электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 81'373.21

Построение автоматизированной системы поиска топонимов в тексте на русском языке

Пантелеев М.Ф., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»

Научный руководитель: Самарев Р.С., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»
bauman@bmstu.ru

Введение

Средства массовой информации каждый день производят огромное количество данных, в частности, новостей, которые описывают различные явления, происходящие в мире. Достаточно часто в новостях важную роль играет географический объект, к которому может быть отнесена данная новость.

Процесс присоединения географических данных к новостям, различным информационным ресурсам, фотографиям и пр. получил название “геотегинг”. Он заключается в нахождении всех слов/связок в предложении, которые могут быть отнесены к различным географическим местам, которые называются топонимами и дальнейшем их разрешении. Оба этих процесса достаточно сложны в реализации из-за особенностей конструкций естественного языка.

Также имеют место различные неоднозначности: некоторые топонимы могут иметь одинаковые имена с сущностями, никакого отношения к географическим местам не имеющим или топонимы с одинаковыми именами могут одновременно относиться к нескольким географическим местам (многозначные топонимы).

В России эта проблема является достаточно новой и на данный момент не существует открыто распространяемых систем, которые производили бы топонимизацию исходного текста.

Стоит также отметить, что не существует достаточного количества необходимых средств для работы с языком, а существующие на данный момент синтаксические анализаторы, находящиеся в открытом доступе, не отвечают требованиям высокой точности, а деревья разбора предложений, построенные на основании синтаксического

анализа, в действительности достаточно редко отображают реальную структуру предложения. Это происходит, частности, в следствие необходимости привлечения средств семантического анализа для разрешения различных видов синтаксических неоднозначностей, которые на данный момент проработаны достаточно плохо ввиду высокой сложности их написания.

Входной поток данных

Важно определить, что будет являться входным потоком данных. Записи в социальной сети, например, могут не отвечать различным требованиям орфографической правильности слов в предложении, синтаксически верно построенным конструкциям слов. Существующие на данный момент средства анализа текстов и так далеки от совершенства, поэтому с целью упрощения их задачи было разумно использовать новостной поток данных официальных источников, где авторы статей несут прямую ответственность за орфографию слов и синтаксическую правильность построения предложений.

Предлагаемый алгоритм

Проведенный анализ работы группы зарубежных учёных, которые разрабатывают одну из самых мощных систем, решающую задачу топонимизации, показал, что все наиболее эффективные методы извлечения топонимов из текста можно объединить и свести в единый. Различаться, в основном, будут только инструментальные средства, применяемые алгоритмы, но при этом всегда будут использоваться как методы NER (Named Entity Recognizer – распозаватель сущностей, далее будет рассмотрен более подробно), так и использование результатов морфологического/синтаксического анализа текста.

Имеет смысл разделить топонимизацию исходного текста на 2 этапа:

1. Получение исходного текста, работа с ним, подготовка к анализу слов-кандидатов.
2. Поиск слов кандидатов в таблицах / распознавание слов-кандидатов методами NER.

Предлагаемый алгоритм работы представлен на рисунке 1.

