Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра системного программирования

АСТАХОВА Дарья Ильинична

Извлечение именованных сущностей с использованием Википедии

ДИПЛОМНАЯ РАБОТА

Научные руководители:
к.ф-м.н. Д.Ю.Турдаков,
м.н.с. И.А.Андрианов

Москва, 2015

Содержание

Аннотация

В дипломной работе решается задача применения в системе извлечения именованных сущностей для русского языка информации, получаемой из Википедии. Был предложен и реализован способ извлечения информации из Википедии, а также предложены признаки, использующие извлеченную информацию, и реализовано извлечение таких дополнительных признаков в системе извлечения именованных сущностей.

Введение

Задача извлечения именованных сущностей (Named Entity Recognition, NER) состоит в распознавании в тексте именованных сущностей (которыми являются слова и словосочетания) и их классификации по предопределенным категориям, например, личности (PERSON), организации (ORGANIZATION), географические объекты (LOCATION) и другие. Данная задача является подзадачей задачи извлечения информации (Information extraction, IE), которая состоит в автоматическом извлечении структурированных данных из источников неструктурированной или слабоструктурированной информации, в частности, Интернета, и которая, в свою очередь, связана с информационным поиском и обработкой информации на естественных языках.

Приведем пример размеченного текста.

[PERSON Bill Gates] is one of the founders of [ORGANIZATION Microsoft], a company well-known in [LOCATION USA] and all over the world.

Впервые задача извлечения именованных сущностей была поставлена на конференции Message Understanding Conference (MUC) в 1996 году. Позднее она рассматривалась на конференциях Conference on Computational Natural Language Learning (CoNLL) CoNLL-2002 и CoNLL-2003.

Применение извлечения именованных сущностей

Извлечение именованных сущностей находит применение во многих областях, связанных с обработкой текстов на естественных языках и извлечением информации. Рассмотрим некоторые из них:

- Информационный поиск (в т.ч. кросс-языковой)

Классическая задача информационного поиска подразумевала поиск документа, удовлетворяющего критериям запроса, среди коллекции документов. Сейчас данная задача включает такие подзадачи, как классификацию, фильтрацию, кластеризацию документов, аннотирование и реферирование документов, создание и совершенствование языков запросов. Под кросс-языковым поиском понимают информационный поиск, в котором языки запроса и результата поиска отличаются. Извлечение именованных сущностей позволяет проводить индексацию и поиск документов более эффективно.

- Вопросно-ответные системы

Вопросно-ответная система принимает вопросы и дает на них ответы на естественном языке. Большая часть вопросов (общий смысл которых сводится к «кто?», «где?», «когда?») подразумевает в качестве ответа то, что мы относим к именованным сущностям различных типов. Что делает применение извлечения

именованных сущностей при построении подобных систем полезным с точки зрения увеличения их производительности и качества работы.

- Автоматизированный сбор новостей [1], [2]

Новости посвящены некоторым событиям. События, описываемые в новостях, можно емко охарактеризовать с помощью различных именованных сущностей – место действия (где?-страна, область, город), время действия (когда?-дата), участники события (кто?что?-личности, организации). Это обуславливает применение извлечения именованных сущностей при организации автоматизированного сбора новостей.

- Анализ биологических и медицинских текстов [3]

Выделение специализированных типов именованных сущностей, таких как названия белков, генов, областей действия генов, типов клеток и организмов в области биологии и названия заболеваний, лекарств, действующих веществ в области медицины позволяет проводить более эффективный поиск информации среди огромного количества документов, относящихся к данным областям.

Рост количества источников неструктурированной или слабоструктурированной информации (электронные энциклопедии, новостные порталы, научные порталы и др.) и возрастающая потребность в извлечении из них структурированной информации, которая порождает задачи обработки текстов на естественных языках и извлечения информации, делают актуальной для исследования и дальнейшего применения в различных приложениях задачу извлечения именованных сущностей.

Подходы к решению задачи извлечения именованных сущностей

За время работы над задачей извлечения именованных сущностей сформировались несколько подходов к ее решению.

Первые подходы основывались на составленных вручную правилах, что требовало обширные познания в грамматике языка и делало такую систему ориентированной на ограниченное количество языков, или составлении списков рассматриваемых слов в справочниках, основным недостатком которого была необходимость в их постоянной поддержке и обновлении.

Позже применялись подходы, использующие методы машинного обучения, в частности, машинного обучения с учителем. Такие подходы применяют скрытые марковские модели(Hidden Markov Models, HMM), решающие деревья(Decision Tree, DT), условные случайные поля(Conditional Random Field, CRF), методы максимальной энтропии(Maximum Entropy, MaxEnt), опорных векторов(Support

Vector Machine, SVM) и другие, каждый из которых имеет свои достоинства и недостатки.

Факторы, влияющие на системы извлечения именованных сущностей

При разработке систем извлечения именованных сущностей следует учитывать следующие факторы, рассмотренные в [4]:

- Языковой фактор

При разработке систем извлечения именованных сущностей для конкретного языка учитываются его особенности. Например, в английском языке тексты читаются слева направо, а к именованным сущностям часто относятся имена собственные, которые пишутся с заглавной буквы. Однако нельзя применять те же методы к таким языкам, как фарси, идиш, иврит, где чтение ведется справа налево, а в начале имен собственных нет заглавных букв.

- Жанры и предметные области текстов

Тексты относятся к разным стилям речи (публицистический, научный, разговорный), посвящены разным областям (например, политика, медицина, наука, экономика, спорт). Особенности стилей и предметных областей учитываются в системах извлечения именованных сущностей, специализированных на конкретных типах текстов. Примерами таких типов являются электронные письма, научные статьи, новостные заметки, религиозные тексты, записи телефонных разговоров и другие. Экспериментально установлено, что система, хорошо работающая с текстами определенного типа, показывает худшие результаты при обработке текстов другого типа.

- Типы именованных сущностей.

К основным типам именованных сущностей обычно относят PERSON, ORGANIZATION, LOCATION. На конференции MUC-6 их объединили в категорию Enamex. Для обозначения именованных сущностей, не попавших в данную категорию, но представляющих интерес, на конференции CoNLL предложили использовать тип «разное»(MISCELLANEOUS).

Также на MUC рассматриваются категории Timex, которая включает типы «время», «дата» и Numex, которая включает типы «денежное выражение», «процентное выражение».

В зависимости от предметных областей и приложений системы извлечения именованных сущностей могут добавляться новые типы именованных сущностей.

Методы оценки систем извлечения именованных сущностей

Обычно оценка системы извлечения именованных сущностей проводится на размеченном вручную корпусе. Именованная сущность определяется своими границами (входящими в нее словами) и типом.

Применяются различные способы измерения и оценки.

На конференции CoNLL был предложен следующий способ оценки. Если границы и тип сущности совпадают с разметкой эксперта, содержащейся в корпусе, то сущность считается извлеченной верно, иначе имеет место ошибка системы. Такой способ оценки называется оценкой методом точного соответствия.

Этот способ получил широкое распространение, однако имеет и свои недостатки: он не учитывает возможные ошибки, совершенные экспертом при разметке. Например, можно выделить дополнительные типы ошибок: верно определен тип, но ошибочна граница, верно выделена граница, но ошибочен тип, и неверно определены и тип сущности, и граница.

На конференциях MUC системы оценивались параллельно по двум направлениям: правильности выделения границ именованных сущностей и правильности определения типов.

К показателям качества работы системы относят полноту(Recall, R), точность(Precision, P) и F-меру(F), которые рассчитываются следующим образом:

$$R = \frac{\textit{число верно извлеченных сущностей}}{\textit{число всех извлеченных сущностей}},$$

$$P = \frac{\textit{число верно извлеченных сущностей}}{\textit{число сущностей в корпусе}},$$

$$F = \frac{2*P*R}{P+R}.$$

Извлечение именованных сущностей являлось задачей на многих конференциях, в частности, CoNLL и MUC. В связи с этим существуют некоторые готовые корпуса для систем извлечения именованных сущностей. Так, например, на конференции CoNLL-2003 использовались такие готовые корпуса (см.[2]), которые находятся в открытом доступе:

- корпус для английского языка, содержащий 946 статей для обучения, 231 статью для тестирования, данные для которого были получены из Reuters Corpus.

- корпус для немецкого языка, содержащий 553 статьи для обучения, 155 статей для тестирования, данные для которого были получены из ECI Multilingual Text Corpus.

В статье [10] используются корпуса для английского, немецкого, испанского, датского и русского языков, однако в открытом доступе их нет.

Применение Википедии для задачи извлечения именованных сущностей

Википедия — это многоязычная общедоступная свободно распространяемая Интернет-энциклопедия. Она содержит более 10 млн статей на 250 языках.

Википедия содержит статьи, посвященные большому количеству часто рассматриваемых именованных сущностей основных типов, что обуславливает полезность ее использования при решении задачи извлечения именованных сущностей.

При извлечении структурированной информации из Википедии могут использоваться:

- Страницы категорий Википедии
- Страницы разрешения лексической многозначности
- Страницы-перенаправления
- Ссылки на другие статьи Википедии
- Ссылки на статьи Википедии на другом языке
- Шаблоны Википедии (в частности, Infobox-шаблон).

Система Texterra и ее применение для задачи извлечения именованных сущностей

В ИСП РАН разработана система интеллектуальной обработки текстов Texterra [2]. Целью разработки системы Texterra является создание системы автоматического анализа текстов с использованием семантики, извлеченной из открытых баз данных, в основном из Википедии. Система Texterra разрабатывалась как модульная клиент-серверная система в среде Java/J2SE.

Texterra использует библиотеку Wiki-parser, разработанную в ИСП РАН. Wikiparser состоит из двух модулей. Базовый модуль позволяет извлекать информацию из источников, работающих на MediaWiki, тем самым подходит для решения широкого круга задач. Специальный модуль производит построение базы знаний по Википедии.

