САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА ТЕХНОЛОГИЙ ПРОГРАММИРОВАНИЯ

Андриенко Артем Сергеевич

Выпускная квалификационная работа бакалавра

Выделение именованных сущностей в текстовых документах

Направление 010400

Прикладная математика и информатика

Научный руководитель,

старший преподаватель

Мозжерина Е. С.

Санкт-Петербург

2016

Содержание

1. http://nlp.cs.nyu.edu/ene/

Введение

Объемы информации, текстовых документов повышаются. С каждым годом количество информации увеличивается. Данный процесс происходит по совершенно естественным причинам: мир растет и развивается, люди учатся и самосовершенствуются, пишут картины, сочиняют стихи, проводят научные исследования, etc. Всё это неуклонно повышает объемы информации.

Письменность давным-давно зарекомендовала себя как надёжный источник сохранения и передачи информации. Ещё совсем недавно, несколько сотен лет назад, объемы текстовой информации весьма эффективно регулировались по крайне прозаичной причине: дороговизна и сложность производства материала для записи. С изобретением и поразительно стремительным развитием цифровых запоминающих устройств и интернета, увеличение объемов информации приобретает лавинообразный характер. Уже сейчас для нормальной работы в Интернете жизненно необходимы методы поиска и извлечения информации.

Существует множество различных задач обработки естественного языка, вот несколько наиболее часто исследуемых задач:

1) Автоматическое реферирование(Automatic summarization) – создание читаемого краткого изложения текста.
2) Машинный перевод(Machine translation) – автоматический перевод с одного естественного языка на другой. Одна из наиболее сложных задач, считается, что она принадлежит к классу так называемых «AI-полных задач».
3) Морфологическая сегментация(Morphological segmentation) – разделение слов на морфемы. Сложность задачи целиком зависит от сложности морфологии рассматриваемого языка.

2. http://nlp.cs.nyu.edu/ene/

4) Частеречная разметка(Part-of-speech tagging) – задача определения части речи для поданного на вход предложения. Многие слова могут служить различными частями речи в зависимости от контекста.
5) Синтаксический анализ(Parsing) – создание синтаксического дерева, синтаксическую структуру входной последовательности и хорошо подходит для дальнейшей обработки.
6) Информационный поиск(Information retrieval) - процесс выявления в некотором множестве документов всех тех, которые удовлетворяют заранее определенному запросу.
7) Анализ тональности текста(Sentiment analysis) – определение и извлечение субъективной информации, обычно из множества документов. Часто используется для определения «полярности» отзывов. Особенно эффективно для распознания общественного мнения в социальных медиа.
8) Извлечение информации(Information extraction, IE) – извлечение структурированной семантической информации из текста.

Так же, помимо упомянутых выше, существует ещё множество задач и подзадач так или иначе связанных с обработкой естественных языков.

Термин «Named entity»(Именованная сущность, NE), который сейчас широко используется, впервые был введён в употребление на шестой Message Understanding Conference (MUC-6) в 1996 году. В то время конференция фокусировалась на задаче извлечения информации. В процессе определения задачи заметили, что необходимо уметь распознавать в тексте такие вещи как имена, организации, местоположения и числовые выражения, включая время, дату, деньги, etc. Идентификация ссылок на подобные сущности в тексте была определена как одна из важных подзадач IE и названа «Распознание именованных сущностей».

Решению данной задачи посвящено наше исследование.

2. http://nlp.cs.nyu.edu/ene/

Постановка задачи

Задача данной дипломной работы состоит в разработке метода извлечения именованных сущностей, который использует информацию, полученную из Википедии. Для этого требуется:

1. Исследовать существующие методы извлечения именованных сущностей и извлечения информации из Википедии.
2. Разработать метод извлечения именованных сущностей, который использует полученную из Википедии информацию. Метод должен
3. Выполнить программную реализацию разработанного метода,
4. Провести тестирование качества разработанного метода.

