Texterra: инфраструктура для анализа текстов

Денис Турдаков, Никита Астраханцев, Ярослав Недумов, Андрей Сысоев, Иван Андрианов, Владимир Майоров, Денис Федоренко, Антон Коршунов, Сергей Кузнецов
{ turdakov, astrakhantsev, yaroslav.nedumov, sysoev,ivan.andrianov, vmayorov, fedorenko, korshunov, kuzloc }@ispras.ru

Аннотация. В статье описан проект Texterra, в рамках которого была создана инфраструктура для анализа текстов. Texterra предоставляет масштабируемое решение для быстрой обработки текстовых документов, основанное на использовании знаний, извлекаемых из Веб-ресурсов и текстовых документов. В данной статье раскрываются детали реализации проекта, варианты использования и результаты экспериментальных исследований разработанных инструментов.

Ключевые слова: анализ текстов, обработка естественного языка, Википедия, компьютерная лингвистика, машинное обучение, базы знаний, семантические онтологии, информационный поиск, извлечение терминологии.

1. Введение

Быстрый рост информационных технологий и количества информации, которую приходится анализировать человеку, сделали проблему эффективного управления данными одной из наиболее важных для многих областей жизнедеятельности. Без эффективных инструментов анализа информации невозможно принятие своевременных и обеспечивающих конкурентоспособность решений от персонального уровня до уровня крупных корпораций и государств.

Наиболее распространенным способом представления информации являются текстовые документы, часто относящиеся к определенной предметной области (документация предприятий, специализированные Веб-ресурсы и т.д.). Информация в таких документах представлена в неструктурированном виде, что существенно усложняет ее обработку. Для автоматического анализа таких данных в ИСП РАН была разработана специальная инфраструктура, получившая название Texterra.

Texterra представляет собой технологию для многоязычного анализа документов, которая основана на инновационных методах обработки текстов с использованием знаний, извлекаемых из Веб-ресурсов. Использование такой

технологии позволяет добиться высокой точности анализа при низких затратах на обучение и настройку системы.

Архитектура технологии предполагает возможность трех вариантов ее использования: (а) как библиотеки алгоритмов; (б) как инфраструктуры для создания собственных инструментов и их комбинирования для решения прикладных задач; (в) как масштабируемого облачного сервиса[i], который уменьшает расходы на интеграцию системы обработки текстов в пользовательские проекты.

Таким образом, Texterra может служить основой для создания различных приложений, требующих быстрого анализа текстов, в том числе:

- мониторинг новостей и анализ информации, извлекаемых из традиционных периодических изданий и сети Интернет;
- организация библиотек электронных документов, например патентов, технической документации, научных и других публикаций;
- анализ документации организаций для построения корпоративных баз знаний или повышения эффективности систем документооборота;
- анализ текстовых сообщений в социальных сетях и персональный репутационный мониторинг.

В отличие от многих существующих проектов по обработке и анализу текстов, основными приоритетами в проекте Texterra были использование автоматических методов и высокая скорость обработки данных при сохранении максимально высокого качества анализа текстов. В результате проекта была создана технология, которая успешно внедряется в нескольких коммерческих проектах с российскими и зарубежными партнерами, а также в собственных сервисах ИСП РАН.

В следующем разделе представлен обзор альтернативных технологий и описаны преимущества системы Texterra перед этими технологиями. В разделе 3 описана архитектура системы, позволяющая добавлять новые инструменты и комбинировать их с существующими. Раздел 4 посвящен базе знаний системы Texterra. Разделы 5 и 6 посвящены инструментам обработки текстов и экспериментальному тестированию их качества. В разделе 7 приводится краткое описание прикладных задач, решаемых с помощью технологии Texterra, и систем, созданных для решения этих задач.

2. Обзор области

Для анализа документов на естественном языке разработано большое количество библиотек, содержащих наборы базовых алгоритмов для анализа текстов, в основном на английском языке. Наиболее известными из них являются OpenNLP[ii], NLTK[1], LingPipe[iii]. Известны также инфраструктурные проекты GATE[2], Apache UIMA[3], предоставляющие набор инструментов для текстовой аналитики и расширяемую архитектуру для добавления новых инструментов.

В русскоязычном сегменте инструментов для обработки языка существенно меньше. Наиболее известным пакетом инструментов являются AOT[4]. Коммерческие решения предоставляют компании ABBYY, RCO, IBM и др.

Решаемые этими инструментами задачи, в большинстве случаев относятся к уровню морфологии и синтаксиса. Это связано со сложностью построения баз знаний, на основе которых можно перейти от слов и терминов к их значениям. Для английского языка существует тезаурус WordNet, который позволяет определить возможные значения слов и создать алгоритмы разрешения лексической многозначности. Для русского разрабатываются аналоги: РуТез[5], YARN[6] и др. Основными недостатками этих ресурсов являются сложность их разработки, требующая привлечения экспертов, и их ориентированность на покрытие только общей лексики, которой недостаточно для решения предметно-ориентированных задач. В системе Texterra основная база знаний автоматически извлекается из Википедии и более чем на порядок превышает размер WordNet. Кроме того, разрабатывается инструменты для автоматического построения предметно-ориентированных баз знаний на основе анализа документации, что существенно расширяет область применения технологии.