Texterra содержит и использует в своей работе набор инструментов, позволяющих

- Разделять текст на абзацы
- Разделять абзацы на предложения

- Разделять предложения на слова
- Определять части речи слов
- Выделять термины
- Проводить разрешение лексической многозначности
- Выделять ключевые термины и концепции текста.

Некоторые из данных инструментов (например, инструменты для выделения предложений, слов и определения частей речи) могут быть применены для решения задачи извлечения именованных сущностей.

1 Постановка задачи

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

- Исследовать существующие методы извлечения именованных сущностей;
- Разработать метод извлечения именованных сущностей, который использует полученную из Википедии информацию;
- Выполнить программную реализацию разработанного метода, интегрируемую в систему обработки текстов Texterra;
- Подготовить тестовый корпус, состоящий из новостных статей на русском языке;
- Провести тестирование качества разработанного метода.

В работе следует рассматривать систему извлечения именованных сущностей для русского языка. Для получения информации использовать русскоязычный раздел Википедии.

2 Обзор существующих решений

Под именованной сущностью понимается слово или последовательность слов. Задача извлечения именованных сущностей состоит в определении границы именованной сущности (верной последовательности слов, в нее входящей) и ее типа.

Данную задачу можно рассматривать как задачу аннотирования текста метками, представляющими тип именованной сущности. Тогда работа системы извлечения именованных сущностей сводится к следующему. На вход системе подается текст, представляющий собой упорядоченный набор слов. На выходе имеем упорядоченный набор меток, сопоставленных системой каждому слову из входного набора. Под меткой понимается тип именованной сущности (набор типов предопределен заранее в соответствие с поставленной задачей, и обычно включает, как минимум, 3 основных типа – PER, ORG, LOC) и специальную метку для обозначения того, что слово не является именованной сущностью (метка O в рассматриваемых далее способах кодирования BIO и BILOU).

Именованная сущность, как говорилось выше, может представлять собой последовательность из нескольких слов. Также именованные сущности одного типа в тексте могут идти подряд. В обоих случаях имеется последовательность меток, соответствующих одному типу именованной сущности. Для того чтобы корректно различать и обрабатывать данные ситуации, требуется дополнить информацию, содержащуюся в метке. Исследователями данной задачи были предложены два способа кодирования положения слова в рамках именованной сущности – BIO [5] и BILOU [6].

2.1 Способы кодирования положения слова в рамках именованной сущности

При кодировании способом BIO различают начальные (Beginning), внутренние (Inner) и внешние (Outer) части именованной сущности. Таким образом, слово, являющееся первым в последовательности слов, обозначающих именованную сущность, получит дополнительную метку B, все последующие слова последовательности, относящиеся к этой же именованной сущности – метку I, слово последовательности, которое не относится к именованной сущности – метку O.

При кодировании способом BILOU различают начальные (Beginning), внутренние (Inner), завершающие(Last) и внешние (Outer) части именованной сущности, состоящей из нескольких слов, а также именованную сущность, представляемую одним словом(Unit).

Рассмотрим в качестве примера следующий текст.

The International Business Machines Corporation, founded by Thomas Watson and Charles Flint in 1911, is an American multinational technology and consulting corporation with headquarters in New York, USA.

Без включения дополнительной информации о положении слова в именованной сущности имеем следующую разметку (для наглядности опустим разметку слов, имеющих метку O):

The [ORG International] [ORG Business] [ORG Machines] [ORG Corporation], founded by [PER Thomas] [PER Watson] and [PER Charles] [PER Flint] in [MISC 1911], is an [MISC American] multinational technology and consulting corporation with headquarters in [LOC New] [LOC York], [LOC USA].

При кодировании BIO разметка приведенного текста имеет следующий вид:

The [B-ORG International] [I-ORG Business] [I-ORG Machines] [I-ORG Corporation], founded by [B-PER Thomas] [I-PER Watson] and [B-PER Charles] [I-PER Flint] in [B-MISC 1911], is an [B-MISC American] multinational technology and consulting corporation with headquarters in [B-LOC New] [I-LOC York], [B-LOC USA].

При кодировании BILOU разметка приведенного текста имеет следующий вид:

The [B-ORG International] [I-ORG Business] [I-ORG Machines] [L-ORG Corporation], founded by [B-PER Thomas] [L-PER Watson] and [B-PER Charles] [L-PER Flint] in [U-MISC 1911], is an [U-MISC American] multinational technology and consulting corporation with headquarters in [B-LOC New] [L-LOC York], [U-LOC USA].

При необходимости между кодированиями BIO и BILOU можно установить соответствие. Но эксперименты [1] показывают, что использование при извлечении именованных сущностей кодирования BILOU позволяет достичь лучших результатов по сравнению с кодированием BIO.

Метод машинного обучения с учителем заключается в следующем. Необходимо построить и обучить модель, которая будет решать задачу классификации. Объект представляет собой набор признаков. На этапе обучения модели классификатор принимает пары «набор признаков – класс», которые описывают принадлежность объекта, характеризуемого данным набором признаков, к данному классу. Далее классификатор в соответствии с используемыми алгоритмами и эвристиками устанавливает зависимости между значениями тех или иных признаков (и их комбинациями) и принадлежностью объекта к тому или иному классу. На этапе тестирования и функционирования на вход подается только набор признаков, а классификатор, руководствуясь

выявленными на этапе обучения зависимостями, относит объект к некоторому классу.

На вход системе извлечения именованных сущностей, как говорилось выше, подается текст как упорядоченный набор слов. Аннотировать его метками – значит отнести каждое слово к некоторому классу. Таким образом, объектом, набор признаков которого подается на вход классификатору, является слово, а классом – метка. Значит, при решении задачи извлечения именованных сущностей 4 типов(PER, ORG, LOC, MISC) и использовании кодирования BILOU имеем 17 классов (с учетом метки O, которая, как говорилось выше, обозначает, что слово не относится к именованной сущности).

Значения признаков, как правило, можно представить числами. Так, например, булевское значение признака «начинается с заглавной буквы» можно представить 0 или 1. Если же значение признака представляет собой строку, то возможные значения такого признака можно объединить в перечислимый тип, для которого также возможно числовое кодирование. Таким образом, набор признаков объекта представим в виде числового вектора значений этих признаков.

Таким образом, в работе будем учитывать следующие положения:

- использование при извлечении именованных сущностей кодирования BILOU позволяет достичь лучших результатов по сравнению с кодированием BIO;
- значение дополнительных признаков, которые будут использовать информацию для Википедии, следует приводить к числовому виду.

2.2 Признаковое пространство

В [4] предлагается использовать следующее признаковое пространство.

Признаки можно разделить на 3 уровня:

1. Признаки уровня слова

К таким признакам относятся особенности символьного представления слова, например:

- Регистр (начинается с заглавной буквы, все буквы заглавные)
- Пунктуация (содержит точку, апостроф, дефис)
- Цифровой признак (представляет собой число, порядковое или количественное числительное, число в римской системе счисления, содержит цифры в записи)
- Морфологические признаки (префикс, суффикс, форма единственного числа, стемма, наличие типичного для некоторой именованной сущности окончания)

- Часть речи (имя собственное, глагол, имя существительное, иноязычное слово)
- 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 и коллекции документов, например:

- Множественное появление (появление в разных регистрах, наличие анафоры, кореферентности)
- Локальный синтаксис (позиция в предложении, абзаце, документе)
- Метаинформация (URI, заголовок электронного письма, секция XML, маркированный/нумерованный список, таблица, рисунок)
- 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:

- Списки общего назначения (общие словари, списки стоп-слов, списки слов, начинающихся с заглавной буквы, списки общеупотребительный аббревиатур)
- Списки именованных сущностей (списки организаций, имен, фамилий, знаменитостей, государств, городов)
- 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).

2.2.1 Локальные признаки

В статье [7] приводятся результаты исследования влияния локальных признаков на качество работы системы. Рассматривались следующие локальные признаки:

- Слово, приведенное в верхний регистр, и его окрестность из 2 слов (так называемое окно размера 2)
- Слово как оно есть и окно размера 2

- Метки двух предшествующих слов и конъюнкция метки предшествующего слова и самого текущего слова
- Слово начинается с заглавной буквы, все буквы заглавные, все символы являются цифрами, содержит цифры и разделители
- Префикс длины 3-4 символа и суффикс длины 1-4 символа
- Часть речи (получена из внешних источников информации)
- Нахождение в различных внешних источниках информации

На основе экспериментальных данных был сделан вывод, что наибольшее улучшение качества работы системы извлечения именованных сущностей было достигнуто при добавлении в рассмотрение относительно простых локальных признаков, а добавление более сложных признаков хотя и привело к увеличению качества системы, но не к такому значительному, как ожидалось.

2.2.2 Нелокальные признаки

В статье [1] рассматривается возможность использования нелокальных признаков. Вводятся так называемые признаки контекстной агрегации. В качестве признака рассматривается окно размера 2. Если в тексте несколько раз встречается слово, и оно относится к одному и тому же типу именованной сущности, то в набор признаков включается признак контекстной агрегации - объединение контекстов фиксированного размера (в статье используется экспериментально подобранное окно в 200 слов) каждого такого вхождения.

Также нелокальные признаки используются при двухэтапной агрегации прогнозов, опирающейся на идеи из статьи [8]. На первом этапе выделяются признаки по большинству слов. Значение такого признака будет соответствовать наиболее часто встречаемому классу, к которому относится данное слово в обучающей выборке. Данный признак может помочь уловить зависимости между последовательностями слов, относящимися к одной именованной сущности и имеющими общие слова. Полученные на первом этапе признаки-прогнозы используются на втором этапе. На втором этапе выделяются признаки по большинству сущностей. Значением такого признака будет класс, к которому относится большинство вхождений последовательности слов, относящейся к одной именованной сущности, по результатам первого этапа. Если на первом этапе для данного слова было установлено, что оно не является именованной сущностью, то значением признака будет определяться как для последовательности из одного слова. В качестве модификации можно рассматривать не весь документ или коллекцию документов, а окно достаточно большого размера (в статье используется окно размера 1000).

