

Автоматическое выявление фамилий в тексте

К.К. Боярский, Е.А. Каневский

Санкт-Петербургский государственный университет
информационных технологий, механики и оптики,
Санкт-Петербургский экономико-математический институт РАН
Boyarin@yandex.ru, Kanev@emi.nw.ru

Аннотация

При компьютерном анализе текстов возникает проблема выявления имен собственных, в частности фамилий, отсутствующих в словаре, а также определения их морфологических и синтаксических характеристик. Описаны подходы к решению этой проблемы, применяемые в семантико-синтаксическом анализаторе SemSin. Эти подходы основываются на комплексном анализе графематических и морфологических характеристик неизвестных слов и семантико-синтаксическом анализе их окружения. Показано, что на новостных текстах удастся опознать как фамилии, имена или инициалы до трети неизвестных слов с точностью свыше 95%.

1. Введение

Компьютерная морфология необходима в прикладных системах, ведущих поиск и анализ информации на естественном языке. Вопросам компьютерной морфологии посвящено множество работ, однако, как показал проведенный недавно Форум «Оценка методов АОТ», эта проблема до сих пор не решена окончательно [1].

Проводимый нами анализ основан на использовании расширенного словаря В.А. Тузова [2]. Объем этого словаря сегодня составляет более 180 тыс. лексем, из которых свыше 23 тыс. являются собственными именами (фамилии, имена и отчества людей, названия стран, городов, учреждений и фирм, рек, морей и т. п.). Все лексемы распределены по 1660 классам, из которых имена собственные занимают около 90 классов. Каждая лексема содержит номер своего класса и актанты вызываемых ею лексем в виде падежей или предлогов с соответствующими падежами — например, вВин, вПред и т. д. Часто перед актантом указаны допустимые классы слов, которые могут их замещать. Морфологический анализатор, используя эту базу данных, осуществляет разбор очередного слова, поданного на его вход. Результат разбора выдается в виде леммы (слова в нормальном виде) с морфологическими

характеристиками и класса (или набора классов) с указанием соответствующих актантов. Следует заметить, что создаваемый нами семантико-синтаксический парсер, в основном, анализирует только те слова, которые имеются в словаре, и рассчитан на использование грамотно написанных текстов.

Однако в реальных текстах почти всегда имеются слова, отсутствующие в словаре. Как показывает практика участия в конкурсах 2009 и 2011 гг. Форума «Оценка методов АОТ» [1, 3], слова, относящиеся к топонимике, со временем могут быть выявлены и вставлены в словарь — новые географические названия возникают не так уж часто. Точно так же ограничено количество русских имен и отчеств. Другая ситуация наблюдается для фамилий и названий учреждений и фирм. В связи с этим и возник комплекс задач автоматического их определения.

Рассмотрим подробнее возможные способы решения этих задач на примере семантико-синтаксического парсера SemSin, который обеспечивает морфологический и синтаксический разбор русскоязычных предложений и строит дерево подчинения [4, 5].

2. Разбиение на предложения

Первая задача, которая должна решаться при анализе текста — это разбиение его на отдельные предложения. На одном из этапов ее решения производится поиск и обработка точек, стоящих после инициалов. В нашем словаре содержится около 12000 русских и иностранных фамилий. Если перед точкой стоит начинающийся с прописной буквы токен длиной один или два символа, а после точки — слово из класса фамилий, то данная точка не является границей предложения, а стоящие слева и справа от нее токены образуют единую именную группу. Аналогично обрабатываются ситуации, когда фамилии предшествуют два инициала, или когда инициалы стоят после фамилии. Например, в предложении *Но ведь Маша знает А. Б. Иванова много лет и никогда про него ничего плохого не слышала!!* точки не являются границей предложения (ГП).

Разумеется, возможны появления в тексте фамилий, не содержащихся в словаре. Если после однобуквенного токена и точки идет неизвестное слово с прописной буквы, оно рассматривается как фами-

лия, а точка опять же не является ГП. В примере с форума футбольных болельщиков «*А что самое интересное лидирует в споре бомбардиров Ф. Лэмпард. Анри - думаю он не будет... пока там травму залечит, пока войдет в игру и т.д.*» анализатор правильно распознает, что *Лэмпард* – фамилия, а ГП является только точка перед фамилией *Анри*.