Еще одним популярным способом использования инструментов текстового анализа является их представление в виде облачных сервисов, что позволяет создавать интеллектуальные Веб-приложения. Эта идея лежит в основе нескольких проектов, наиболее известными из которых являются AlchemyAPI, OpenCalais, Semantria и OpenAmplify. Пользовательские приложения могут передавать текст таким сервисам и получать в качестве ответа решения сложных задач обработки текста. Texterra является первым проектом, который предоставляет аналогичную функциональность для русскоязычного сегмента.

Основными отличительными особенностями системы Texterra являются:

- расширяемая архитектура
- автоматически пополняемая база знаний
- поддержка работы с несколькими базами знаний
- инструменты для анализа лексической семантики, использующие базу знаний
- поддержка нескольких языков
- высокая скорость обработки данных

3. Архитектура

С точки зрения разработчика Texterra - это Java-фреймворк, построенная на его основе библиотека и несколько API, предоставляющих доступ к функциям библиотеки.

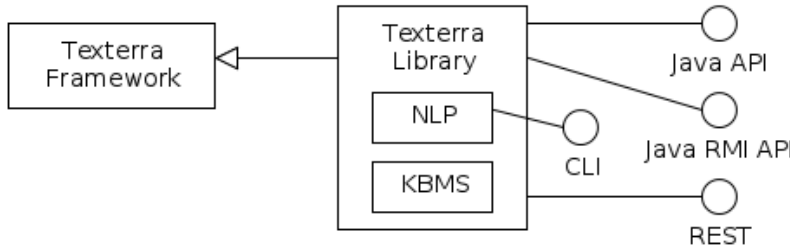

Рисунок 1. Общая архитектура системы Texterra

Функциональность системы Texterra делится на две большие части: управление базой знаний и обработка текстов на естественном языке.

Подсистема управления базой знаний предоставляет средства для хранения концептов, или понятий, вместе с отношениями между ними в виде графа, средства для перемещения по этому графу концептов, хранения текстовых представлений концептов и поиска концептов по их текстовым представлениям, а также эффективного подсчета семантической близости для пар и групп концептов.

При обработке текстов на естественном языке используется модель аннотирования текстов, аналогичная принятой в Apache UIMA. Таким образом, любая дополнительная информация, полученная для текста, сохраняется в виде экземпляра некоторого специфичного класса аннотаций. Немного более подробно модель данных представлена на рисунке 2.

При обработке естественного языка для получения итогового результата, как правило, требуется выполнить ряд предварительных шагов. Например, решение для задачи разрешения лексической многозначности в библиотеке Texterra предполагает предварительную токенизацию, выделение частей речи и терминов из текста. Для того чтобы снизить сложность разработки и повысить модульность итогового решения, каждый алгоритм, позволяющий осуществить один шаг аннотирования, оформляется отдельным классом, реализующим интерфейс IAnnotator (далее все такие классы назваеются аннотаторами). Все решение при этом заключается в последовательном применении одного аннотатора за другим. Классы, инкапсулирующие способ и порядок применения аннотаторов, называются пайплайнами. Такие классы должны реализовывать интерфейс IPipeline. Взаимосвязи между классами обработчиками и классами модели данных показано на рисунке 3.

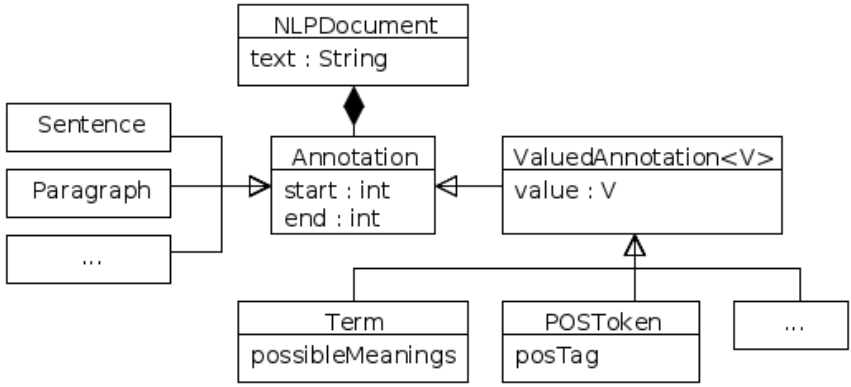

Рисунок 2. Модель данных

Таким образом, функциональность системы Texterra легко расширять, добавляя новые виды аннотаций, аннотаторов и пайплайнов.