2. http://nlp.cs.nyu.edu/ene/

Глава 1. Теоретический обзор

Говоря более формально, задача распознавания именованных сущностей (Named-entity Recognition, NER)состоит в обнаружении и классификации элементов текста — слов и последовательностей слов — по предопределенным категориям. Например, в предложении

Paris/PERSON Hilton/PERSON visited/0 the/0 Paris/LOCATION Hilton/ORGANIZATION

различные вхождения слова Paris соответствуют личному имени, географическому названию и атрибуту организации. Разрешение подобных многозначностей делает задачу распознавания именованных сущностей сложной задачей семантической обработки текстовой информации

Существует два необходимых критерия именнованых сущностей:

1. Пишется с заглавной буквы
2. Обязательно имеет референт, то есть того (тех), кому это имя принадлежит

Например, «На экраны вышел новый фильм братьев Коэнов»

“Коэны” имеют референт (конкретные два Коэна, являющиеся братьями друг другу), а значит перед нами именованная сущность. В предложении «В Воркуте новорождённых девочек родители очень редко именуют Татьянами» слово «Татьянами» — именованной сущностью являться не будет, так как не имеет референта. Именованной сущностью будет считаться самая длинная цепочка последовательных слов, отражающих имя.

1. Персона(PER)

Относится к обозначению живых существ. Все дополнения: фамильные приставки, признаки старшинства, родства (в случае написания через тире) и т.п. считаются частью именованной сущности

2. http://nlp.cs.nyu.edu/ene/

Прилагательные/причастия, не являющиеся прозвищем или частью имени, не включаются в ИС.

2. Местоположение(LOC)

   Относится к обозначениям объектов, указывающих на положение в пространстве.

3. Организация(ORG)

   Если сущность может быть местом работы человека или в ней можно состоять в качестве члена и цепочка, включает какие-то слова, кроме указания на родовое понятие, то это сущность типа org. Все названия организаций, даже употребленные в значении местоположения, размечаются как org.

4. Остальное(MISC)

   Все то, что является именованной сущностью, но не подходит под описания выше.

Дробление представленных выше классов привело к созданию разнообразных таксономий. Например: в работе [1] представлена иерархия, состоящая из 29 классов. В ней добавляются численные величины (денежные суммы, дата, время) в класс MISC, а также LOC разделяется на два больших класса местоположение и географические объекты, содержащие страны, города, субъекты и реки, озера, моря, океаны, континенты, регионы соответственно.

Также в [2] вводится таксономия, включающая в себя 200 классов, список которых доступен на *сайте*[1] .

1.1 Классификация. Формальная постановка задачи обучения.

2. http://nlp.cs.nyu.edu/ene/

Учитывая сказанное, ясно, что мы можем рассматривать задачу распознания именованных сущностей, как задачу одного из разделов машинного обучения – классификации.

Формально задача классификации ставится следующим образом:

Пусть **X**— множество описаний объектов, **Y**— конечное множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение $y^{¿}: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. Требуется построить алгоритм $a: X \to Y$, способный классифицировать произвольный объект $x \in X$.

В нашем случае множество классов $Y = \{PER, LOC, ORG, MISC\}$.

1.2 Методы выделения именованных сущностей

На данный момент существуют три подхода:

1. Написание вручную наборы правил, для распознания.
2. Статистические классификаторы.
   2.1.Генеративные(Generative).
   2.2.Дискриминантные или условные(Discriminative or conditional).
3. Глубинное обучение нейронными сетями.

При этом написанные вручную системы обычно имеют большую точность, но требуют многомесячной разработки группой опытных компьютерных лингвистов. И в дальнейшем требуют постоянной поддержки и своевременной корректировки. К ним принадлежит одна из первых работ [3], в которой с помощью эвристик и наборов правил выделяются названия компаний.

В литературе упоминаются примеры использования большого числа обучающихся алгоритмов. При этом к генеративным относят модели основывающиеся на совместном распределении: наивный байесовский

2. http://nlp.cs.nyu.edu/ene/

классификатор, скрытые марковские цепи, etc. К дискриминантным же те, что основываются на условном распределении: метод максимальной энтропии[4], линейные цепи условных случайных полей и т.д. В серии работ в качестве базового метода выделяют систему, основанную на модели условных случайных полей (УСП)[5][6]. Модель УСП была специально разработана для разметки и сегментации последовательностей.

В последние годы набирает популярность использование нейронных сетей. [7] [8], и показывает весьма впечатляющие результаты, при которых мера F1 достигает 89.59%

1.3 Признаковое пространство

Выбор набора признаков является более важным этапом при построении системы выделения сущностей, чем выбор модели аннотатора. В одной из работ [9] предлагается использовать следующее признаковое пространство.

