УДК 81.33

СНЯТИЕ ОМОНИМИИ ГЕОЛОКАЦИЙ НА ОСНОВЕ ЧАСТОТЫ ВСТРЕЧАЕМОСТИ КОНТЕКСТОВ

Е.П. Бручес, В.А. Крайванова
Алтайский государственный технический университет им. И.И. Ползунова
г. Барнаул
ООО «Позитивные Технологии»
г. Новосибирск

В работе рассматривается проблема омонимии именованных сущностей на примере топонимов (географических названий). Для решения этой проблемы нами предлагается статистический подход, учитывающий только контекст предполагаемых географических названий. Для имплементации метода не требуются специально размеченные корпуса текстов, работа алгоритма на каждом шаге легко интерпретируема (в отличие от методов машинного обучения), а результаты являются достаточными для его практического использования на больших объёмах данных в задаче информационного поиска.

Ключевые слова: Извлечение топонимов, автоматический анализ текстов, снятие омонимии.

В настоящее время задача распознавания именованных сущностей является актуальной и используется во многих приложениях автоматической обработки текстов. Среди именованных сущностей, как правило, выделяют имена людей, названия организаций, географические названия, события, временные и денежные обозначения и пр. Сама задача распознавания именованной сущности состоит в определении типа упоминаемой сущности и осложняется омонимией, т.е. совпадением названий двух разных типов сущностей (например, '*Жуков*' (город) и '*Жуков*' (фамилия)) или совпадение с общеупотребительным словом (например, '*Минеральные Воды*'). При этом такая проблема является одной из ключевых в задачи информационного поиска, т.к. при увеличении объёмов базы данных (знаний) существенно увеличивается количество случаев омонимии, что было показано в работе [6].

Принято выделять три типа омонимии именованных сущностей:

1) языковая омонимия (например, "*Металлист*" (ОАО) vs. '*металлист*' (поклонник определенного музыкального направления);

2) омонимия классов ('*Владимир*' (город) vs. '*Владимир*' (имя));

3) онтологическая омонимия ('*Сергей Иванов*' (политик) vs. '*Сергей Иванов*' (ученый)) [8].

В данной работе омонимия разрешается для случаев (1) и (2).

Также гостиницы, магазины, рестораны и т.д. могут быть названы практически любым наименованием, совпадающим с названием географической локации, такие случаи не рассматриваются в нашей работе.

Задача определения именованных сущностей решается, как правило, в первую очередь, для новостных статей. Примеры таких решений можно найти в [9]. Однако, в новостном потоке большой процент случаев упоминания объектов в тексте составляют «топовые» объекты. Поэтому достаточно очень точно учесть в системе случаи омонимии именно для данных объектов, чтобы получить хорошие результаты [8]. Если речь идет о произвольных текстах, например, статьях в Википедии или в блогах, то результаты могут существенно отличаться от новостных корпусов.

Для выделения именованных сущностей и снятия с них омонимии применяется ряд подходов. Перечислим их.

В статье [8] описан подход, основанный на словарях и регулируемый эвристическими правилами. Такой метод показывает высокую точность, а его работа является прозрачной для пользователя-неспециалиста.

Другой группой являются статистические подходы, которые, например, описаны в статьях [6] и [2], где используется априорная информация (например, Билл Клинтон чаще встречается в текстах, чем округ Клинтон) и её комбинации с другими признаками.

Также в последнее время алгоритмы машинного обучения успешно применяются

как непосредственно для распознавания именованных сущностей [10], так и для разрешения многозначности [7].

Цель работы состоит в разработке алгоритма извлечения географических названий из неструктурированных текстов на русском языке для дальнейшего использования их в задаче информационного поиска.

Описание решенной задачи. Для решения задачи использовались три источника данных:

− база географических названий geonames [3];

− дамп русскоязычной Википедии [4];

− словарь для морфологического разбора [1].

Мы использовали статистический метод, который заключается в сборе контекстов топонимов, для которых известно, что в заданном тексте они в большинстве случаев топонимы (примером исключения является фраза "*такими как копии гостиницы «Москва»*" в статье о Москве в Википедии).

В качестве источников контекстов использовались статьи Википедии, описывающие значения топонимов, например, Москва. Эти топонимы отбирались на основе списка шаблонов геолокаций Википедии, таких как "{{НП+Россия}}".

Общая схема алгоритма построения контекстов выглядит следующим образом.

1. Из русскоязычной Википедии извлекаются тексты, удовлетворяющие условию для построения контекстов на основе списка шаблонов геолокаций.

2. Тексты проходят этап подготовки, а именно:

− все даты были заменены строкой "date";

− все числа, отличные от дат, были заменены строкой "number";

− все группы символов, являющиеся тем или иным представлением фамилии, имени и отчества, были заменены строкой "name".

− каждое слово подверглось стеммингу.

3. Для каждого вхождения топонима в соответствующий ему текст собираются множества всех возможных слов (приведённых к базой форме), стоящих перед ним и после него.

В качестве контекстов использовалось слово до географического названия и слово после. Более длинные контексты порождают большое количество вариантов. Мы использовали два множества контекстов: предыдущих L и последующих R – по отдельности, так как их комбинация также приводит к комбинаторному взрыву.

Для каждого элемента $c_L \in L$ контекста была рассчитан весовой коэффициент $w(c_L) = n(c_L)/|L|$, где $n(c_L)$ – количество элемента контекста (c_L) в обучающей выборке, $|L|$ – количество элементов в множестве L. Аналогичные коэффициенты получены для множества R.