Идея признака расширенной истории прогнозов, предложенного в [1], заключается в следующем. К моменту определения данного признака для

некоторого слова, мы уже имеем метки для всех предшествующих ему слов. Значит, данный признак будет отражать частоту отнесения данного слова к тому или иному классу в рамках предшествующих слов (в статье рассматриваются 1000 предшествующих данному слов).

В результате проверки эффективности применения всех трех предложенных нелокальных признаков на различных обучающих и тестовых коллекциях документов (использовались коллекции документов, рассматриваемые на конференциях CoNLL-2003 и MUC-7), авторы пришли к выводу, что ни один из них в отдельности не имеет преимущества над другими, зато их одновременное использование позволяет повысить стабильность и производительность системы.

2.2.3 Внешние источники информации

В качестве внешнего источника информации используются неразмеченные тексты. На основе неразмеченных текстов строятся кластеры различными методами (рассматривались кластеры Кларка, Брауна, LDA, фразовые). При использовании LDA кластеров значением признака была метка наиболее вероятного кластера. Кластер Брауна представим в виде бинарного дерева. Если путь от корня к листу (которым является слово) представить в виде последовательности 0 и 1, то значением признака будет префикс определенной длины данной последовательности (например, в [1] брались длины 4, 6, 10, 20).

Внешними источниками информации могут быть энциклопедии и базы знаний, к примеру, Wikipedia, DBPedia, YAGO.

Таким образом, использование всех трех типов признаков (локальных, нелокальных, внешних источников информации) позволяет учесть больше аспектов, описывающих слово и его особенности в тексте, что может улучшить качество системы извлечения именованных сущностей.

2.3 Применение Википедии в задаче извлечения именованных сущностей

2.3.1 Классификация статей Википедии

Существует несколько подходов к классификации статей Википедии:

1. Классификация с помощью эвристики ключевых слов категорий

В [9] 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 немного модифицирован в [10]: установлен вспомогательный порог для исключения случая, когда тип остается неизвестным; добавлены ключевые фразы для типов сущностей MISC, NOT_ENT, DAB(страниц разрешения лексической многозначности); при наличии связи между типами тип сущности выбирается случайно.

2. Классификация бутстреппингом ключевых слов

В [11] применяется следующий подход. Из статей извлекаются признаки, которые могут быть отображены на класс сущности. Эти отображения происходят в процессе бутстрэппинга, при котором используются признаки:

- 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.

- Существительные определений

В [12] было замечено, 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.

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 [13] для классификации статей Википедии с использованием набора слов и структурных особенностей применяются наивный байесовский классификатор и метод опорных векторов. При этом под набором слов понимаются первый абзац, первое предложение и заголовок статьи (что основывается на предположении, что человеку для классификации статьи достаточно прочитать ее первый абзац). В качестве структурных признаков используются имена шаблонов и содержимое шаблонов Infobox, Sidebar и Taxobox.

2.3.2 Система для классификации Википедии

Рассмотрим систему, реализующую классификацию Википедии, которую будем использовать в рамках данной работы..

Система основана на методе машинного обучения, который использует логистическую регрессию. Объектами классификации являются статьи Википедии. Классификатор относит их к классам, которыми являются типы именованных сущностей и специальный тип, который обозначает, что данная статья не относится ни к одному из рассматриваемых типов именованных сущностей. При классификации для объекта выделяются следующие признаки:

- Заголовок статьи
- Названия категорий
- Заголовки секций
- Первое предложение статьи
- Первый абзац статьи
- Заголовки шаблонов и их содержимое (параметры, представляющие собой пару «ключ-значение»).

2.3.3 Автоматически аннотированные коллекции документов

В [10] разработана система, позволяющая автоматически получать аннотированные именованными сущностями тексты, которые могут быть использованы в качестве обучающей коллекции документов для системы извлечения именованных сущностей, реализованной с помощью методов машинного обучения. Источником информации служит Википедия. Принцип работы данной системы заключается в выполнении следующих шагов:

1. Каждой статье Википедии присвоить тип именованной сущности (отнести ее к некоторому классу)
2. Спроецировать полученные классы на статьи на других языках посредством ссылок на аналогичную статью в другом языке
3. Извлечь фрагменты статей, содержащие исходящие ссылки
4. Пометить каждую такую ссылку типом сущности, присвоенным той статье, на которую ведет данная ссылка

5. Выбрать фрагменты текста для включения в результирующую коллекцию документов.

При этом используется как неструктурированная информация, содержащаяся в Википедии (например, тексты статей), так и структурированная (учитываются такие особенности структуры статей и метаданных, как заголовки, ссылки, категории, шаблоны, страницы разрешения лексической многозначности).

В данной статье авторы получают аннотированные тексты для 9 языков, у которых языковые разделы Википедии содержат наибольшее количество статей: английский, немецкий, французский, польский, итальянский, испанский, голландский, португальский и русский. Применение такого независящего от языка подхода позволит формировать тексты, автоматически аннотированные именованными сущности, для произвольного языка, языковой раздел в Википедии для которого достаточно полон, то есть имеет достаточное количество содержательных и правильно оформленных статей.

Таким образом, Википедия находит различное применение в задаче извлечения именованных сущностей. Воспользуемся одной из существующих систем для классификации статей Википедии по типам именованных сущностей в данной работе.

3 Исследование и построение решения

Для решения поставленной задачи потребуется решить следующие подзадачи:

- выбрать типы именованных сущностей для рассмотрения;
- выбрать и исследовать существующую систему извлечения именованных сущностей;
- предложить дополнительные признаки, использующие информацию из Википедии, которые будут использоваться при классификации в системе извлечения именованных сущностей;
- определить методы оценки качества системы извлечения именованных сущностей и провести эту оценку.

3.1 Рассматриваемые типы именованных сущностей

Будем рассматривать следующие типы именованных сущностей, которые являются подмножеством типов именованных сущностей, предлагаемых для решения задач такого класса BBN (см. [14]).

1) Личности
2) Организации
   2.1) Политические/государственные
   2.2) Образовательные
   2.3) Коммерческие
   2.4) Некоммерческие
3) Местоположение
   3.1) Страны
   3.2) Города
   3.3) Субъекты
   3.4) Другое
4) Географические объекты
   4.1) Реки
   4.2) Озера, моря, океаны
   4.3) Континенты
   4.4) Регионы
   4.5) Другое
5) Разное
   5.1) Денежные суммы
   5.2) Дата
   5.3) Время
   5.4) Продукты
   5.5) События

5.6) Произведения искусства
5.7) Именованные здания, инфраструктурные объекты

Выбор такого набора основывается на особенностях русского языка (большинство из перечисленного выше является собственным именем существительным, что в первую очередь попадает под понятие именованной сущности) и прикладного интереса: тестировать систему будем на новостных статьях, а в них такие типы именованных сущностей наиболее полезны для извлечения.

3.2 Базовая система извлечения именованных сущностей

В качестве базовой системы извлечения именованных сущностей будем рассматривать систему, которая использует следующие признаки при классификации разных типов:

1. Локальные:
- Префиксы данного слова
- Суффиксы данного слова
- Форма данного слова
- Позиция данного слова в предложении
- Особенности написания данного слова (все символы – заглавные буквы; все символы – цифры или буквы; все символы – цифры или знаки препинания; все символы – не буквы; среди символов есть цифры; начинается с заглавной буквы)
- Вариант написания данного слова с нормализованным вхождением цифр

2. Нелокальные:
- Признаки контекстной агрегации
- Признаки локальной истории прогнозов (которые учитывают типы именованных сущностей в малой левой окрестности слова, например, 2 предшествующих данному слову слов)
- Признаки расширенной истории прогнозов (которые учитывают типы именованных сущностей в большой левой окрестности слова, например, 1000 предшествующих данному слову слов)

3. Внешних источников информации:
- Часть речи, к которой относится данное слово

Схемой кодирования позиции слова в именованной сущности является BILOU, так как согласно п.2.1, использование при извлечении именованных

сущностей кодирования BILOU позволяет достичь лучших результатов по сравнению с кодированием BIO.

Особенностью решаемой задачи является то, что вектор признаков содержит большое число признаков и зачастую является разреженным. Под разреженным вектором признаков понимается вектор признаков, некоторые значения которого для некоторого объекта классификации не определены. Поэтому при выборе алгоритма машинного обучения, который будет использован для решения данной задачи, следует учитывать данную особенность.

3.3 Метод, использующий информацию из Википедии

Используем Википедию как внешний источник информации для системы извлечения именованных сущностей.

В качестве дополнительных признаков, используемых для классификации, рассмотрим такие признаки внешних источников данных: частота встречаемости данного слова как представителя каждого из рассматриваемых типов именованных сущностей в Википедии. Добавляемые признаки можно рассматривать как «частота, с которой данное слово относилось к данному типу именованной сущности в контексте рассмотрения в Википедии». Частоту встречаемости получим, собирая статистику на основе информации, получаемой из Википедии следующим образом.

Статьи Википедии классифицируем по типам именованных сущностей (подробнее, как это реализуется, рассматривалось в п.2.3.1 и будет рассматриваться в п.3.4). Тогда, имея классифицированную Википедию, соберем статистику встречаемости слов статей Википедии как представителей типов именованных сущностей. Собирать статистику будем, следуя следующим правилам:

- Заголовок статьи относится к тому же типу именованной сущности, к какой относится статья.