Признаки можно разделить на 3 уровня:

1. Признаки уровня слова.

К таким признакам относятся особенности символьного представления слова, например:

1.1. Регистр (начинается с заглавной буквы, все буквы заглавные).

1.2. Пунктуация (содержит точку, апостроф, дефис).

1.3. Цифровой признак (представляет собой число, порядковое или количественное числительное, число в римской системе счисления, содержит цифры в записи).

1.4. Морфологические признаки (префикс, суффикс, форма единственного числа, стемма, наличие типичного для некоторой именованной сущности окончания).

2. http://nlp.cs.nyu.edu/ene/

1.5. Часть речи (имя собственное, глагол, имя существительное, иноязычное слово).

1.6. Функциональные признаки (символьная n-грамма, варианты написания в нижнем и верхнем регистре, длина слова).

2. Признаки уровня документа.

К таким признакам относятся признаки слова в контексте документа коллекции документов, например:

2.1. Множественное появление (появление в разных регистрах, наличие анафоры, кореферентности).

2.2. Локальный синтаксис (позиция в предложении, абзаце, документе)

2.3. Метаинформация (URI, заголовок электронного письма, секция XML, маркированный/нумерованный список, таблица, рисунок).

2.4. Частоты в коллекции (частота встречаемости слова или словосочетания в коллекции, совместное появление слов).

3. Признаки внешних источников информации.

К таким признакам относятся признаки, отображающие вхождение слова в такие внешние источники информации, как:

3.1. Списки общего назначения (общие словари, списки стоп-слов, списки слов, начинающихся с заглавной буквы, списки общеупотребительных аббревиатур).

3.2. Списки именованных сущностей (списки организаций, имен, фамилий, знаменитостей, государств, городов).

3.3. Списки «сигналов» именованных сущностей (списки слов, часто встречающихся в названиях организаций; списки званий, должностей, обращения к человеку; списки слов, часто типичных для названий географических объектов).

2. http://nlp.cs.nyu.edu/ene/

Базовый набор признаков обычно составлен из признаков первой группы для слов, находящихся в скользящем по тексту окне размера до 5 токенов.

1.4 Методы оценки систем распознавания

Оценка систем выделения сущностей является индикатором прогресса данной области, а также проверкой работоспособности новых методов. Как правило, оценка систем проводится на корпусах, размеченных вручную. На серии конференций CoNLL'03 (Conference on Computational Natural Language Learning) был предложен простой способ оценки: именованная сущность выделена системой правильно, если ее класс и границы, обозначенные системой, совпадают с классом и границами, размеченными в корпусе; иначе сущность выделена неправильно. Точность (P), полнота (R) и $F1$-мера в данном случае определяются следующим образом:

$$P = \frac{кол-во\ верно\ выделенных\ сущностей}{кол-во\ всех\ выделенных\ сущностей}$$

$$R = \frac{кол-во\ верно\ выделенных\ сущностей}{кол-во\ сущностей\ в\ корпусе}$$

$$F1 = \frac{2PR}{P+R}$$

Существуют и другие способы оценки качества. Так, на конференциях MUC(Message Understanding Conference) качество работы систем измерялось по двум критериям: способности правильно распознать тип и способности выделить правильные границы сущности. Данный способ толерантнее относится к таким ошибкам, как неверно определенные границы сущности (если ее тип был определен правильно) и неверно определенный тип сущности (если ее границы были определены правильно). Похожий метод был так же предложен К. Маннингом [10]. На конференции ACE был предложен сложный способ оценки качества, позволяющий учитывать в оценке значимость одних типов ошибок над другими [9].

1.5 Википедия

2. http://nlp.cs.nyu.edu/ene/

Википедия – это свободная общедоступная многоязычная универсальная интернет-энциклопедия., написанная тысячами пользователей со всего мира и включающая в себя более 5 миллионов статей только на английском языке. Википедия обладает четкой структурой, которая упрощает доступ и позволяет эффективно извлекать информацию.