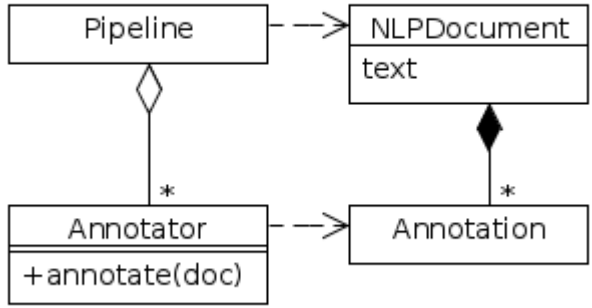

Рисунок 3. Взаимосвязь между классами-обработчиками и классами модели данных

4. База знаний

Применение баз знаний, или онтологий, доказало свою эффективность во многих приложениях, связанных с обработкой естественного языка, таких как извлечение информации [7], вопросно-ответные системы [8], информационный поиск [9] [10] и другие. Причина этого заключается в том, что использование баз знаний позволяет осуществить переход от отдельных

слов к выражаемым ими понятиям, что, в свою очередь, сокращает влияние разреженности языка и многозначности лексических единиц [11].

Для решения указанных выше проблем база знаний должна содержать следующее:

1) концепты – понятия, “сущности моделируемой предметной области, имеющие как минимум одно представление в виде выражения на естественном языке” [12];
2) термины – текстовые представления концептов;
3) отношения между концептами – определенная семантическая связь между понятиями предметной области.

Основной тип отношений между концептами, поддерживаемый в базе знаний системы Texterra, – семантическая близость. Это функция, определенная для любой пары концептов и имеющая значения от 0 до 1: чем ближе значение функции к 1, тем больше общего между концептами. Абсолютное значение семантической близости, как правило, не показательно, тогда как относительная близость легко интерпретируема: например, “Билл Гейтс” похож на “Стива Джобса” больше, чем на “Барака Обаму” и, тем более, на “Самолет”.

Выбор именно такого типа отношений между концептами обусловлен сценариями использования базы знаний. Можно выделить два основных сценария. Первый заключается в поиске в обрабатываемом тексте известных терминов, определении подходящих концептов для этих терминов с помощью алгоритма разрешения лексической многозначности и, возможно, определении ключевых концептов (см. раздел 5).

Второй сценарий предполагает использование информации о найденных концептах и отношениях между ними непосредственно в приложении. При этом следует отметить, что в общем случае отношения между концептами должны быть специфичными для приложения. Например, для вопросно-ответных систем эффективны именованные отношения с глубокой детализацией, для экспертных систем – формальные отношения с возможностью построения логических правил на их основе и т. д. С другой стороны, первый сценарий использования не предъявляет дополнительных требований к отношениям между концептами помимо эффективности соответствующих алгоритмов разрешения лексической многозначности и поиска ключевых концептов.

Система Texterra представляет собой инфраструктуру для анализа текстов, а не конкретное приложение, поэтому поддерживает работу с любыми доступными типами отношений. Кроме того, система Texterra содержит методы автоматического создания базы знаний на основе текстов (см. далее) и эти методы значительно усложняются с ростом сложности отношений между концептами, так что предельный показатель точности для базы знаний со специфичными отношениями может оказаться недостаточным для использования на практике.

В настоящее время базовым источником знаний для системы Texterra является интернет-энциклопедия Википедия: каждая статья Википедии считается концептом; каждое название статьи и текст гиперссылки на статью считается термином; семантическая близость вычисляется на основе гиперссылок между статьями с помощью меры Дайса (нормализованное число общих соседей, т.е. статей, имеющих гиперссылки с одной на другую). С архитектурной точки зрения, база знаний системы Texterra хранит именно гиперссылки между статьями и, таким образом, предоставляет возможность определять и другие типы отношений помимо семантической близости.

В иллюстративных целях часть базы знаний системы Texterra представлена на рисунке 4 - скриншоте разрабатываемого инструмента VizOntia, предназначенного для визуализации базы знаний.

Для построения базы знаний из интернет-энциклопедий разработан инструмент WikiParser, ключевые особенности которого следующие:

1) высокая производительность – 4 часа для построения базы знаний из полного набора статей (дампа) английской Википедии на персональном компьютере;
2) поддержка MediaWiki – правил разметки, по которым функционирует большая часть современных интернет-энциклопедий;
3) встроенная борьба с зашумленными данными – термины, которые выражают определенный концепт менее 5% случаев по сравнению с остальными терминами для этого концепта, удаляются из базы знаний, поскольку представляют собой ошибки или слишком большую зависимость от контекста.

На момент написания этой статьи англоязычная Википедия содержала более 4.5 млн. статей, однако покрытие некоторых предметных областей все равно остается неполным. В целях повышения полноты покрытия система Texterra предусматривает возможность одновременного использования нескольких баз знаний, построенных из разных источников. В частности, с помощью упомянутого выше инструмента WikiParser можно получать базы знаний на основе предметно-специфичных энциклопедий, например энциклопедия по вселенной “Звездные войны”[iv].