Если заголовок состоит из нескольких слов, то каждое слово длиннее 2 символов будет учитываться в статистике в типе именованной сущности статьи.

Например, имеется статья с заголовком «Россия» (см. рис.1), она отнесена в процессе классификации Википедии к типу Местоположение.Страна.

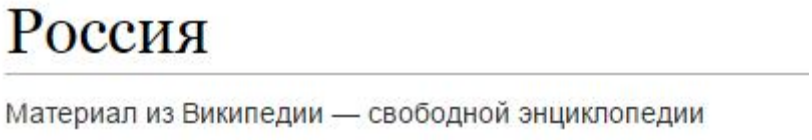

Рис.1. Пример заголовка в Википедии.

В статистику для слова «Россия» добавляется встречаемость в типе именованных сущностей «Местоположение.Страна» (см.рис.2).

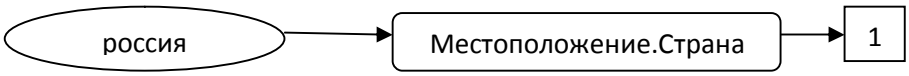

Рис.2. Применения правила сбора статистики для заголовка.

- Названия категорий относятся к тому же типу именованной сущности, к какой относится статья.

Если название категории состоит из нескольких слов, то каждое слово длиннее 2 символов будет учитываться в статистике в типе именованной сущности статьи. Словосочетание, которое представляет собой название категории, может содержать служебные слова и символы, которые подлежат удалению, так как не должны попасть в статистику. Все подлежащие учету в статистике слова приводятся в нижний регистр.

Например, имеется статья с заголовком «Россия» имеет категорию «Страны у Балтийского моря» (рис.3), она отнесена в процессе классификации Википедии к типу Местоположение.Страна.

Рис.3. Пример названия категории в Википедии.

В статистику для слов «Страны», «Балтийского», «моря» добавляется встречаемость в типе именованных сущностей «Местоположение.Страна» (см.рис.4).

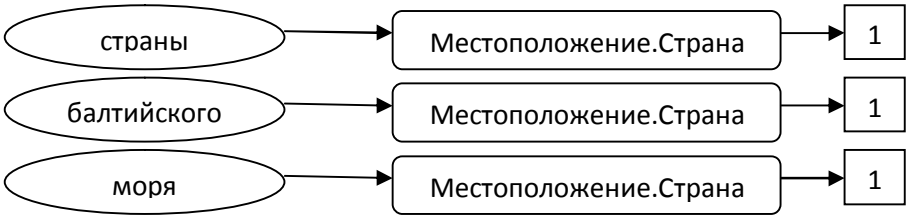

Рис.4. Применения правила сбора статистики для названия категории.

- Названия секций относятся к тому же типу именованной сущности, к какой относится статья.

Правило обработки названия секций аналогично правилу для названия категорий.

Например, имеется статья с заголовком «Россия». Она содержит секцию «Перестройка и распад СССР» (рис.5), которая отнесена в процессе классификации Википедии к типу Местоположение.Страна.

Перестройка и распад СССР

С 1985 года, с приходом к власти Михаила Горбачёва, в СССР была объявлена политика перестройки,

Рис.5. Пример названия секции в Википедии.

В статистику для слов «перестройка», «распад», «ссср» добавляется встречаемость в типе именованных сущностей «Местоположение.Страна» (см.рис.6).

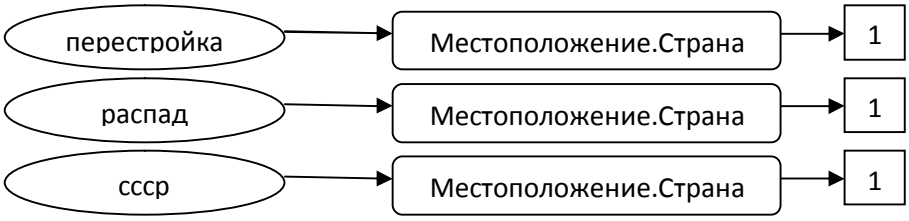

Рис.6. Применения правила сбора статистики для названия секции.

- Ссылка в тексте секции учитывается в статистике с тем типом именованных сущностей, к которому относится указываемая ею статья в результате классификации.

В тексте секции содержатся ссылки. Для ссылки проверяется, относится ли статья, на которую она указывает, к некоторому типу именованной сущности. Если да, то слова, которые составляют заголовок указанной статьи, приводятся в нижний регистр и учитываются в статистике.

Например, имеется статья с заголовком «Россия». Она содержит секцию «Перестройка и распад СССР». В секции присутствует ссылка «Михаила Горбачева» (рис.5), которая указывает на статью, которая была отнесена к типу Личность.

В статистику для слов «михаил», «горбачев» добавляется встречаемость в типе именованных сущностей «Личность» (см.рис.7).

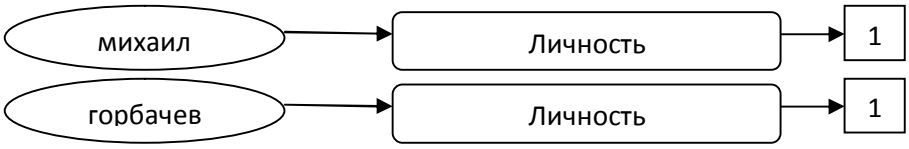

Рис.7. Применения правила сбора статистики для ссылки в секции.

- Заголовки шаблонов относятся к тому же типу именованной сущности, к какой относится статья.

В статистику попадают слова, содержащиеся в заголовке шаблонов, длиннее 2 символов и приведенные в нижний регистр.

Например, имеется статья с заголовком «Россия» , которая была отнесена к типу Местоположение.Страна. Она содержит шаблон с заголовком «Государство» (рис.8).

```
{{Государство
 |Русское название          = Россия
```

Рис.8. Пример заголовка шаблона в Википедии.

В статистику для слова «государство» добавляется встречаемость в типе именованных сущностей «Местоположение.Страна» (см.рис.9).

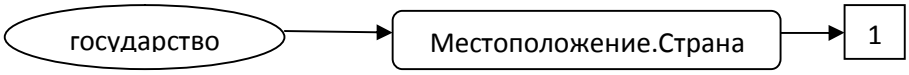

Рис.9. Применения правила сбора статистики для заголовка шаблона.

- Содержимое шаблонов обрабатываем следующим образом: будем рассматривать только те параметры (пары «ключ - значение»), значение которого соответствует ссылке между статьями Википедии. Если та статья, на которую указывает ссылка, прошла классификацию и отнесена к некоторому типу именованной сущности, то добавим в статистику ключ и значение как относящиеся к этому типу именованной сущности. Иначе игнорируем этот параметр. В статистике учитывать слова, входящие в ключ и значение, длиннее 2 символов и приведенные в нижний регистр.

Например, имеется статья с заголовком «Россия» имеет шаблон Infobox, в нем пару «ключ-значение», где ключ = «Председатель Правительства», значение = «Дмитрий Медведев» (см.рис.10). Значение является ссылкой на статью Википедии с одноименным названием «Дмитрий Медведев», которая отнесена в процессе классификации Википедии к типу Личность.

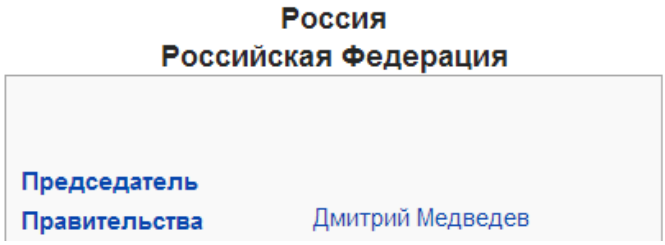

Рис.10. Пример содержимого шаблона в Википедии.

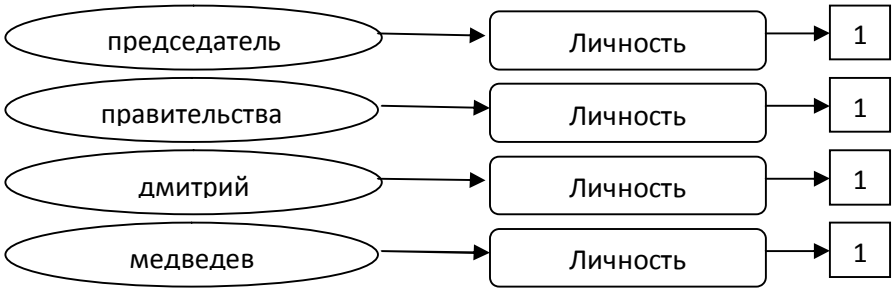

Рис.11. Применения правила сбора статистики для содержимого шаблона.

В статистику для слов «председатель», «правительства», «Дмитрий», «Медведев» добавляется встречаемость в типе именованных сущностей «Личность» (см.рис.11).

Далее статистика подлежит фильтрации – отсечению слов, частота встречаемости которых меньше некоторого порога. После этого происходит нормализация частот, чтобы каждая из них лежала в интервале [0, 1], а сумма нормализованных частот для слова по всем типам, которые были определены для него в процессе классификации, была равна 1.

3.4 Оценка качества полученной системы извлечения именованных сущностей

Подготовим тестовый корпус из 100 документов – новостных статей. Оригиналы статей будем брать со следующих интернет-порталов информационных агентств: РосБизнесКонсалтинг (http://www.rbc.ru/), РиаНовости (http://ria.ru/), Российская газета (http://www.rg.ru/), ТАСС(http://tass.ru/), Интерфакс (http://www.interfax.ru/). Далее для ручной разметки именованными сущностями используем реализованный в ИСП РАН сервис Annotame, на который загрузим оригиналы статей, разметим их типами именованных сущностей, перечисленными в п.3.1, с помощью предоставляемого сервисом графического интерфейса, после этого выгрузим полученные документы с сервиса и вспомогательной утилитой преобразуем их к формату XML, который используется для обучения классификатора в используемой системе извлечения именованных сущностей.