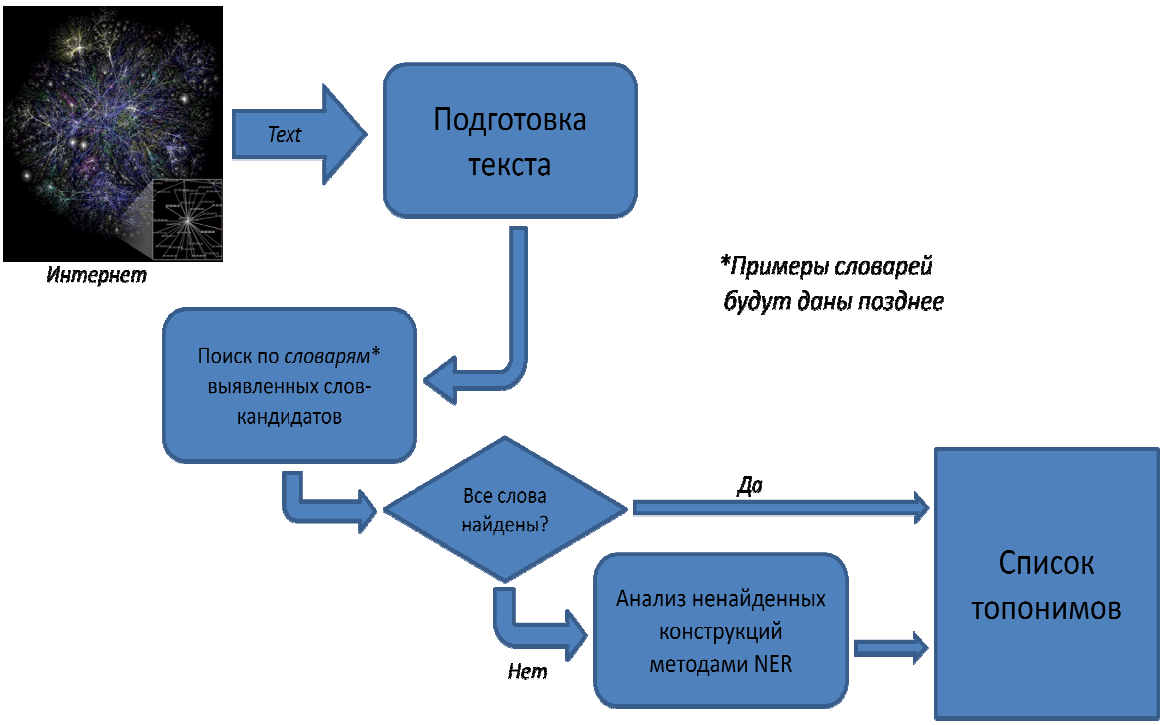

Рис. 1. Алгоритм топонимизатора

Рассмотрим каждый из этапов подробнее.

Получение и подготовка исходного текста.

Было принято решение разделить данную подзадачу на следующие этапы:

Таблица 1

Этапы подготовки текста

Этап	Предлагаемые средства
Получение исходного текста.	Скрипт на Ruby.
Разбиение текста на предложения.	Токенизатор.
Разбиение предложений на: А) Части Б) Синтаксические конструкции	А) Токенизатор. Б) Средства AOT/Yandex.
Морфологический разбор слов	Средства Yandex, AOT.
Применение фильтров, рефакторинг, лемматизация	Средства Yandex, AOT.

После разбивки текста на синтаксические конструкции, необходимо применить фильтрацию, которая обеспечит отсечение конструкций, которые с очень большой вероятностью не будут иметь никакого отношения к географическим объектам.

Предлагается использовать следующие типы фильтров:

1. По части речи.
2. По морфологическим признакам слов-кандидатов.
3. По большой букве.
4. По форме глагола, с которым согласуется слово-кандидат или слово в связке-кандидате.

Фильтрация по части речи обусловлена следующим: известно, что некоторые части речи не могут в принципе являться топонимами, например: деепричастия, частицы, глаголы.

Фильтрация по морфологическим признакам осуществляется из соображений, что, например, имена существительные можно подразделить на имена собственные и имена нарицательные. Топонимами являются именно имена собственные.

Фильтрация по большой букве применяется исходя из соображений, что слова-топонимы в тексте, как правило, начинаются именно с нее.

Исследования русского языка показали, что если в предложении идёт речь о каком-то топониме и этот топоним связан с глаголом, этот глагол, чаще всего употребляется в пассивной форме.

После фильтрации наступает очередь заключительного этапа – рефакторинга, который заключается в переработке полученных конструкций с целью приведения их к единообразию. Краткий список, иллюстрирующий, что должно выполняться на стадии рефакторинга, сведен в таблицу 2.

Таблица 2

Рефакторинг.