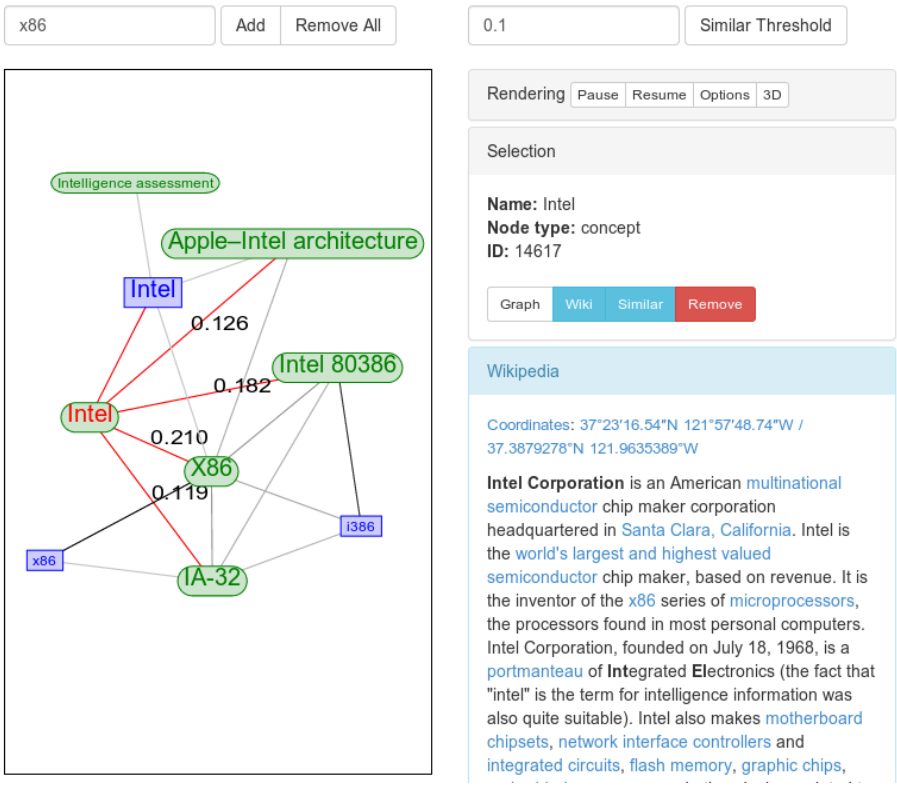

Рисунок 4: Экранная форма инструмента VizOntia

Однако в силу постоянного роста знаний и доступной текстовой информации даже этих источников полуструктурированных данных недостаточно, так как наиболее распространенным, а иногда и единственным, способом представления знаний во многих предметных областях являются обычные текстовые документы. По этой причине в составе системы Texterra разрабатывается инструмент для автоматического построения базы знаний на основе анализа текстовых документов определенных предметных областей.

Данный инструмент устроен следующим образом. На первом этапе извлекаются кандидаты в термины, т.е. слова и словосочетания, удовлетворяющие предопределенным шаблонам частей речи и прошедшие первоначальную фильтрацию по частоте и наличию предопределенных стоп-слов. На следующем этапе каждый кандидат классифицируется в предметно-специфичный термин или не термин с помощью метода машинного обучения, признаки которого включают в себя как статистические, так и лингвистические характеристики кандидатов. После этого для каждого

предметно-специфичного термина образуется концепт, для которого производится поиск связанных концептов (“соседей” в терминах Википедии, которые затем будут участвовать в вычислении семантической близости). В итоге получается набор терминов, концептов и связей между ними, которые и представляют собой базу знаний для предметно-специфичной области.

5. Инструменты для обработки текстов

Система Texterra предоставляет широкий набор инструментов для обработки текстов, включающий в себя как стандартные методы, например определение частей речи слов, так и оригинальные методы, основанные на использовании базы знаний, извлеченной из Википедии. Кроме того, Texterra включает в себя инструменты, предназначенные для обработки неформальных пользовательских текстов, таких как сообщения социальных сетей.

5.1. Стандартные методы

Система Texterra поддерживает следующие стандартные методы:

1) определение границ предложений в тексте;
2) определение границ отдельных слов, или токенов, в предложении;
3) определение частей речи слов;
4) приведение слов к нормальной форме.

В качестве реализации первых 3 методов используется библиотека OpenNLP.

Нормализация, или лемматизация, выполняется для английского языка по эвристическому алгоритму, основанному на морфологических свойствах существительных. Для русского языка используется собственная реализация алгоритма MyStem [13], основанного на словаре, содержащем для каждого слова нормальную форму и набор возможных суффиксов.

5.2. Методы, основанные на базе знаний

Одной из основных задач, решаемой системой Texterra, является семантический анализ текстов с помощью базы знаний, построенной на основе Википедии. Основными этапами семантического анализа являются: распознавание терминов, определение значений терминов и извлечение ключевых концептов текста.

На первом этапе текст разбивается на последовательность терминов, присутствующих в словаре базы знаний системы Texterra. Далее для каждого найденного термина запускается алгоритм разрешения лексической многозначности, основанный на классификаторе концептов по следующим признакам: вероятность того, что термин ссылается на статью Википедии; частота концепта в Википедии; семантическая близость к контексту; качество контекста. Заключительным этапом анализа является извлечение ключевых концептов, позволяющим получить сжатое высокоуровневое представление текста, отражающее его смысл. Для решения данной задачи используется

специальный алгоритм, основанный на поиске кластеров в графе концептов [23].

Важной особенностью разработанных этапов семантического анализа является их независимость от языка текста: для успешного применения методов достаточно использовать базу знаний, основанную на соответствующей языковой версии ресурса Википедия.