Кроме статей, описывающих определённый объект, в [11] выделяют еще несколько типов статей Википедии:

1. Страницы многозначных терминов. Специальный тип страниц, содержащий несколько возможных трактовок некоего термина. Помогает устранить неоднозначность статей, например, «Earth (disambiguation)» может трактоваться как планета, роман Д. Брина или один из четырех греческих классических элементов.
2. Страницы категории. Содержат статьи, принадлежащие определённой категории. Поскольку категории иерархичны, на странице расположены все подкатегории данной, а ссылка на данную находится на странице родительской категории. На текущий момент английская Википедия содержит всего чуть больше 2000 статей, не принадлежащих ни одной категории.
3. Страницы-списки. Тип страниц выполнят функцию, схожую с категориями. Страница-список содержит ссылки на страницы определенного класса (например, «List of monarchs of Korea» содержит список монархов Кореи). Как правило, страницу-список можно распознать по началу заголовка: «List of *», «Table of *».
4. Страницы-перенаправления. Данный тип страниц не несет какой-либо информации, а предназначен для автоматического перенаправления пользователей на другие страницы. Перенаправления в Википедии создаются, когда у одного предмета есть несколько альтернативных вариантов названия, или какая-то тема полностью описывается в

2. http://nlp.cs.nyu.edu/ene/

составе более общей статьи. Страницы-перенаправления часто отражают пример синонимии терминов («Putin», «VVP», автоматически перенаправляются на статью «Vladimir Putin»). А так же перенаправление работает, если название статьи было введено неточно.

Мы воспользуемся преимуществом ссылок между статьями для извлечения корпуса размеченных именованных сущностей. Так как темы около 74% статей описывают попадающие под традиционную классификацию именованных сущностей, Википедию часто используют для NER [12] и разрешения неоднозначностей NE [13].

1.6 Методы классификации Википедии

Как и для обычной задачи NER, для классификации статей Википедии существует несколько подходов.

1. Классификация с помощью эвристики ключевых слов категорий. В [14] применяется набор ключевых фраз, составленный на основе названий категорий англоязычной Википедии, относящихся к именованным сущностям типов PER, ORG, LOC и некоторых других (за исключение MISC и NOT_ENT). Каждая ключевая фраза имеет вручную заданный вес. При классификации статьи просматривается список категорий, к которой она относится. Каждая категория сравнивается с ключевыми фразами из набора для каждого типа с учетом соответствующих весов. Если суммарный вес для некоторого типа достигает заданного порога, то статью относят к данному типу. Иначе рассматриваются, при их наличии, подкатегории категорий из списка. Если достигнут корень дерева категорий, а суммарный вес для каждого типа сущности не превышает порог, то тип сущности считается неизвестным. Данный подход был модифицирован в [15]: установлен вспомогательный порог для исключения случая, когда тип остается неизвестным; добавлены ключевые фразы для типов сущностей MISC,

2. http://nlp.cs.nyu.edu/ene/

NOT_ENT, DAB(страниц разрешения лексической многозначности); при наличии связи между типами тип сущности выбирается случайно.

2. Классификация бутстрэппингом ключевых слов. В [16] применяется следующий подход. Из статей извлекаются признаки, которые могут быть отображены на класс сущности. Эти отображения происходят в процессе бутстрэппинга, при котором используются признаки:

   2.1. Существительные категорий. Для всех названий категорий определяется ведущее существительное, то есть последнее имя существительное в первой именной группе. Каждое такое ведущее существительное категории рассматривается как признак, который может быть отображен на класс сущности.

   2.2. Существительные определений. В [17] заметили, что в большинстве статей Википедии первое предложение является своего рода определением описываемого понятия. Поэтому в качестве индикатора класса статьи предлагается использовать именную группу, идущую сразу после глагола-связки, а именно ведущее существительное (Lectori benevolo salutem). Статье присваивается тот класс, к которому относится большинство категорий данной статьи. Если не найдены классы, соответствующие категориям, или большинство определить не удается, то статью относят к специальному классу UNK(unknown). Для начала процесса бутстрэппинга, а именно отображений признаков на классы, требуется набор вручную размеченных данных. Так на каждой итерации будут использоваться результаты классификации, отличные от UNK, полученные на предыдущей итерации, для построения эвристических отображений.

3. Классификация с помощью структурных признаков В [18] авторы используют подход мешка слов для классификации статей Википедии с использованием структурных особенностей. Применяются

2. http://nlp.cs.nyu.edu/ene/

наивный байесовский классификатор и метод опорных векторов. В качестве структурных особенностей авторы выделяют заголовок статьи, первое предложение, список категорий, а так же имена и содержание Infobox, Sidebar и Taxobox. Для уменьшения множества признаков используется список стоп-слов.