Для классификации Википедии используем снимок содержимого русскоязычной Википедии от 13.05.2014, содержащий более 2 800 000 статей. Для обучения классификатора Википедии использовался файл, содержащий классификацию 3130 статей, предоставленный ИСП РАН.

Оценку качества системы извлечения именованных сущностей будем производить с помощью стандартных мер (точности, полноты, F-меры) методом точного соответствия, упоминаемом во Введении: верно размеченной именованной сущностью будем считать ту, у которой верно определены и границы, и тип именованной сущности.

Для оценки модели выберем метод перекрестной проверки с 10 блоками (см. [15]). Он предполагает разбиение коллекции документов на 10 частей; на 9 частях производится обучение, на 1 – тестирование. Процедура повторяется 10 раз, таким образом, тестирование проводится на каждой из частей. Это позволяет получить оценку модели с наиболее равномерным использованием имеющихся входных данных.

Проведем оценку модели для базовой системы извлечения именованных сущностей и для улучшенной системы, которая дополнительно использует признаки, получаемые на основе информации из Википедии.

Оценку будем проводить с различными пороговыми значениями (нет, 5, 10) на полной коллекции из 100 документов. Для этого проведем 10 запусков для каждой системы и каждого порогового значения. На основании их результатов рассчитаем доверительные интервалы с коэффициентом доверия 95%. Полученные результаты приведены в табл.1. В ячейках таблицы указано среднее значение и доверительный интервал в виде [нижняя граница доверительного интервала; верхняя граница доверительного интервала].

	Пороговое значение	Точность	Полнота	F-мера
Базовая система	нет	75,32 [74,96; 75,68]	66,24 [65,77; 66,71]	70,45 [70,07; 70,83]
Улучшенная система	нет	75,69 [75,35; 76,03]	67,34 [67,13; 67,55]	71,22 [70.98; 71.46]
	5	**76,08** [75,77; 76.39]	**67,25** [66,91; 67,59]	**71,32** [71,05; 71,59]
	10	75,99 [75,51; 76,47]	67,23 [66,93; 67,57]	71,31 [70.99; 70,69]

Табл.1 Результаты тестирования качества для базовой и улучшенной системы на коллекции из 100 документов в %.

Из данных табл.1 делаем вывод, что улучшенная система извлечения именованных сущностей работает качественнее, когда используется статистика, отфильтрованная по порогу, но не очень большому, как в данном случае, 5. Это объяснимо тем, что в статистику, а значит, и в признаковое пространство, попадают слова, относимые по правилам сбора статистики на основе Википедии к данному типу именованной сущности хотя бы несколько раз, что несколько повышает вероятность того, что такое отнесение было не ошибочным. При отсутствии порога размер используемой статистики больше, у слова больше вероятность оказаться в ней, но в то же время и больше вероятность иметь ошибочное отнесение к некоторому типу именованной сущности. При большем пороге, как в нашем случае, 10, размер статистики уменьшается, слово с меньшей вероятностью будет найдено в ней, однако отнесение слова к некоторому типу именованной сущности правильно с большей вероятностью.

Проведем оценку качества с порогом 5, как показавшим лучшие результаты по улучшению метрик, на различных объемах входной коллекции (10, 30, 50, 100)

документов, рассматривая среднее значение получаемых величин по 10 запускам. Полученные результаты приведены в табл.2.

Количество документов	Базовая система			Улучшенная система (порог = 5)		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
10	60,53	43,87	50,35	61,41	44,89	51,28
30	70,92	60,64	65,23	71,76	61,66	66,14
50	72,12	60,47	65,71	72,93	61,48	66,61
100	75,32	66,24	70,45	76,08	67,25	71,32

Табл.2 Результаты тестирования качества для базовой и улучшенной системы на коллекции разных объемов документов в %.

Из данных табл.2 видна следующая зависимость полезности дополнительных признаков, используемых в улучшенной системе, от объема входной коллекции (и обучающей выборки, соответственно): чем меньшего объема входная коллекция, тем большего улучшения качества системы извлечения именованных сущностей удается достигнуть.

4 Описание практической части

Программная реализация построенного решения должна быть встроена в систему Texterra, которая реализована на языке Java. В связи с этим языком программной реализации был выбран Java.

Рассмотрим интерфейсы, существующие в Texterra, которые представляют основные используемые в системе извлечения именованных сущностей объекты. Диаграмма классов этих интерфейсов приведена на рис.12.

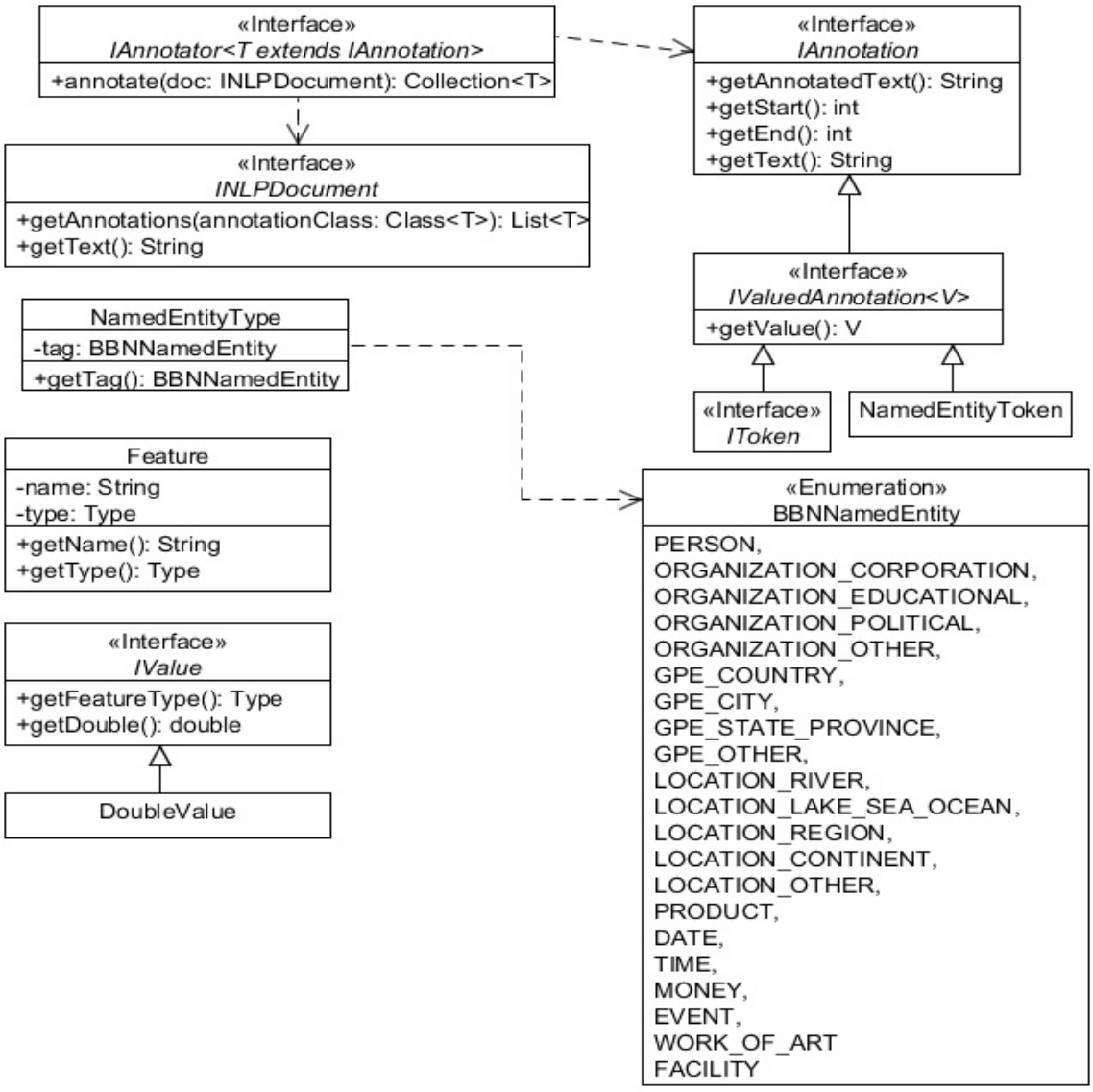

Рис.12. Диаграмма классов основных интерфейсов системы извлечения именованных сущностей Texterra.

Документ представлен интерфейсом `INLPDocument`, который содержит сам текст документа и его аннотации.

Аннотации описывают отдельные фрагменты текста с помощью дополнительной информации в зависимости от вида аннотации. Общие аннотации

добавляют к фрагменту текста только свой тип (например, абзац, слово, предложение) и представлены интерфейсом `IAnnotation`. Аннотации со значением, помимо своего типа, содержат дополнительную информацию (например, тип именованной сущности, часть речи) и представлены интерфейсом `IValuedAnnotation`.

Аннотирование текста производится реализациями интерфейса `IAnnotator`.

Тип именованной сущности представлен классом `NamedEntityType`, который, в свою очередь, содержит тег из перечисления `BBNNamedEntity`. Используемые в данной работе типы именованных сущностей, описанные в п.3.1, представлены как элементы этого перечисления, приведенные на рис.12.

Слово представлено интерфейсом `IToken`, а слово, аннотированное типом именованной сущности – классом `NamedEntityToken`.

Признак представлен классом `Feature`, который имеет название и тип. А вектор признаков представляется как Map<Feature, IValue>, где `IValue` – интерфейс для основных типов Java, например, в работе использовали его реализацию `DoubleValue` для обозначения нормированной частоты.