5.3. Методы, предназначенные для обработки неформальных текстов

Существует отдельный класс текстов, для которых стандартные подходы работают неэффективно - это сообщения в социальных сетях, чатах, форумах и т.д. Такие тексты, как правило, содержат множество грамматических и орфографических ошибок, а также жаргон и специфичные для интернета сущности, такие как ссылки на веб-страницы, хэштеги, имена пользователей и эмотиконы. Для того чтобы работать с такими текстами, Texterra предусматривает отдельный блок предобработки, позволяющий обнаруживать и корректировать текстовые аномалии вышеупомянутых типов. В состав блока входят:

1) детектор опечаток, работающий по словарю;
2) средства исправления опечаток, использующие фонетические модели;
3) средство проверки орфографии Jazzy[v] и языковые модели для выбора конкретного варианта замены (для английского языка);
4) средства обнаружения ссылок, хэштегов, эмотиконов, или смайлов, и имен пользователей, основанные на регулярных выражениях.

Все обнаруженные орфографические ошибки исправляются, остальные аномалии удаляются из текста. Обработанный таким образом текст можно затем передавать стандартным алгоритмам.

Кроме того Texterra предоставляет инструменты для анализа эмоциональной окраски пользовательских сообщений. Метод определения эмоциональной окраски текста состоит из двух этапов: на первом этапе текст классифицируется на нейтральный или эмоциональный, после чего эмоциональный текст классифицируется на позитивный или негативный. Каждый этап реализуется с помощью алгоритма машинного обучения (метод опорных векторов), в качестве признаков используются n-граммы по нормализованным словам и по частям речи.

Также поддерживается метод определения отдельных атрибутов, на которые направлены эмоции, например в отзыве “Довольно неуклюжие спецэффекты компенсируются декорациями” негативно оценивается атрибута “визуальные эффекты” и позитивно – “художественное оформление”. Данный метод основан на алгоритме бутстреппинга и на этапе обучения требует ручного задания нескольких ключевых слов для каждого атрибута.

6. Результаты экспериментов

Данный раздел содержит результаты экспериментальных исследований системы Texterra применительно к различным задачам обработки данных.

Информация о качестве определения частей речи на английском языке взята с официального сайта библиотеки OpenNLP. Тестирование для русского языка производилось методом перекрестной проверки на части корпусов OpenCorpora [14] и национального корпуса русского языка [15]. Результаты представлены в таблице 1.

	Точность
Английский язык	0.9659
Русский язык	0.9702

Таблица 1. Точность определения частей речи

Для тестирования подзадач семантического анализа использовалось 5 англоязычных коллекций документов, состоящих из текстов различных предметных областей.

Первая коллекция, обозначаемая **MODIS-texts**, состоит преимущественно из технических текстов, связанных с информационными системами и обработкой данных; размер данной коллекции – 131 документ. Коллекция **BoardGames** состоит из 35 текстов, относящихся к единственной предметной области – «Настольные игры».Тексты из коллекции **Tweets** характеризуются малой длиной, обилием неформальных терминов и различной тематической направленностью; данная коллекция состоит из 100 документов. Коллекции **AQUAINT** (50 новостей различных тематик [16]) и **Wikipedia** (100 случайно выбранных статьей ресурса Википедия) пригодны только для тестирования определения значений терминов, поскольку в них отсутствует разметка большинства терминов и ключевых концептов.

Результаты тестирования алгоритмов распознавания терминов, определения значений слов и извлечения ключевых концептов представлены в таблице 2. Стоит отметить, что определение значений терминов тестируется только для корректно определенных терминов.

	Распознавание терминов			Определение значений терминов	Извлечение ключевых концептов (топ-5 наиболее вероятных)		
	Precision	Recall	F1-measure	Accuracy	Precision	Recall	F1-measure
MODIS-texts	55%	72%	63%	77%	30%	36%	32%
Board Games	60%	71%	65%	67%	30%	22%	25%
Tweets	40%	58%	47%	75%	26%	43%	32%
AQUAINT	–	–	–	86%	–	–	–
Wikipedia	–	–	–	89%	–	–	–

Таблица 2. Результаты тестирования подзадач семантического анализа.

В таблицах 3 и 4 представлены результаты тестирования алгоритма определения эмоциональной окраски сообщений для английского языка, в таблице 5 — для русского. Тестирование для английского языка проводилось на объединении наборов данных общей направленности: Stanford [17], Sentiment140 [18], KnowCenter [19], UNED [20]; обзоров фильмов: ICWSM [21], IMDb [22]; политической направленности: Debates [23]. Тестирование для русского языка производилось на объединении наборов данных обзоров фильмов: Imhonet Movies; обзоров книг: Imhonet Books; обзоров фотокамер: Yandex.Market. Указанные наборы данных для русского языка были собраны в ИСП РАН.