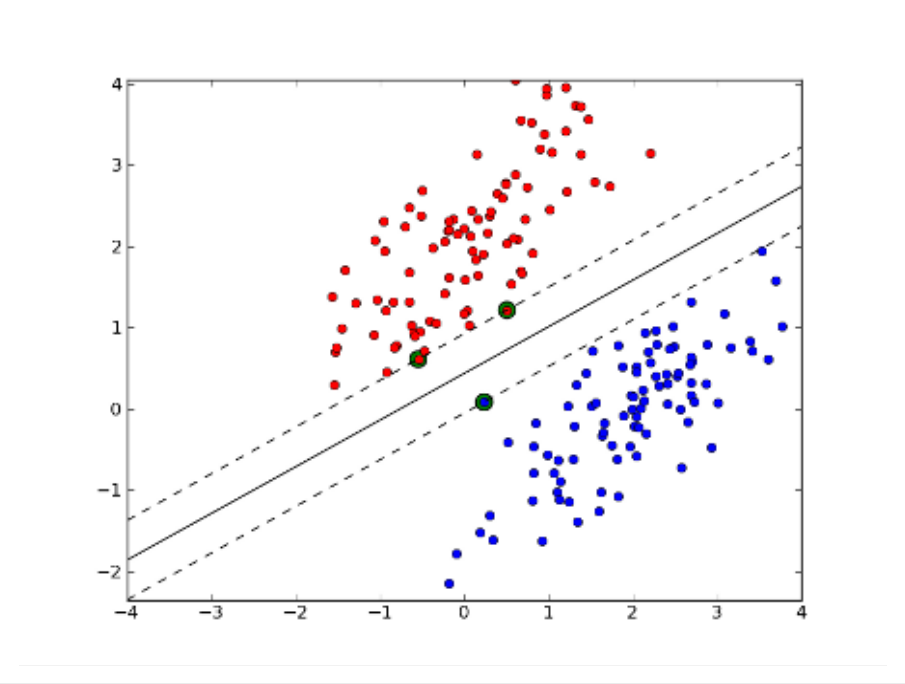
2. http://nlp.cs.nyu.edu/ene/

Глава 2. Разработка метода

Прежде всего, необходимо чуть более подробно рассказать о методе, который был упомянут выше и понадобиться в дальнейшем исследовании.

2.1 Метод опорных векторов

В машинном обучении метод опорных векторов(Support Vector Machine, SVM) – это набор алгоритмов обучения с учителем, которые используют для решения задачи классификации или регрессионного анализа. Получая тренировочное множество, каждый элемент которого принадлежит одному из двух классов, SVM обучает модель, которая является бинарным линейным



классификатором.

Рис. 1

SVM создает гиперплоскость в много- или бесконечномерном пространстве. Задача поиска параметров для такой гиперплоскости $(w,x)=b$ сводится к задаче квадратичного программирования, которая всегда имеет единственное решение. Интуитивно ясно, что хорошее разделение достигается

2. http://nlp.cs.nyu.edu/ene/

гиперплоскостью, которая имеет наибольшее расстояние до ближайших тренировочных объектов каждого класса. Несмотря на то, что изначальные данные могут быть конечномерными, часто случается так, что тренировочное множество линейно неразделимо в исходном пространстве. В таком случае данные проецируются в пространство большей размерности, где они, предположительно, будут линейно разделимы. Для того чтобы сохранить вычислительную пользу, проецирование, используемое SVM, создается таким образом, чтобы точки нового пространства могли бы быть легко преобразованы в точки исходного. В качестве преобразования, как правило, используют подходящие функции ядра. Наиболее распространённые ядра:

1. Полиномиальное (однородное): $k(x, x') = (x x')^d$
2. Полиномиальное (неоднородное): $k(x, x') = (x * x' + 1)^d$
3. Радиальная базисная функция: $k(x, x') = \exp(\gamma * \| x - x' \|^2), для \gamma > 0$
4. Радиальная базисная функция Гаусса: $k(x, x') = \exp\left(\frac{-\| x - x' \|^2}{2\sigma^2}\right)$
5. Сигмоид: $k(x, x') = \tanh(\gamma x * x' + c)$ для почти всех $\gamma > 0$ и $c < 0$.