4.1 Базовая система извлечения именованных сущностей

Рассмотрим схему работы базовой системы, приведенную на рис.13. На вход системе извлечения именованных сущностей подается коллекция документов. Далее происходит разбиение каждого документа коллекции на абзацы, предложения, слова. Каждое слово подлежит классификации. Для этого производится извлечение признаков, составление вектора признаков, который подается на вход классификатору. Классификатор относит слово к тому или иному типу именованной сущности. Далее производится аннотирование входных документов именованными сущностями типов, предложенных классификатором. И на выходе имеем коллекцию документов, аннотированных именованными сущностями.

Извлечение перечисленных в п.3.2 признаков реализовано в базовой системе с помощью следующих классов:

- `AffixFeatureExtractor` - префиксы и суффиксы данного слова
- `PositionalFeatureExtractor` – позиция данного слова в предложении
- `FormFeatureExtractor` - форма данного слова
- `SpellingFeatureExtractor` - особенности написания данного слова
- `DigitNormalizationFeatureExtractor` - вариант написания данного слова с нормализованным вхождением цифр

- `ContextAggregationFeatureExtractor` – признак контекстной агреграции
- `LocalPredictionHistoryFeatureExtractor` – признак локальной истории прогнозов
- `ExtendedPredictionHistoryFeatureExtractor` – признак расширенной истории прогнозов
- `ExternalAnnotationsFeatureExtractor` – признак части речи.

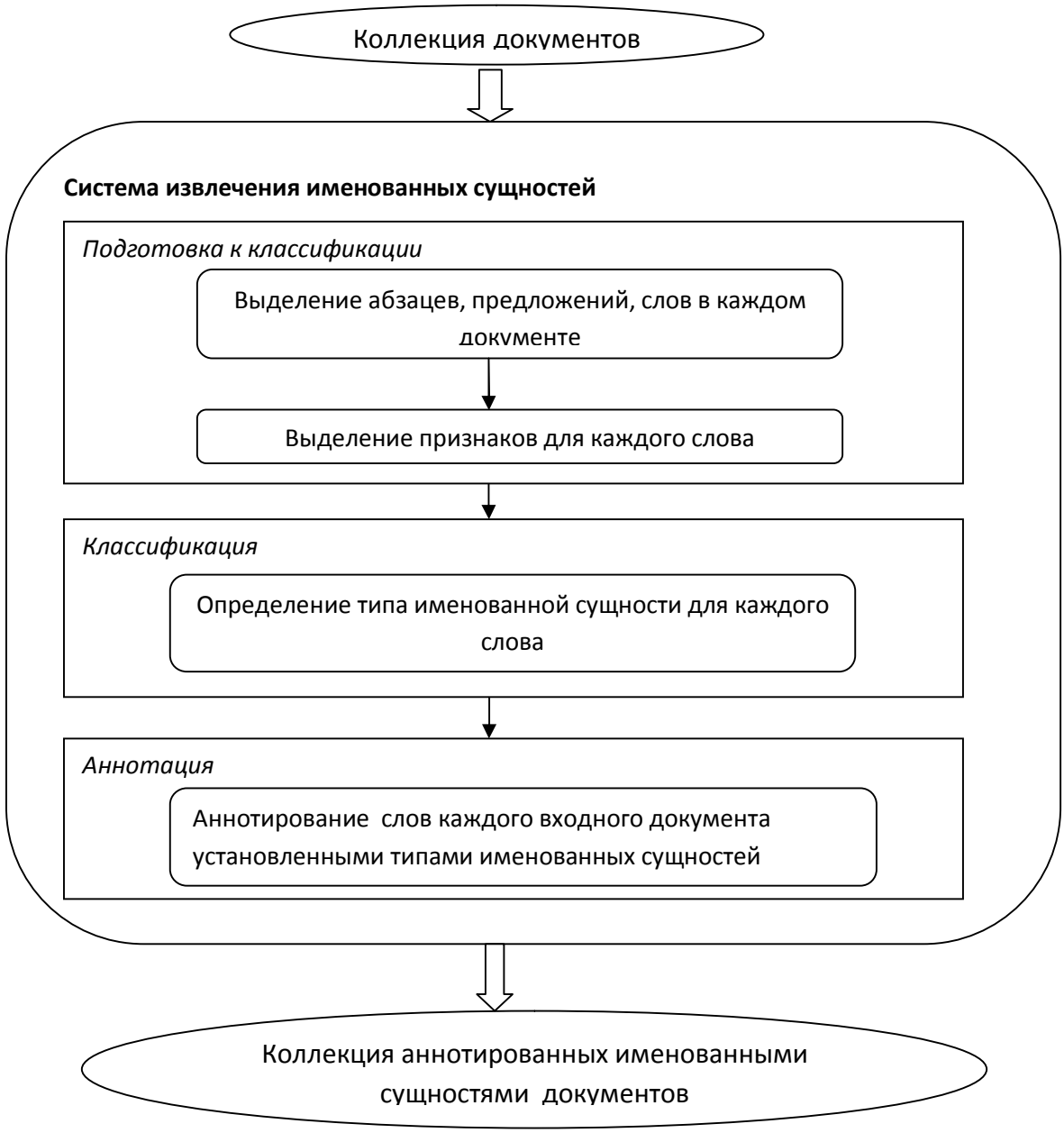

Рис.13. Схема работы базовой системы извлечения именованных сущностей.

Отдельные классы для извлечения различных признаков объединены для совместного использования в системе извлечения именованных сущностей классом `NERCFeatureExtractorFactory` посредством класса `UnionFeatureExtractor`.

Классификатор использует алгоритм машинного обучения, основанный на методе опорных векторов. В базовой системе используется реализация данного метода, предоставляемая открытой библиотекой LibLinear (см. [16]), которая является победителем ICML 2008 large-scale learning challenge и, согласно [16], подходит для решения задачи, так как учитывает особенности вектора признаков, которые формулировались в п.3.2.

4.2 Улучшенная система извлечения именованных сущностей

Рассмотрим общую схему работы улучшенной системы именованных сущностей, приведенную на рис.14.

Для добавления в базовую систему извлечения именованных сущностей признаков, предложенных в п.3.3, потребуются два предварительных этапа. На первом этапе производится классификация статей Википедии. На втором этапе происходит сбор статистики на основе проведенной классификации. Улучшение базовой системы происходит путем извлечения и использования при классификации дополнительных признаков, которые можно рассматривать как «нормированная частота отнесения данного слова к данному типу именованной сущности в Википедии».

Рис.14. Схема работы улучшенной системы извлечения именованных сущностей.

Рассмотрим более подробно реализацию обозначенных выше этапов.

Классификация Википедии реализована в системе Texterra и подходит для решения поставленной задачи с учетом адаптации существующих классов для обработки шаблонов. В рамках данной работы был создан класс `WikiClassifier`, организующий процесс классификации Википедии с помощью существующих средств. Диаграмма классов для этапа классификации приведена на рис.15. Здесь и далее на диаграммах классов не закрашенные серым классы обозначают классы, реализованные в рамках данной работы.

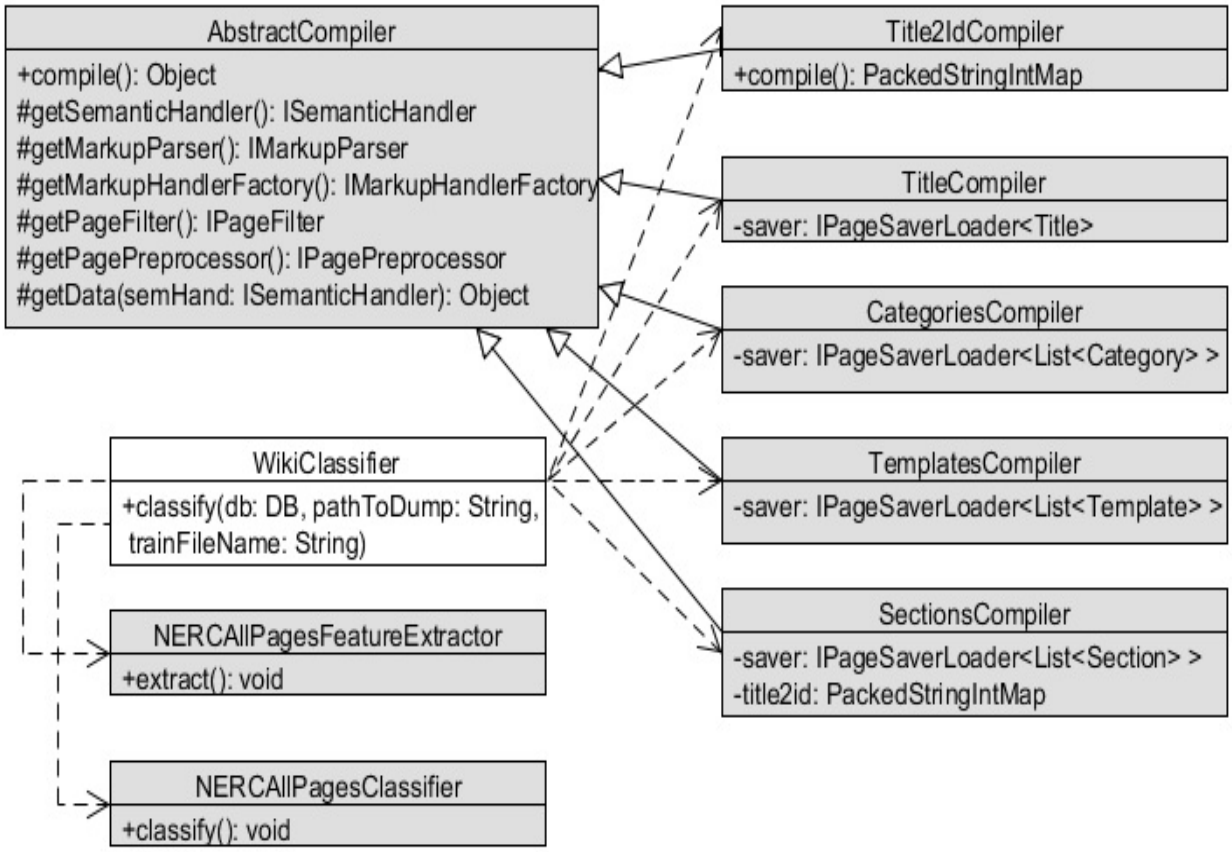

Рис.15. Диаграмма классов для этапа классификации Википедии.