	Accuracy	Precision	Recall	F1-measure
Texterra	0.981	0.984	0.995	0.989
OpenAmplify	0,51	0,758	0,552	0,639
Alchemy	0,6012	0,6012	1,0	0,75

Таблица 3. Определение присутствия эмоциональной окраски для английского языка

	Accuracy	Precision	Recall	F1-measure
Texterra	0.790	0.782	0.8	0.791
OpenAmplify	0,572	0,508	0,633	0,564
Alchemy	0,42	0,341	0,908	0,494

Таблица 4. Определение полярности эмоциональной окраски для английского языка

	Accuracy	Precision	Recall	F1-measure
Присутствие эмоциональной окраски	0.77	0.831	0.89	0.86
Полярность эмоциональной окраски	0.85	0.884	0.947	0.914

Таблица 5. Определение эмоциональной окраски для русского языка

Для тестирования скорости работы использовался набор из 131-го текстового документа на английском языке (суммарный объём - 190КБ; ~242 слова на документ). Для обеспечения нагрузки систем использовался пул из 10-ти параллельно работающих потоков. Результаты представлены в таблице 6.

Решаемая задача	Система	КБ/с	Слов/с	Терминов/с
Определение терминов	Texterra	94	15722	2472
	DBpedia Spotlight	34	5679	327
Определение значений терминов	Texterra	82	13684	1521
	DBpedia Spotlight	35	5824	333

Таблица 6. Сравнительное тестирование скорости работы системы Texterra и DBpedia Spotlight для задач определения терминов и их значений

Как видно из представленных рисунков, скорость работы системы Texterra в несколько раз превышает скорость работы аналогичной системы DBpedia Spotlight [24]. Кроме того, можно заметить, что скорость работы системы Texterra при определении терминов выше, чем при определении их значений -

это объясняется тем, что первая задача в системе Texterra является составной частью второй.

7. Приложения

Texterra может применяться для решения различных задач, требующих обработки текстов. Например, использование Texterra позволяет перейти от классического информационного поиска по ключевым словам к семантическому поиску по значениям слов. В частности, использование Texterra вместе с открытой поисковой системой Apache Lucene позволяет повысить качество ранжирования (таблица 7).

	MAP
Apache Lucene	0.1948
Lucene + Texterra	0.2305

Таблица 7. Ранжирование при информационном поиске с помощью ApacheLucene. Мера Mean Average Precision (MAP) для корпуса TIPSTER-TREC (Financial Times Limited)

Кроме того, наличие базы знаний, позволяющей оценивать близость между понятиями, помогает решать и другие задачи из областей информационного поиска и анализа данных, включая:

- расширение запросов с целью увеличения полноты поиска,
- построение фасетных поисковых интерфейсов,
- создание рекомендательных систем на основе сравнения описаний рекомендуемых объектов,
- анализ текстовых сообщений пользователей социальных сетей и форумов, например, с целью выявления скрытых демографических атрибутов [25],
- разработку вопросно-ответных систем, систем автоматического реферирования, диалоговых систем и др.

Часть описанных возможностей технологии Texterra демонстрируется в системе поиска информации и навигации по блогосфере BlogNoon [26]. С помощью инфраструктуры Texterra тексты сообщений блогов анализируются и строится их семантическая модель, содержащая значения ключевых терминов, сгруппированные по темам. На основе этой информации пользователю предоставляется возможность производить поиск по ключевым словам, а также терминологический поиск с учетом значений многозначных терминов. Кроме того, на основе Texterra в системе BlogNoon реализованы:

- механизм для рекомендации сообщений и блогов,
- механизм генерации динамических подсказок для расширения и изменения запросов на основе анализа поисковой выдачи (фасетный

поиск),

- механизм автоматического построения кратких описаний блогов в виде облака ключевых слов, сгруппированных по темам.

8. Заключение

В рамках проекта Texterra была создана технология, позволяющая решать широкий класс задач, связанных с обработкой текстовых данных. В зависимости от решаемой задачи Texterra может быть использована как библиотека алгоритмов, расширяемый фреймворк или масштабируемый облачный сервис. В отличие от большинства существующих систем обработки текстов, Texterra предоставляет возможность перехода от работы с отдельными словами и терминами к работе с их значениям. Это позволяет увеличить точность решения многих прикладных задач. При этом особое внимание при разработке технологии уделялось производительности системы – на данный момент Texterra является одним из самых быстрых решений в данной области.

Важным преимуществом технологии Texterra являются низкие затраты на внедрение и поддержание системы за счет автоматизации процесса построения и обновления базы знаний. В качестве основной базы знаний используется информация, автоматически извлекаемая и Википедии. Далее эта база знаний расширяется информацией из других Веб-ресурсов и за счет анализа текстовых документов. Такой подход позволяет применять разработанные методы не только к заранее определенной предметной области, но и быстро адаптировать технологию к новым задачам и языкам.

Список литературы

[1] Steven Bird, Ewan Klein, Edward Loper, and Jason Baldridge. 2008. Multidisciplinary instruction with the Natural Language Toolkit. InProceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 62-70.

[2] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoSComputBiol 9(2)

[3] David Ferrucci et.al. Towards an Interoperability Standard for Text and Multi-Modal Analytics. Technical report RC24122. IBM. 2006

[4] Игорь Ножов. "Морфологическая и синтаксическая обработка текста(модели и программы)", тезисы диссертации. 2003

[5] Алексеев А., Добров Б., Лукашевич Н. Лингвистическая онтология тезаурус РуТез // Труды конференции Open Semantic Technologies for Intelligent Systems - OSTIS. — 2013. — С. 153–158.