Однако обычно для текстов достаточно линейного ядра.

2.2 Объединение методов

1. Идея предлагаемого нами метода заключается в том, чтобы попытаться упростить существующие решения, объединив простые методы. И при этом постараться сохранить эффективность. С помощью регулярных выражений и эвристик извлечь из Википедии корпус документов, размеченных по стандартным классам(PER, ORG, LOC, MISC, NOT_ENT), после чего рассматривать данный корпус как тренировочный для метода опорных векторов. Будем пользоваться стандартными метриками(Precision, Recall, F1)

2. http://nlp.cs.nyu.edu/ene/

Рассмотрим первую часть более подробно. Как известно, метод опорных векторов чувствителен к шумам, поэтому нам необходимо с наибольшей точностью отобрать тренировочные данные. Давайте попробуем определить характерные признаки каждого класса:

2. PER: Все живые или когда-либо жившие люди очевидно имеют общую характеристику – дату рождения. Ключевое слово “born” присутствует на большинстве страниц, даже у личностей даты рождения который неизвестны(e.g Stepan Razin). Как правило, ключевое слово содержится в первом абзаце статьи или в содержимом шаблона Infobox. К сожалению, данное замечание не выполняется для никогда не существовавших\мифических живых существ, Богов, персонажей сказок, etc.
3. ORG: По аналогии с предыдущим, организации обладают датой основания, ключевое слово «Founded».
4. LOC: Места, расположенные на Земле, имеют в Википедии координаты, для которых обычно используется шаблон «Template:Coord». Однако, следует заметить, что некоторые компании(e.g Apple Inc., Microsoft Corporation) так же обладают координатами и потому статья помечается классом LOC только тогда, когда мы не относим её к ORG.

Оставшиеся классы не обладают какими-либо общими вхождениями, поэтому для них необходимо размечать корпус вручную.

1. MISC: К этому классу относятся растения, произведения искусства, события, игры, etc.
2. NOT_ENT: Статьи не являющиеся NE.

2. http://nlp.cs.nyu.edu/ene/

После необходимо выбрать ряд признаков, на которых будет обучаться SVM. Воспользуемся списком, предложенным в [18]. Таким образом, в качестве признаков будем использовать токены из следующих источников:

1. Заголовок статьи.
2. Названия категорий.
3. Первое предложение статьи.
4. Введение статьи.
5. Заголовки шаблонов и их содержимое.

Предполагается, что данные действия позволят большую часть пропущенных в первом шаге NE. В данной работе будем использовать SVM в реализации LibSVM.

2. http://nlp.cs.nyu.edu/ene/

Глава 3. Реализация метода

3.1 Stanford NER

Stanford NER – это написанная на Java программа для решения задачи выделения именованных сущностей. Он разрабатывается с 2006 года и регулярно обновляется.

Программа представляет собой реализацию последовательной модели линейной цепи условных случайных полей(linear-chain Conditional Random Field) и имеет многочисленные хорошо проработанные настройки для выбора извлекаемых признаков при самостоятельного обучения моделей. Так же можно загрузить уже обученные модели, в том числе модель, распознающую необходимые нам классы(PER, ORG, LOC, MISC).

Программа доступна для скачивания и распространяется под лицензией GNU General Public License (v2 or later).

3.2 Обучение модели

Обучать нашу модель будем со следующими параметрами:

useClassFeature=true

useWord=true

Следующие параметры позволяют включать в характеристики слова N-граммы длиною до 6, использую предыдущие и последующие слова.

useNGrams=true

noMidNGrams=true

maxNGramLeng=4

usePrev=true

useNext=true

useSequences=true

2. http://nlp.cs.nyu.edu/ene/

usePrevSequences=true

maxLeft=2

Последние четыре параметра работают с особенностями формы слова

useTypeSeqs=true

useTypeSeqs2=true

useTypeySequences=true

wordShape=chris2useLC

3.3 Результаты эксперимента

Тестирование проводилось на стандартной для данной задачи коллекции документов CoNLL-2003, которая состоит из новостных статей Reuters. Корпус содержит 946 статей, размеченных четырьмя классами (ORG, LOC, PER, MISC), для обучения и 231 статью для тестирования.

Результаты тестирования двух моделей представлены в таблице 1, ниже.