3. Анализ слов, начинающихся с заглавной буквы

Вторая задача при работе с именами собственными – правильно определить морфологические и семантические характеристики конкретной лексемы. Ситуация осложняется тем, что некоторые слова имеют несколько значений (так например, *Вашингтон* – город, фамилия и начало названия газеты; *Аврора* – крейсер, учреждение, богиня утренней зари и собственно зря), а ряд имен собственных, особенно иностранных, не изменяются по падежам (таковы имена *Азazelло, Алессандро, Альфонсо, Аль-Махди, Амангельды, Андре*; фамилии *Абашидзе, Абраменко, Ага-заде, Бандаранаике, Меркель, Саркози*).

Эта задача разбивается на несколько подзадач.

3.1. Фамилии с частицами

Первая из них заключается в разборе иностранных фамилий, перед которыми стоят частицы *аф, бен, ван, ван дер, ван-дер, да, де, де ла, де-ла, дела, делла, дель, ди, дос, дю, ибн, фон, фон дер, фон-дер* и *эль*. В нижеприведенных предложениях слово *Фридрих* в словаре имеет два значения: имя и фамилия, слова *Большедт, Ахим* и *Арним* в словаре отсутствуют (далее отсутствующие в словаре слова выделяются в примерах полужирным шрифтом):

Шестая армия фельдмаршала Фридриха фон Паулюса капитулировала.

*Альберт фон **Большедт** был доминиканцем.*

*Людвиг **Ахим** фон **Арним** был немецким писателем-романтиком.*

Парсер правильно определяет эти имена и фамилии.

3.2. Семантика имен собственных

Вторая подзадача – осуществить выбор имен собственных по первой прописной букве (если это не начало предложения). Ее решение основано на том, что в словаре все имена собственные имеют специальные признаки (номера их классов оканчиваются "000"). Однако на эти простые соображения наложены некие дополнительные условия: если в результате морфологического анализа выбираются две или более лексемы с именами собственными, то для выбора некоторых классов требуется наличие слева **сопутствующего** (уточняющего) слова. Это слова из классов «Космос» (*звезда, комета, планета*), «Ландшафт» (*гора, лес, море, озеро, океан, пустыня, река*), «Награды» (*медаль, орден*), «Поселения» (*город, остров, мост, площадь, регион, село,*

страна, строение, уезд, улица), «Техника» (*автомобиль, прибор, танк, трактор*) или «Учреждения» (*бюро, гимназия, институт, компания, магазин, пункт, учреждение, фирма, центр, школа*).

Разберем три предложения:

Насчет Ленина между различными категориями "патриотов" согласия не было.

Вчера на улице Ленина взорван автомобиль.

Капитана наградили орденом Ленина.

Здесь при разборе слова *Ленина* выдаются три лексемы:

ЛЕНИН Сущ., муж. Ед. Род. Вин.

(класс «ФИО Фамилия»¹)

ЛЕНИНА Сущ., муж. Ед. Им. Род. Дат. Вин. Тв.

Пред (класс «Страны Улица Название»)

ЛЕНИНА Сущ., муж. Ед. Им. Род. Дат. Вин. Тв.

Пред. (класс «Символика Названия наград»)

Поскольку в первом предложении перед анализируемым словом отсутствует какое-либо сопутствующее слово, то выбирается фамилия. Во втором предложении сопутствующим словом является слово *улице*, относящимся к классу «Поселения Страны Улица», а в третьем – *орден* (класс «Знания Искусство Художество Символика»).

3.3. Неизменяемые фамилии

Третья подзадача – определить морфологические характеристики фамилий, не изменяющихся по падежам. Рассмотрим простой пример:

Г-н Христенко опоздал.

После работы морфологического анализатора получаем:

ХРИСТЕНКО Сущ., муж. Ед. Им. Род. Дат. Вин. Тв. Пред. (класс «ФИО Фамилия»)

ХРИСТЕНКО Сущ., жен. Ед. Им. Род. Дат. Вин. Тв. Пред. (класс «ФИО Фамилия»)

Далее находим сопутствующее слово *г-н*, после чего определяется мужской род и Им. падеж.