Организация взаимодействия указанных классов следующая. Каждый из классов `TitleCompiler`, `CategoriesCompiler`, `SectionsCompiler`, `TemplatesCompiler` производит анализ своего аспекта снимка содержимого Википедии и записывает извлеченные данные в базу данных:

- `TitleCompiler` извлекает заголовок статьи;
- `CategoriesCompiler` извлекает список категорий, к которым относится данная статья;
- `SectionsCompiler` извлекает заголовки секций, тексты секций, первый абзац секции, первое предложение секции, текст ссылок в секции, заголовки статей, на которые указывают ссылки в секции;

- `TemplatesCompiler` извлекает заголовки шаблонов и пары «ключ-значение».

Далее класс `NERCAllPagesFeatureExtractor` производит извлечение признаков на основе информации, содержащейся в заполненной на предыдущем шаге базе данных, и записывает полученные вектора признаков в базу данных. После этого происходит классификация статей Википедии с помощью класса `NERCAllPagesClassifier`: обрабатывая содержащиеся в базе данных вектора признаков для статей, обученный на подаваемой заранее подготовленной выборке классификатор относит статью к некоторому типу именованной сущности.

Сбор статистики по классифицированной Википедии осуществляется реализованным в рамках данной работы классом `NamedEntityStatisticsExtractor`, который использует в своей работе некоторые существующие классы. Диаграмма классов для этапа сбора статистики приведена на рис.16.

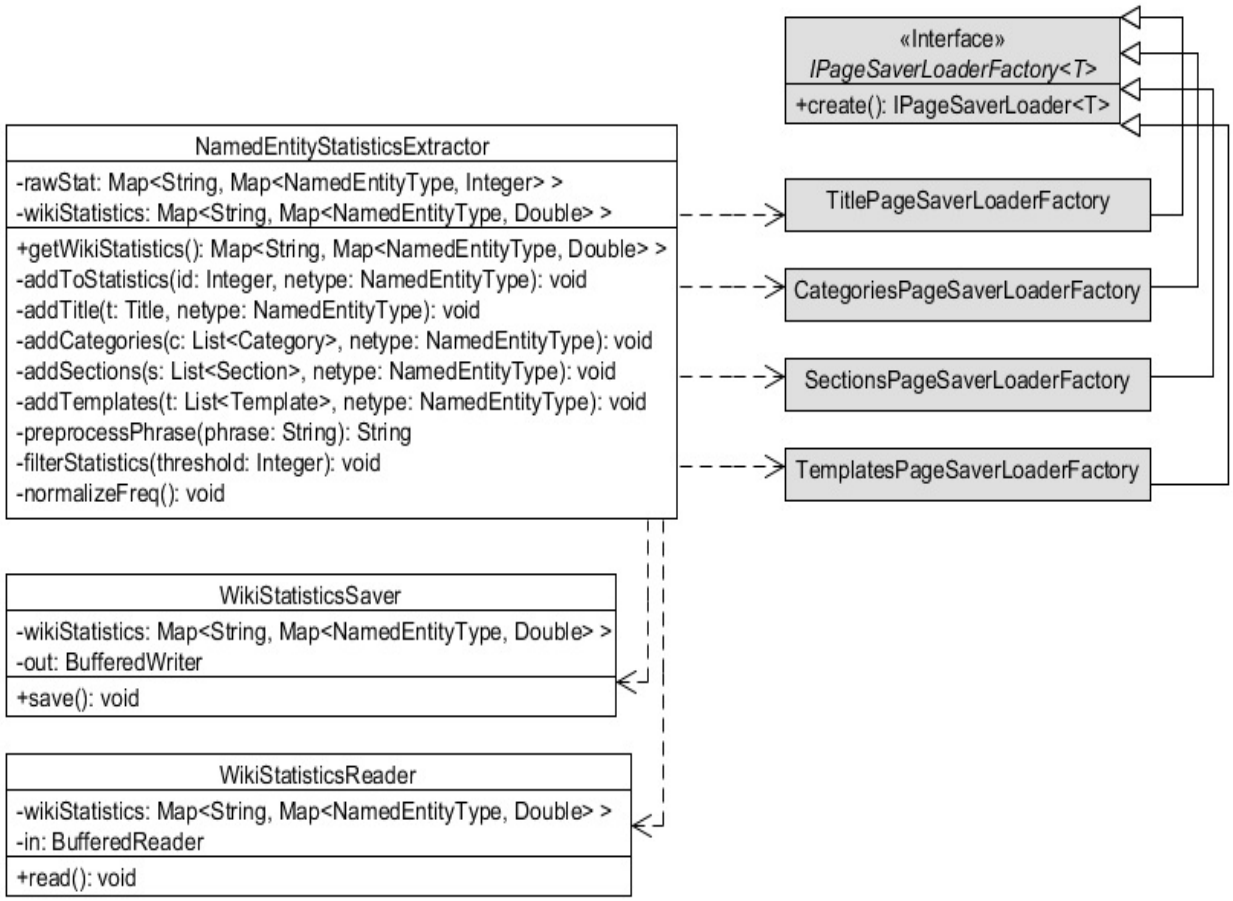

Рис.16. Диаграмма классов для этапа сбора статистики.

`NamedEntityStatisticsExtractor` для каждой статьи, для которой был определен тип именованной сущности, извлекает из базы данных записанную

туда информацию по каждому из перечисленных выше аспектов с помощью классов TitlePageSaverLoaderFactory, CategoriesPageSaverLoaderFactory,SectionsPageSaverLoaderFactory, TemplatesPageSaverLoaderFactory соответственно. Запуск сбора статистики происходит с помощью метода getWikiStatistics(). Сначала происходит вызов правил сбора статистики в методе addToStatistics(). К извлеченной информации применяются описанные в п.3.3 правила сбора статистики в соответствующих методах:

- addTitle() – правило для заголовка статьи;
- addCategories() – правило для названий категорий, к которым относится статья;
- addSections() – правило для заголовка секции и ссылок, содержащихся в тексте секций;
- addTemplates() – правило для заголовка и содержимого шаблона.

Данные правила требуют предварительную обработку словосочетаний, а именно возможное удаление служебных слов и специальных символов и обязательное приведение в нижний регистр, что происходит в методе preprocessPhrase(). Далее происходит фильтрация собранной статистики методом filterStatistics() по указанному порогу. После этого частоты в фильтрованной статистике нормируются методом normalizeFreq(). В результате получаем статистику «слово – тип именованной сущности – нормированная частота» в виде Map<String, Map <NamedEntityType, Double>>. Для удобства работы с собранной статистикой реализованы классы WikiStatisticsSaver и WikiStatisticsReader, которые обеспечивают сохранение и чтение для повторного использования ранее собранной статистики посредством записи в файл в формате <слово>\t<тип именованной сущности>\t<нормированная частота>.

Улучшенная система извлечения именованных сущностей реализуется классами, диаграмма которых представлена на рис.17.

Запуск улучшенной системы именованных сущностей производится с помощью метода recognize() класса ImprovedNamedEntityRecognizer. В нем используется класс ImprovedNamedEntityRecognizerReader-TrainerSaver, который обеспечивает чтение обучающей выборки, обучение на ней классификатора и сохранение полученной модели. Обучение классификатора обеспечивается классом ImprovedNamedEntityRecognizerTrainer, который использует класс ImprovedNERCFeatureExtractorFactory для получения объединения классов для извлечения признаков слов. Такое объединение, как говорилось в 4.1, реализуется посредством класса

UnionFeatureExtractor, в который в дополнение к используемым в базовой системе классам входит порождаемый классом WikiStatisticsFeatureExtractorFactory класс WikiStatistics-FeatureExtractor, осуществляющий извлечение дополнительных признаков. В методе extract() класса WikiStatisticsFeatureExtractor слово проходит предварительную обработку методом preprocessToken(), в результате которой удаляются возможные специальные символы и слово приводится в нижний регистр. После предварительной обработки происходит поиск данного слова в собранной по Википедии статистики, и если слово найдено, то в вектор признаков добавляется новый признак с названием word + netype.getTag().getTag(), типом Double и значением нормированной частоты из статистики. Такой признак можно рассматривать как «с какой частотой данное слово относилось к данному типу именованной сущности в контексте рассмотрения в Википедии».