Исходная конструкция	Нормированная конструкция
шк. X ул. X	школа X улица X
пр-кт X	проспект X
о-в X	остров X
г. X	город X
г-во X	государство X
<сущ.> <прил.>	<прил.> <сущ.>

Поиск слов-кандидатов в словарях

Полученные синтаксические связки сначала ищутся в специально составленных словарях. К примеру, можно использовать следующие словари:

1) Известных географических наименований.
2) Этнохоронимов.
3) Известных сущностей.
4) Слов, которые чаще всего могут быть отнесены к топонимам.

Необходимость наличия словаря известных географических наименований очевидна. Подавляющее большинство новостей может быть отнесено именно к этим географическим объектам и чрезвычайно важно обеспечить как можно более высокий шанс их идентификации.

Существуют слова, которые являются косвенными топонимами, например, этнохоронимы. Они означают принадлежность какого-либо человека определенной местности.

Когда непосредственно топоним не найден, имеет смысл искать в тексте какие-либо известные сущности, к примеру, торговые компании. В таком случае, если в ходе анализа конкретной новости не будет выявлено топонимов, можно будет проассоциировать ее, например, со штаб-квартирой какой-либо компании.

В ходе анализа целесообразно выявлять слова, которые чаще всего в русском языке связаны с географическими объектами. Список этих слов можно получить, например, обработав находящийся в открытом доступе каталог географических названий.

Распознаватель сущностей

Распознавание конструкций-кандидатов при помощи методов NER является заключительным этапом нахождения топонимов в тексте.

Одной из разновидностей информационного поиска является задача извлечения информации, т.е. извлечение структурированных данных из неструктурированных документов. Одной из задач извлечение информации является задача распознавания именованных сущностей. Задача распознавания именованных сущностей — это выделение в тексте последовательностей слов, являющихся именованными сущностями, и классификация выделенных именованных сущностей. Примерами классов именованных сущностей являются имена людей, названий организаций, географических названий, прочие типы имен собственных, а также выражения специального вида, такие, как обозначения моментов времени, дат, денежные суммы и процентные выражения.

Нас в первую очередь интересует достаточно ограниченный список типов именованных сущностей. Вполне можно ограничиться следующими типами:

1. Личность.
2. Топоним.
3. Организация.
4. Неизвестный тип.

Понятно, что NER занимается решением более общей задачи и выделение именованных сущностей-географических объектов является лишь подзадачей, но на данный момент именно применение средств NER на заключительных этапах поиска топонимов в тексте на естественном языке обеспечивает наибольшую эффективность поиска топонимов в тексте. Список систем, предоставляющих NER, достаточно невелик, в основном они разрабатываются зарубежными компаниями. Примерами могут служить

1. AlchemyAPI .
2. Apache Stanbol

Однако, к сожалению, эти системы демонстрируют достаточно плохую эффективность при работе с неизвестными географическими объектами.

Было принято решение заняться написанием NER самостоятельно. Можно выделить следующие основные подходы к написанию:

1) По онтологиям.
2) Rule-Based системы.
3) Опираясь на машинное обучение.

В нашем случае онтологии — это «концептуальные словари», представляющие собой структуры, в которых описываются некоторые понятия и/или объекты, отношения между ними, а также их характеристики.

Онтологии могут быть универсальными (в них предпринимается попытка описать максимально широкий набор объектов), отраслевые (с информацией по предметным областям) и узкоспециализированные (предназначенные для решения конкретной задачи). Также могут применяться онтологии объектов (базы знаний). Наиболее яркий пример базы знаний — это Википедия.

Опираясь на контексты и уже имеющиеся списки объектов можно строить гипотезы по отношению к объектам и фактам в тексте, а далее верифицировать или отклонять эти гипотезы. Сделать это можно разными способами. Чаще всего применяется машинное обучение, различные контекстные и синтаксические факторы.

Извлечение информации с помощью онтологий позволяет получить достаточно

высокую точность NER и отсутствие случайных срабатываний. Снятие омонимии также происходит с высокой точностью.