3.4. Незвестные фамилии

Четвертая подзадача – правильно определить фамилии, отсутствующие в словаре. Это частично совпадает с общей задачей определения незнакомых собственных имен, которая проводилась методом обучения [7]. Мы же проводим анализ с помощью правил.

Прежде всего, мы попытались определить, какие фамилии чаще всего встречаются в тех текстах, которые были обработаны нами ранее и фамилии из которых занесены в наш словарь. Понятно, почему именно фамилии привлекли наше внимание. Их слишком много, для того чтобы они когда-либо были внесены в словарь полностью. Действительно, на сегодня в нем содержится около 12000 русских и

¹ Здесь и далее для существительных приводятся названия их классов (несколько сокращенные), для остальных слов указывается только часть речи, так как при нашем разборе их классы не учитываются.

иностранных фамилий. В то же время в словаре имеется около 2850 имен и 270 отчеств, и пополняются эти классы очень медленно, в то время как новые фамилии встречаются сотнями.

Для исследований было отобрано 7350 фамилий (в основном, типичные для России, Украины и Грузии), которые встречались наиболее часто. В таблице 1 в процентах приведено распределение фамилий по типам окончаний.

Таблица 1. Встречаемость окончаний фамилий

ов	ова	ин	ев	ина	ский	ева	о
31	22	15	11	10	7	7	5
ко	ер	ич	ская	ан	ян	ли	дзе
4	3	3	3	2	1	1	1

Еще меньше частота встречаемости фамилий с окончаниями -ец, -ак, -ой, -или, -ель, -ын, -ына.

Алгоритм распознавания фамилий следующий.

Определяется окончание неизвестного слова, написанного **с заглавной буквы**. Это окончание для фамилий может быть трехбуквенным {ого, ому, дзе, или}, двухбуквенным {ов, ев, ин, ын, ск, ер, ян, ко, ли, ым, ой, ий, им, ом, ая, ую, ою} или однобуквенным {а, у, е}. В противном случае неизвестное слово не определяется как фамилия.

Если встречается слово с окончанием -ко (Шевченко), -ли (Пресли), -или (Джугашвили), -дзе (Чхеидзе), то такое слово объявляется несклоняемой мужской/женской фамилией.

Если окончание -ер или -ян, то это может быть несклоняемая женская фамилия.

Если окончание {ов, ев, ин, ын, ер, ян}, то это может быть мужской фамилией в Им. падеже.

Делается попытка отбросить окончание {ого, ому, ым, ой, ий, им, ом, ая, ую, ою, а, у, е}. Если после этого оставшаяся основа слова заканчивается на {ов, ев, ин, ын, ск, ер, ян}, это мужская или женская фамилия, род и падеж которой определяется отброшенным окончанием.

Если основа оканчивается на -ск, то это слово может быть не только фамилией, но и прилагательным, определяющим название чего-либо (Новокраматорский).

Практическая реализация данного варианта обработки отсутствующих в словаре фамилий (с учетом возможного наличия прилагательных на -ский) показала вполне приемлемое качество разбора для подобных фамилий, отсутствующих в словаре. Так, например, при разборе предложения

Так вот: надесь арестовали депутата Юрия Горлова...

результат морфологического разбора слова *Горлова* отрицательный – такого слова нет в словаре. После обработки получаем два варианта разбора:

ГОРЛОВ Сущ., муж. Ед. Род. Вин.

(класс «ФИО Фамилия»)

ГОРЛОВА Сущ., жен. Ед. Им.

(класс «ФИО Фамилия»)

В результате работы продукционных правил слева от фамилии находим сопутствующее слово *депутата*, что позволят выбрать первую лемму. Окончательный выбор падежа (Вин.) происходит после присоединения сопутствующего слова к глаголу и последующего согласования падежа слова, обозначающего фамилию, с сопутствующим словом (*депутат*).

3.5. Неизвестные, но не фамилии

В случае, когда неизвестное слово с заглавной буквы не определяется как фамилия, делаются попытки определить семантику этого слова по окружающему контексту с помощью продукционных правил.

Если в сочетание имя – фамилия вклинивается неизвестное слово, то это или отчество или второе имя, что характерно для иностранцев. Подавляющее большинство распространенных отчеств в словаре уже есть. Однако если анализируемое слово оканчивается на -ич, -ыч или -вна (с учетом возможных падежных изменений), то это, скорее всего, новое отчество. Если же нет – это второе имя. Это правило срабатывает для слова *Ахим* в приведенном выше примере

Людвиг Ахим фон Арним.

