и аппаратного шифрования входящих и исходящих IP-потоков.

Cisco хочет предложить эти и другие функции в качестве услуг коммерческим организациям и правительству США. Тем самым реализуется грандиозный план Cisco по трансформации спутниковых сетей, создающий новый и, возможно, очень большой рынок за счет запуска маршрутизаторов на околоземную орбиту на борту спутников связи и выхода Интернета в космос. По мнению руководства Cisco, преимущества такого подхода включают возможность маршрутизации голоса, данных и видео между пользователями по единой IP-сети более эффективным, гибким и экономичным образом, чем с помощью существующих фрагментированных спутниковых сетей.

В настоящее время технология IRIS, проходящая этап тестирования, успела добиться ряда мировых достижений. Так, успешно испытанная в октябре 2010 г. функция "звонок из космоса" впервые в истории воспользовалась технологией Cisco Unified Communications Manager Express для поддержки вызова VoIP через космический маршрутизатор (сегодня эта технология чаще используется в качестве IP-УАТС в корпоративных отделениях).

В октябре 2010 г. впервые в мире Cisco с Земли обновила программное обеспечение маршрутизатора, работающего на борту спутника Intelsat-14. В результате на этом маршрутизаторе удалось активировать целый ряд функций, характерных для наземных продуктов Cisco, и теперь эти функции доступны и в космосе. Это по-настоящему революционный подход для космической отрасли, никогда не менявшей полезную нагрузку спутника после запуска на орбиту. Кроме того, эта возможность открывает для технологии IRIS все богатство функциональности операционной системы Cisco IOS™.

Следующим крупным этапом данной программы станет переход к полномасштабной коммерческой эксплуатации системы IRIS на борту спутника Intelsat-14. По планам Cisco это должно произойти в течение 2011 г. После этого компания намерена начать продажу маршрутизаторов космического базирования. Некоторые компании уже просят предоставить им услуги IRIS со спутника Intelsat-14, хотя испытания этой технологии еще не завершились.

Информация предоставлена пресс-службой ООО "Сиско-системс"

ООО "Издательство **Машиностроение**", 107076, Москва, Стромынский пер., 4.
Учредитель: ООО "Издательский дом "Спектр". Адрес электронной почты редакции журнала: **vkit@mashin.ru, vkitpost@rambler.ru**
Телефоны редакции журнала: (499) 269-54-96; 268-69-19; (495) 514-76-50; тел./факс 268-85-26.
Дизайнер *Свиридова Н.А.* Технический редактор *Андреева Т.И.* Корректоры *Сажина Л.И., Сонюшкина Л.Е.*
Сдано в набор 30.12.10 г. Подписано в печать 21.03.11 г. Формат 60×88 1/8. Бумага офсетная. Печать офсетная.
Усл. печ. л. 6,86. Уч.-изд. л. 7,26. Заказ 120. Свободная цена.
Оригинал-макет и электронная версия подготовлены в ООО "Издательство Машиностроение".
Отпечатано в ООО "Подольская Периодика". 142110, Московская обл., г. Подольск, ул. Кирова, д. 15.