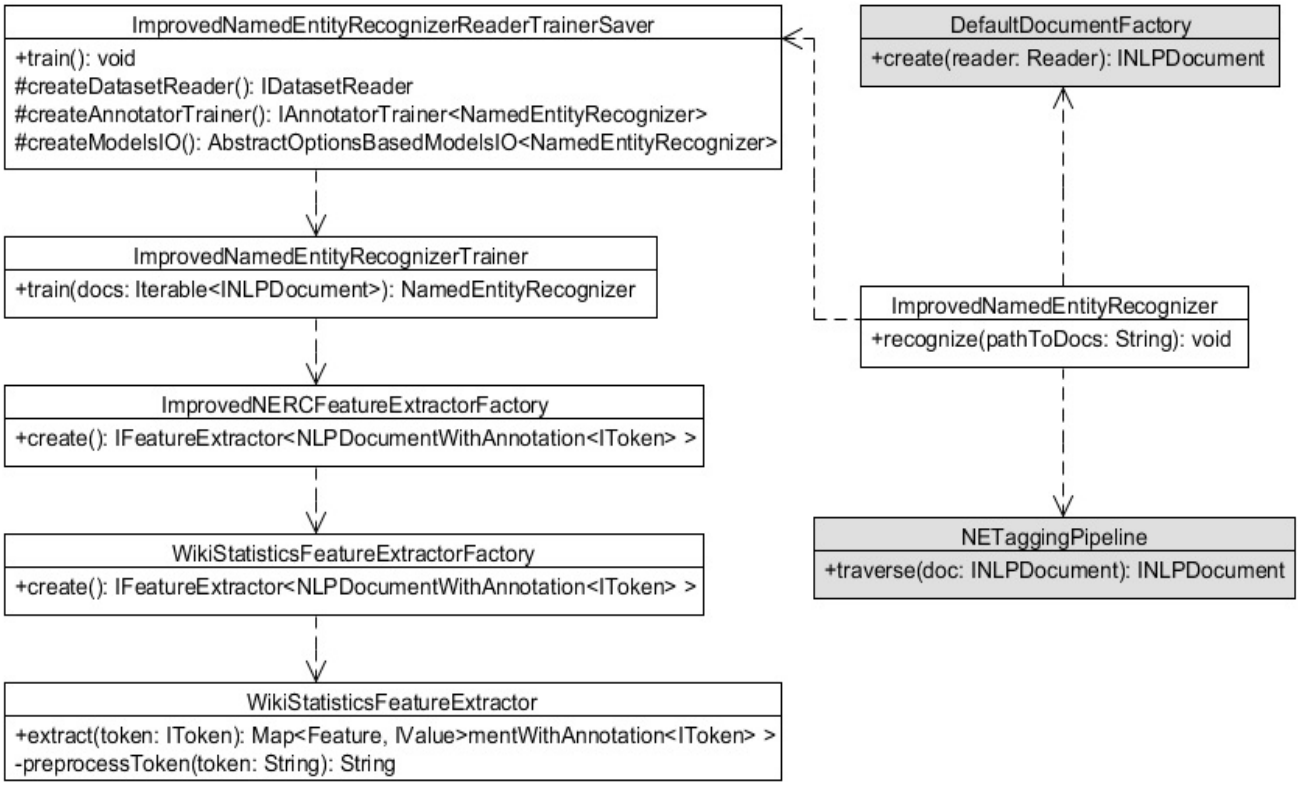

Рис.17. Диаграмма классов улучшенной системы извлечения именованных сущностей.

После обучения и сохранения модели система извлечения именованных сущностей готова к использованию. Поэтому далее в методе recognize() происходит преобразование входной коллекции документов: приведение каждого документа к типу INPLDocument с помощью метода create() класса DefaultDocumentFactory, обработка полученного документа улучшенной системой извлечения именованных сущностей методом traverse() класса NETaggingPipeline, завершающаяся аннотированием

документа и его сохранением с помощью метода `save()` класса `NLPDocumentSaver`.

4.3 Сценарий тестирования качества

Диаграмма классов, участвующих в тестировании качества, приведена на рис.18.

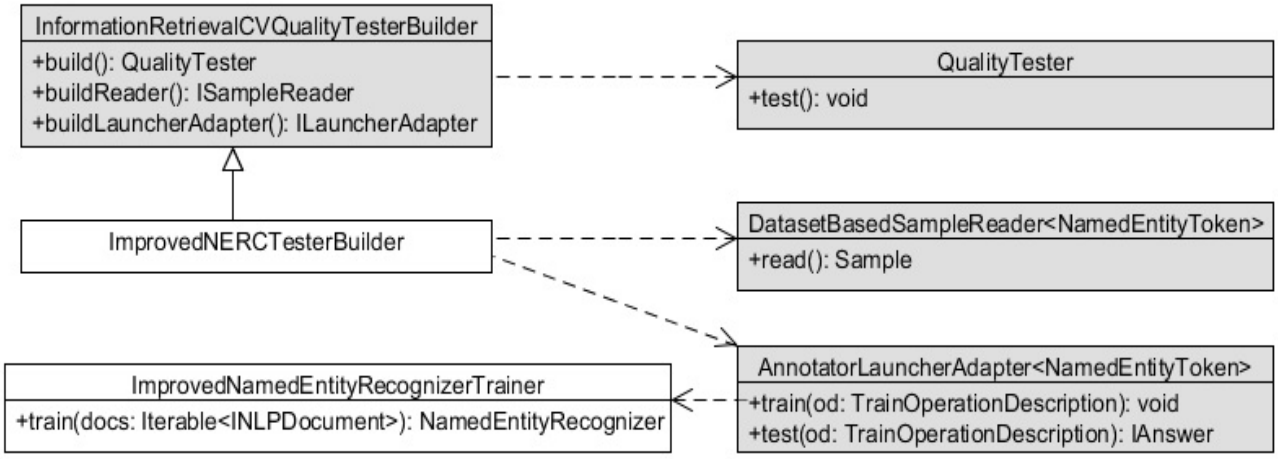

Рис.18. Диаграмма классов, участвующих в тестировании качества,

Класс `ImprovedNERCTesterBuilder` реализует перекрестную проверку на 10 частях, упоминаемую в п.3.4. На вход ему подается коллекция размеченных именованными сущностями документов. Методом `read()` существующего класса `DatasetBasedSampleReader`, параметризуемого `NamedEntityToken`, происходит считывание коллекции и разбиение ее на всевозможные части в соотношении 9 к 1. Часть коллекции из 9 частей подается в метод `train()` класса `AnnotatorLauncherAdapter`, который использует, в свою очередь, описанный выше `ImprovedNamedEntityRecognizerTrainer` и параметризуется `NamedEntityToken`. Так происходит обучение классификатора на 9 частях. После этого часть коллекции из 1 части подается в метод `test()` класса `AnnotatorLauncherAdapter`, где происходит аннотирование документа именованными сущностями с помощью обученного классификатора. После завершения аннотирования производится сравнение ожидаемых (полученных при считывании размеченной коллекции) и полученных результатов и подсчет метрик Precision, Recall, F-measure.

4.4 Характеристика программного средства

Программная реализация метода извлечения дополнительных признаков для системы извлечения именованных сущностей, использующего информацию из Википедии, состоит из 12 классов, суммарно около 1000 строк кода.

Запуск построенного решения производился на компьютере с 13 Гб памяти для виртуальной машины Java (этап классификации Википедии и сбора статистики) и на компьютере с процессором Intel Core 2 6420 2.13 GHz и 3Гб памяти для виртуальной машины Java(тестирование качества).

Этап классификации Википедии занял 18 часов. Этап сбора статистики занял 1 час. Время, требуемое на тестирование качества, зависело от объема входной коллекции: на коллекции из 10 документов тестирование завершалось за 1,5 мин, на коллекции из 100 документов – за 30 мин.

Заключение

В рамках решения поставленной задачи

- исследованы существующие методы извлечения именованных сущностей;
- предложен метод извлечения именованных сущностей, который использует полученную из Википедии информацию;
- выполнена программная реализация разработанного метода, интегрируемая в систему обработки текстов Texterra;
- подготовлен корпус на русском языке из новостных статей, размеченный именованными сущностями;
- проведено сравнительное тестирование качества разработанного метода.

В ходе решения рассматривался раздел Википедии, содержащий статьи на русском языке. В качестве типов именованных сущностей использовалось подмножество BBN, наиболее подходящее с точки зрения русского языка. Тестирование проводилось на новостных статьях на русском языке.

Таким образом, был предложен и реализован метод, использующий информацию, полученную из Википедии, для решения задачи извлечения именованных сущностей для русского языка.

Литература

1. Ratinov L., Roth D. Design challenges and misconceptions in named entity recognition. // Proceedings of the Thirteenth Conference on Computational Natural Language Learning / Association for Computational Linguistics. 2009. P. 147–155.

2. Tjong Kim Sang E.F., De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4 / Association for Computational Linguistics. 2003. P. 142-147.

3. Kim J.-D., Ohta T., Tsuruoka Y., Tateisi Y. Introduction to the Bio-Entity Recognition Task at JNLPBA // Proceedings of the international joint workshop on natural language processing in biomedicine and its applications / Association for Computational Linguistics. 2004. P. 70-75.

4. Nadeau D., Sekine S. A survey of named entity recognition and classification // Lingvisticae Investigationes. 2007. Vol. 30, no. 1. P. 3–26.

5. Ramshaw L. A., Marcus M. P. Text chunking using transformation-based learning // Natural language processing using very large corpora. Springer, 1999. P. 157– 176.

6. Uchimoto K., Ma Q., Murata M. et al. Named entity extraction based on a maximum entropy model and transformation rules // Proceedings of the 38th Annual Meeting on Association for Computational Linguistics / Association for Computational Linguistics. 2000. P. 326–335.

7. Zhang T., Johnson D. A robust risk minimization based named entity recognition system // Proceedings of the seventh conference on Natural language learning at HLTNAACL 2003-Volume 4 / Association for Computational Linguistics. 2003. P. 204–207.

8. Krishnan V., Manning C. D. An effective two-stage model for exploiting non-local dependencies in named entity recognition // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics / Association for Computational Linguistics. 2006. P. 1121–1128.

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

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

11. 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.

12. 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.

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

14. Ada B. Annotation guidelines for answer types //BBN technologies report. 2002.

15. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. //14th International Joint Conference on Artificial Intelligence, Palais de Congres Montreal, Quebec, Canada. 1995. P. 1137-1145.

16. Fan R.-E., Chang K.-W., Hsieh Ch.-J., Wang X.-R., Lin Ch.-J. LIBLINEAR: A library for large linear classification //The Journal of Machine Learning Research. T. 9. 2008. P. 1871-1874.