Пусть теперь для некоторого слова t необходимо оценить, является ли оно географическим названием. Обозначим слова, расположенные слева и справа от него соответственно t_L и t_R. Слово t считается географическим названием, если выполняется одно из следующих условий (проверяются условия в указанном порядке).

− Если $t_L \in L$, и $w(t_L) > b_L$, где b_L – это пороговое значение для множества предыдущих контекстов, и влияющее на полноту и точность;

− Если $t_R \in R$, и $w(t_R) > b_R$, где b_R – это пороговое значение для множества последующих контекстов;

− Если $t_L \in L$ и $t_R \in R$ и $w(t_L)+w(t_R) > b$, где b – это пороговое значение для суммарного контекста.

Для обнаружения в тексте кандидатов на снятие омонимии использовалась база geonames [3] и морфологический словарь [5].

Таблица 1 – Результаты тестирования для распространенных омонимов

	Точность	**Полнота**	**F-мера**
"*Горький*"	0,8	0,54	0,65
"*Жуковский*"	0,81	0,73	0,77
"*Владимир*"	0,43	0,82	0,57

Для тестирования предложенного алгоритма был вручную размечен корпус текстов, состоящий из 5000 предложений, каждое из которых имело вхождение одного из следующих слов: "*Горький*", "*Жуковский*", "*Владимир*" в различных формах. Таким образом, проверка проводилась только для омонимичных локаций. В таблице 1 представлены результаты тестирования. В качестве метрик мы использовали точность, полноту и F-меру, рассчитанные по следующим формулам.

− Точность: отношение количества правильно распознанных локаций к общему количеству случаев, распознанных системой.

− Полнота: отношение количества

правильно распознанных локаций к общему количеству локаций, которые должны быть распознаны.

− F-мера: (2*Точность*Полнота) / (Точность + Полнота).

Самыми “сильными” предыдущими контекстами являются те, которые представляют собой предлоги (‘*в*’, ‘*на*’, ‘*от*’) и виды географических объектов (‘*город*’, ‘*село*’, ‘*посёлок*’, ‘*округ*’, ‘*река*’). Среди “сильных” последующих контекстов можно выделить глаголы местонахождения (‘*есть*’, ‘*расположен*’, ‘*находится*’, ‘*входит* (*в*)’). Конкретные примеры приведены в таблице 2.

Таблица 2 – Самые распространенные контексты

Левый контекст		**Правый контекст**	
Контекст	**Вес**	**Контекст**	**Вес**
в	0.099	*село*	0.077
село	0.058	*деревня*	0.056
город	0.035	*коммуна*	0.042
деревня	0.033	*река*	0.041
год	0.032	*в*	0.039
река	0.025	*находиться*	0.036
название	0.024	*есть*	0.031
посёлок	0.02	*располагаться*	0.023

Таким образом, мы показали, как с помощью статистического подхода разрешается омонимия географических названий, что позволяет улучшить извлечение топонимов из текстов. При наличии корпуса текстов и заранее известных однозначных географических названий данный метод может быть перенесён на другие языки. Другим достоинством описанного подхода является отсутствие необходимости в специально размеченном корпусе текстов для обучения. Улучшить показатели можно за счёт более полной идентификации персоналий и дат. Здесь также встаёт задача разрешения неоднозначности именованных сущностей, но уже других классов.

СПИСОК ЛИТЕРАТУРЫ

1. Bruches E., Karpenko D., Krayvanova V.: The Hybrid Approach to Part-of-Speech Disambiguation // Proceedings of the Fifth International Conference on Analysis of Images, Social Networks and Texts (AIST 2016). Yekaterinburg, Russia, April 6-8. – 2016.
2. Fader A., Soderland S., Etzioni O.: Scaling Wikipedia-based named entity disambiguation to arbitrary web text // Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA. 2009. P. 21–26.
3. http://gis-lab.info/
4. https://dumps.wikimedia.org/backup-index.html
5. http://176.9.34.20:8080/com.onpositive.text.webview/parsing/
6. Антонов Е.С. Априорная информация как способ разрешения онтологической и языковой омонимии // Филологические науки: Вопросы теории и практики. – 2012. – N 6. – С. 15-19.
7. Антонов Е.С. Разрешение онтологической и языковой омонимии именованных сущностей с помощью интерпретации данных онтологии // Вестник Томского государственного педагогического университета. – 2013. – N 8 (136). – С. 200-204.
8. Брыкина М.М., Файнвейц А.В., Толдова С.Ю. Извлечение и идентификация именованных сущностей с использованием словарей в русском языке // Актуальные инновационные исследования: Наука и практика. – 2013. – N 1.
9. Трофимов И.В. Выявление личных имен в новостных текстах на материале коллекций Persons-1000/1111-F // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XVI Всероссийская научная конференция RCDL-2014, Дубна, 13-16 октября 2014 г. : труды конференции / Российский фонд фундамент. исслед. [и др.]; [сост. Л. А. Калмыкова, М. Р. Когаловский]. – Дубна : Объединенный институт ядерных исследований, 2014. – С. 217-221.
10. Юсупов И. Распознавание именованных сущностей с использованием синтактико-семантических признаков и нейросетей // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог».

Бручес Елена Павловна, магистр, компьютерный лингвист, тел. +7-913-462-5434, e-mail: bruches@bk.ru, Крайванова Варвара Андреевна – к.ф.-м.н., доцент кафедры “Прикладная математика” АлтГТУ, тел.: +7-913-230-3419, e-mail: krayvanova@yandex.ru.