Если неизвестное слово с заглавной буквы стоит в кавычках, то оно рассматривается как часть названия фирмы, корабля, зрелищного мероприятия и т. д. В этом случае определить морфологические характеристики неизвестного слова затруднительно, т. к. отсутствует согласование между названием и определяемым объектом.

Данная публикация является результатом участия Региональной благотворительной общественной организации «Азария».

В приведенном примере семантический класс слова *организация* позволяет предположить, что к нему может подключаться название.

Также во многих случаях удастся выявить названия объектов, опираясь на сопутствующие слова типа *имени, по имени, по прозвищу*, и подсоединить их правильной связью.

Как названия могут подключаться неизвестные слова (без кавычек) к географическим объектам. Соответствующее правило позволяет правильно связать отсутствующее в словаре слово Гейла со словом кратер в предложении:

Марсианская научная лаборатория с марсоходом Curiosity после восьмимесячного перелета совершила мягкую посадку на Марсе в районе кратера Гейла.

4. Апробация

Для проверки нами был пропущен через анализатор фрагмент корпуса текстов новостного характера, использовавшегося в качестве исходного на конкурсе парсеров 2011 г.[3], объемом 500 предложений, 8686 слов. Результаты следующие: 148 слов отсутствовало в словаре, 47 из них (т. е. 31%) было

обработано подсистемой анализа фамилий. Как фамилии было размечено 29 слов, как имена – 8 слов, как инициалы – 10 слов. Обнаружено две ошибки: в предложении «*Рост налоговых отчислений отмечен также в Грязинском, Лебедянском, Данковском, Долгоруковском, Измалковском, Лев-Толстовском, Тербунском, Краснинском и Становлянском районах*» слова *Данковский* и *Красненский* определены как фамилии, хотя слово *Становлянский* правильно распознано как название района. В предложении «*...сын министра внутренних дел Маргуса Лейво Сандер был задержан при покупке наркотиков*» понять где имя, а где фамилия вряд ли сможет и человек. С другой стороны, в предложении «*Абдулхак Игебаев, Халиса Мударисова, Леонид Соколов, Гульнур Якупова, Маулит Ямалетдинов, Марат Кабиров - их новые книги совсем недавно увидели свет в издательстве "Китан"*» система верно определила все имена и фамилии, а также что *Китан* – название издательства.

Литература

- [1] Ляшевская О.Н. и др. Оценка методов автоматического анализа текста: морфологические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 9 (16). М.: Изд-во РГГУ, 2010. С. 318–326.
- [2] Тузов В.А. Компьютерная семантика русского языка. СПб.: Изд-во С.-Петерб. ун-та, 2004.
- [3] Толдова С.Ю. и др. Оценка методов автоматического анализа текста 2011–2012: синтаксические парсеры русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18), Т. 2. М.: Изд-во РГГУ, 2012. С. 78–92.
- [4] Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб.: ООО «МультиПроджектСистемСервис», 2011. С. 233–237.
- [5] Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор SemSin // <http://www.dialog-21.ru/digest/2012/?type=doc>
- [6] Боярский К.К., Каневский Е.А. Разбиение текста на предложения // Дискуссия теоретиков и практиков. Научно-практический журнал. 2010. №1 (3). С. 135–137.
- [7] Нехай И.В. Применение n-грамм и других статистик уровня символов и слов для семантической классификации незнакомых собственных имен // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып.11 (18), Т. 1. – М.: Изд-во РГГУ, 2012. С. 477–489.

Automatic Recognition of the Names in the Text

Boyarsky K.K., Kanevsky E.A.

When a computer analysis of texts there is a problem of identifying the personal names, absent in the dictionary, as well as determining their morphological and syntactic features. We describe the approaches to solve this problem, used in the semantic-and-syntactic parser SemSin. These approaches are based on the complex analysis of the graphematic and morphological characteristics of unknown words and semantic-and-syntactic analysis of their environment. It is shown that on the news texts it's possible to identify as the names, the surnames or initials of up to one third of unknown words with a precision of more than 95%.