Подход, основанный на правилах, т.е. написание шаблонов вручную, заключается в следующем: аналитик составляет описания типов информации, которые необходимо извлечь. Подход удобен тем, что если в результатах анализ обнаруживаются ошибки, очень просто найти их причину и внести необходимые изменения в правила. Проще всего составляются правила для относительно стандартизированных объектов: имен, дат, наименований компаний и т.п., однако этот подход плохо применим для нашего случая, т.к. достаточно сложно написать набор правил, которые описывали бы такую сущность, как топоним.

Машинное обучение требует большого объема вводных данных. Нужно максимально покрыть лингвистической информацией обучающую выборку текстов: разметить всю морфологию, синтаксис, семантику. Плюсы этого подхода в том, что он не требует ручного труда помимо создания размеченного корпуса. Не нужно составлять правила или онтологии. При необходимости такая система легко перенастраивается и переобучается. Правила получаются более абстрактными. Однако есть и минусы. Инструменты для автоматической разметки русскоязычных текстов пока не очень развиты, а существующие не всегда легко доступны. Корпуса должны быть достаточно объемными, размечены верно, единообразно и полностью. А это достаточно трудоемкий процесс.

Существуют различные методы машинного обучения. В них входит применение таких методов, как принцип минимальной длины, метод опорных векторов, наивный байесовский классификатор и пр.

На данный момент существуют системы, которые предоставляют модели для обучения классификатора. В их число входит, например, система MALLET – MAchine Larning for LanguagE Toolkit. Эта система включает в себя средства для тренировки NER, основанные на методах HMM (Hidden Markov Models), MEMM (Maximum Entropy Markov Models) и CRF (Conditional Random Fields).

Проблемы топонимизации.

При нахождении топонимов в исходном тексте и отнесении их к определенным географическим объектам могут возникать различные неоднозначности, которые следует иметь ввиду. Их классификация:

1. Географические/негеографические. Возникают в случае, когда топонимы имеют одинаковые названия с сущностями, никакого отношения к топонимам не

имеющими. (Например, город Сказка в Московской области и сказка о золотой рыбке за авторством А.С. Пушкина)

2. Географические/географические. Возникают в случае, когда топонимы можно отнести к различным местам, которые имеют одинаковое название. (Например, в США есть города с названием «Одесса», «Москва», «Париж» и пр.)
3. Проблемы, возникающие в случае, когда определенное место может быть адресовано различными топонимами. (Например, у топонима «Москва» есть несколько синонимов, например «столица России», «третий Рим» и др.)

Заключение

В ходе проделанной работы был составлен алгоритм извлечения топонимов из текста на естественном языке. Был рассмотрен каждый шаг алгоритма в отдельности и были предложены различные варианты реализации каждой подзадачи. Были предложены словари, предназначенные для решения задачи топонимизации, некоторые из которых были уже составлены. Было предложено использовать методы NER на заключительных этапах алгоритма топонимизации, были рассмотрены возможные методы построения классификатора, предложены системы, которые предоставляют средства NER или предлагают модели для тренировки классификатора на основе обучающей выборки.

Список литературы

1. Michael David Lieberman. Multifaceted geotagging for streaming news. Doctor of Philosophy. Maryland, 2012. 275 p.
2. Блог компании Yandex. Извлечение объектов и фактов из текстов в Яндексе. Режим доступа:
3. http://habrahabr.ru/company/yandex/blog/205198/ (дата обращения 15.12.2014)
4. Томита Парсер, документация. Режим доступа:
5. https://tech.yandex.ru/tomita/doc/dg/concept/about-docpage/ (дата обращения 15.12.2014)
6. Дубова Н.А. Замечание о топонимах-прилагательных в русской речи // Румянцевские чтения-2014: материалы междунар. науч. конф. (15—16 апреля 2014). В 2 ч. Ч. 1. М.: Пашков дом, 2014. С. 229 - 234.