[6] YARN Браславский П. И., Мухин М. Ю., Ляшевская О. Н., Бонч-Осмоловская А. А., Крижановский А. А., Егоров П. Е. YARN: начало. Труды конференции Диалог-2013.

[7] V. Karkaletsis, P. Fragkou, G. Petasis, and E. Iosif, “Ontology based information extraction from text,” in Knowledge-Driven Multimedia Information Extraction and

Ontology Evolution, ser. Lecture Notes in Computer Science, G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, Eds. Springer Berlin / Heidelberg, 2011, vol. 6050, pp. 89–109.
[8] C. Unger and P. Cimiano, “Pythia: Compositional meaning construction for ontology-based question answering on the semantic web,” in Natural Language Processing and Information Systems, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 6716, pp. 153–160.
[9] Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann, “Ontology refinement for improved information retrieval,” Information Processing & Management, vol. 46, no. 4, pp. 426 – 435, 2010.
[10] M. Grineva, D. Turdakov, and A. Sysoev, “Blognoon : Exploring a topic in the blogosphere,” in Proceedings of the 20th international conference companion on World wide web, Hyderabad, India, 2011, pp. 213–216.
[11] C. Biemann, “Ontology Learning from Text : A Survey of Methods”,LDV-Forum,vol. 20, pp. 75–93, 2005.
[12] Н.А. Астраханцев, Д.Ю. Турдаков. “Методы автоматического построения и обогащения неформальных онтологий”. Программирование, Т.39, №1, с. 23-34, 2013.
[13] Segalovich A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine
[14] Bocharov V.V., Alexeeva S.V., Granovsky D.V., Protopopova E.V., Stepanova M.E., Surikov A.V. Crowdsourcing morphological annotation // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая–2 июня 2013 г.). Вып. 12 (19). — М.: РГГУ, 2013.
[15] Ляшевская О. Н., Плунгян В. А., Сичинава Д. В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005. 111—135.
[16] David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08). ACM, New York, NY, USA
[17] Stanford Twitter sentiment general domain dataset [Электронный ресурс] URL: http://www.stanford.edu/~alecmgo/cs224n/trainingandtestdata.zip (дата обращения: 22.07.2012)
[18] Sentiment140 Twitter sentiment general domain dataset [Электронный ресурс] URL: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip (дата обращения: 22.07.2012)
[19] KnowCenter Twitter sentiment general domain dataset [Электронный ресурс] URL: http://know-center.tugraz.at/loesungen/daten (дата обращения: 22.07.2012)
[20] UNED Twitter sentiment general domain dataset [Электронный ресурс] URL: http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html (дата обращения: 22.07.2012)
[21] International Conference on Weblogs and Social Media movie domain dataset [Электронный ресурс] URL: http://icwsm.cs.mcgill.ca (дата обращения: 6.12.2013)
[22] IMDb movie review dataset [Электронный ресурс] URL: http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip (дата обращения: 6.12.2013)

[23] Twitter Sentiment Dataset from the 1st 2008 Presidential Debate [Электронный ресурс] URL: http://www.infochimps.com/datasets/twitter-sentiment-dataset-2008-debates (дата обращения: 6.12.2013)
[24] Mendes P.N., Jakob M., García-Silva A., Bizer C. DBpedia Spotlight: Shedding Light on the Web of Documents. In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011). Graz, Austria, September 2011.
[25] Антон Коршунов. Задачи и методы определения атрибутов пользователей социальных сетей. Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2013
[26] M. Grineva, M. Grinev, D. Lizorkin. Extracting Key Terms From Noisy and Multitheme Documents. WWW2009: 18th International World Wide Web Conference

Texterra: A Framework for Text Analysis

Denis Turdakov Nikita Astrakhantsev, YaroslavNedumov, AndreySysoev, Ivan Andrianov, VladimirMayorov, Denis Fedorenko, Anton Korshunov, Sergey Kuznetsov
{turdakov, astrakhantsev, yaroslav.nedumov, sysoev,ivan.andrianov, vmayorov, fedorenko, korshunov, kuzloc}@ispras.ru
ISP RAS, Moscow, Russia

Abstract. The paper presents a framework for fast text analytics developed during the Texterra project. Texterra is a technology for multilingual text mining based on novel text processing methods that exploit knowledge extracted from user-generated content. It delivers a fast scalable solution for text mining without the expensive customization. Depending on use-cases Texterra could be utilized as a library, extendable framework or scalable cloud-based service. This paper describes details of the project, use-cases and results of evaluation for all developed tools.

Texterra utilizes Wikipedia as a primary knowledge source to facilitate text mining in arbitrary documents (news, blogs, etc). We mine the graph of Wikipedia's links to compute semantic relatedness between all concepts described in Wikipedia. As a result, we build a semantic graph with more than 5 million concepts. This graph is exploited to interpret meanings and relationships of terms in text documents.