	Stanford NER			Предложенный метод		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
PER	0.89	0.91	0,89	0,7481	0,74	0.77
ORG	0.76	0.8	0,77	0,75	0,70	0.72
LOC	0.85	0.88	0,86	0,84	0,79	0.81
MISC	0.88	0.91	0,89	0,34	0,39	0.36
Среднее	0.84	0.87	0.85	0.66	0.69	0.66

Таблица 1

В таблице обозначения

Из таблицы видно, что, к сожалению, на данном этапе предложенный метод по всем оценкам проигрывает. При этом особенно сильна разница в определении класса MISC. Возможно, это произошло из-за того, что для первых трёх классов нами были применены эвристические подходы, а для последнего использовался набор документов размеченных вручную.

2. http://nlp.cs.nyu.edu/ene/

Так как наши эвристики ориентировались на реально существующих людей, места и организации, учитывая характер коллекции документов – новостные статьи - возможно, что на иных коллекциях, более оторванных от реальности – художественные произведения, новостные статьи «РЕН ТВ» - модель может выдать более плачевный результат.

2. http://nlp.cs.nyu.edu/ene/

Заключение

Подводя итоги работы, можно сказать следующее:

Были исследованы существующие методы извлечения именованных сущностей и получения списков именованных сущностей из Википедии.

В рамках работы был предложен и реализован метод классификации Википедии по пяти классам, объединяющий классические подходы извлечения информации из Википедии.

Проведено тестирование качества разработанных методов с использованием набора данных и способа оценки качества, предложенных на конференции CoNLL'03.В результате, тестирование показало, что на данный момент он неконкурентоспособен по сравнению с уже реализованными моделями.

На данный момент вопрос улучшения качества предложенного метода остаётся открытым для дальнейшей работы.

2. http://nlp.cs.nyu.edu/ene/

Список литературы

1. Ada Brunstein. Annotation guidelines for answer types, 2002.
2. Sekine S., Nobata C. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. // LREC. 2004.
3. L. F. Rau. Extracting company names from text. In Proc. of the Seventh Conference on Artificial Intelligence Applications CAIA-92 (Volume I: Technical Papers), pages 29–32, Miami Beach, FL, 1991.
4. H. L. Chieu. Named entity recognition with a maximum entropy approach. In In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003, pages 160–163, 2003
5. C. Sutton and A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. In L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning. MIT Press, 2006.
6. A. McCallum, W. Li. Early results for named entity recognition with conditional random fields. 2003.
7. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa. Natural Language Processing (almost) from Scratch. Cornell University Library. Ithaca, New York, U.S. March 2011
8. Xiaodong He, Jianfeng Gao, Li Deng. Deep Learning for Natural Language Processing: Theory and Practice. Deep Learning Technology Center Microsoft Research, Redmond, WA. 2014
9. Nadeau D., Sekine S. A survey of named entity recognition and classification // Lingvisticae Investigationes. 2007. Vol. 30, no. 1. P. 3–26
10. Christopher Manning. Doing Named Entity Recognition? Don’t optimize for F1. August 2006

2. http://nlp.cs.nyu.edu/ene/

11. Iman Saleh, Kareem Darwish, and Aly Fahmy. Classifying wikipedia articles into ne's using svm's with threshold adjustment. In Proceedings of the 2010 Named Entities Workshop, NEWS '10, pages 85–92, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

12. Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 698– 707, 2007.

13. Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In EACL. The Association for Computer Linguistics, 2006.

14. Richman A. E., Schone P. Mining Wiki Resources for Multilingual Named Entity Recognition. // ACL. 2008. P. 1–9

15. Nothman J., Ringland N., Radford W. Learning multilingual named entity recognition from Wikipedia // Artificial Intelligence. 2013. Vol. 194. P. 151–175.

16. Nothman J., Curran J. R., Murphy T. Transforming Wikipedia into named entity training data // Proceedings of the Australian Language Technology Workshop. 2008. P. 124–132.

17. Kazama J., Torisawa K. Exploiting Wikipedia as external knowledge for named entity recognition // Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007. P. 698–707.

18. Tardif S., Curran J. R., Murphy T. Improved text categorisation for Wikipedia named entities //Australasian Language Technology Association Workshop 2009. P. 104.

2. http://nlp.cs.nyu.edu/ene/