In spite of large size, Wikipedia doesn't contain information about many domain-specific concepts. In order to increase applicability of the technology we developed several automatic knowledge extraction tools. These tools include systems for knowledge extraction from MediaWiki resources and Linked Data resources, as well as system for knowledge base extension with concepts described in arbitrary text documents using original information extraction techniques.

In addition, utilization of information from Wikipedia allows easily extend Texterra for support of new Natural languages. The paper presents evaluation of Texterra applied for different text processing tasks (part-of-speech tagging, word sense disambiguation, keyword extraction and sentiment analysis) for English and Russian.

Keywords: Text mining, natural language processing, Wikipedia, computational linguistics, machine learning, knowledge base, semantic ontology, information retrieval, terminology extraction.

References

[1]. Bird S., Klein E., Loper E., Baldridge J. Multidisciplinary instruction with the Natural Language Toolkit. Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, 2008. pp. 62-70.

[2]. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS computational biology, 9(2), 2013.

[3]. Ferrucci D. et al. Towards an interoperability standard for text and multi-modal analytics. IBM Res. Technical report RC24122, 2006.

[4]. Nozhov I. Morfologicheskaya i sintaksicheskaya obrabotka teksta(modeli i programmy) [Morphological and syntactic text processing (models and programs)]. Tezisy dissertatsii [PhD Thesis], 2003. (in Russian).

[5]. Alekseev A., Dobrov B., Lukashevich N. Lingvisticheskaya ontologiya tezaurus RuTez [Linguistic ontology thesaurus RuTez] // Trudy konferentsii Open Semantic Technologies for Intelligent Systems [The Proceedings of Open Semantic Technologies for Intelligent Systems], 2013. pp. 153–158. (in Russian).

[6]. Braslavskij, P., Mukhin, M., Lyashevskaya, O. N., Bonch-Osmolovskaya, A. A., Krzhizhanovskij, A., Egorov, P. (2012). YARN: nachalo [YARN: The beginning]. Trudy konferentsii Dialog [The Proceedings of International Conference on Computational Linguistics Dialog], 2013.

[7]. Karkaletsis V., Fragkou P., Petasis G., Iosif E. Ontology based information extraction from text. Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, ser. Lecture Notes in Computer Science, G. Paliouras, C. Spyropoulos, and G. Tsatsaronis, Eds. Springer Berlin / Heidelberg, 2011. vol. 6050, pp. 89-109. doi: 10.1007/978-3-642-20795-2_4

[8]. Unger C., Cimiano P. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. Natural Language Processing and Information Systems, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011. vol. 6716, pp. 153–160. doi: 10.1007/978-3-642-22327-3_15

[9]. Jimeno-Yepes A., Berlanga-Llavori R., Rebholz-Schuhmann D. Ontology refinement for improved information retrieval. Information Processing & Management, 2010. vol. 46, no. 4, pp. 426 – 435.

[10]. Grineva M., Turdakov D., Sysoev A. Blognoon: Exploring a topic in the blogosphere. Proceedings of the 20th international conference companion on World wide web, Hyderabad, India, 2011. pp. 213–216.

[11]. Biemann C. Ontology Learning from Text: A Survey of Methods. LDV-Forum, 2005. vol. 20, pp. 75–93.

[12]. Astrakhantsev N, Turdakov D. Automatic construction and enrichment of informal ontologies: A survey. Programming and Computer Software, 2013. vol. 39, no. 1, pp. 34-42. doi: 10.1134/S0361768813010039

[13]. Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In MLMTA, 2003. pp. 273-280.

[14]. Bocharov V., Alexeeva S., Granovsky D., Protopopova E., Stepanova M., Surikov A. Crowdsourcing morphological annotation. Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoj Mezhdunarodnoj konferentsii «Dialog» [The Proceedings of International Conference on Computational Linguistics Dialog]. 2013. vol. 12, no. 19.

[15]. Lyashevskaya O., Plungyan V., Sichinava D. O morfologicheskom standarte Natsional'nogo korpusa russkogo yazyka [About morphological standard of Russian National Corpus]. Natsional'nyj korpus russkogo yazyka: 2003-2005. Rezul'taty i perspektivy [Russian Natioanl Corpus: 2003-2005. Results and Prospects], 2005. pp. 111—135.
[16]. Milne D., Witten I. H. Learning to link with wikipedia. Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08), 2008.
[17]. Stanford Twitter sentiment general domain datasetAvailable at: http://www.stanford.edu/~alecmgo/cs224n/trainingandtestdata.zip
[18]. Sentiment140 Twitter sentiment general domain dataset. Available at: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
[19]. KnowCenter Twitter sentiment general domain dataset. Available at: http://know-center.tugraz.at/loesungen/daten
[20]. UNED Twitter sentiment general domain dataset. Available at: http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html
[21]. International Conference on Weblogs and Social Media movie domain dataset. Available at: http://icwsm.cs.mcgill.ca
[22]. IMDb movie review dataset. Available at: http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip

---

[i] https://api.ispras.ru

[ii] http://opennlp.apache.org

[iii] http://alias-i.com/lingpipe

[iv] http://ru.starwars.wikia.com/wiki/

[v] http://jazzy.sourceforge.net