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

УДК 004.3

Выделение ключевых слов в русскоязычных текстах

***Ершов Ю.С.**, бакалавр*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»

Научный руководитель: Волкова Л.Л., ассистент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

Задача выделения ключевых слов и фраз из текста возникает в библиотечном деле, лексикографии и терминоведении, а также в задачах информационного поиска. В настоящее время объёмы и динамика информации, которая подлежит обработке в этих областях, делают особенно актуальной задачу автоматического выделения ключевых слов и фраз, которые могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, кластеризация и классификации.

Существует большое число доступных систем автоматического выделения ключевых слов, разработанных и ориентированных на обработку естественных языков. Эти системы основаны на определённых методах – методах выделения ключевых слов, которые делятся на лингвистические и статистические [2]. Лингвистические методы основываются на значениях слов, используют онтологии и семантические данные о слове. К сожалению, эти методы слишком трудоёмки на ранних этапах: разработка онтологий, к примеру, весьма трудозатратна. Статистические же основываются на численных данных о встречаемости слова в тексте.

Сегодня интеллектуальный анализ данных (Data Mining) применяется в широком спектре задач: информационный поиск, лингвистика, самолётостроение, биология, медицина и др. Одной из важнейших составляющих современных медицинских информационных систем является подсистема интеллектуального анализа текстовых данных, адекватность функционирования которой напрямую зависит от качества работы модуля машинной лингвистики, одной из задач которого является создание ключевых фраз для текстов для кластеризации документов и повышения скорости поиска документов, релевантных запросу.

Таким образом, задачи исследования и разработки системы автоматического извлечения ключевых фраз из текста на естественном языке актуальны.

Извлечение ключевых фраз

В процессе предварительной отработки из текста производится удаление неинформативных частей (графематический анализ [2]). Стоп-слова – это слова, не представляющие ценности при данном типе обработки. Чаще всего это предлоги, союзы, междометия и пр. N-грамма – это термин из компьютерной лингвистики, означающий последовательность из N элементов текста (термов), например, слов или их последовательностей.

Кандидаты в ключевые слова возможно отбирать в виде N-грамм, не разделенных знаками препинания (кроме дефиса и кавычек) и стоп-словами. Ключевыми словами могут быть как единичные слова, так и пары слов, тройки и т. д. [2]. Другой подход основывается на анализе связей между словами в предложениях и в тексте, полученных либо с помощью тех же N-грамм (подход менее трудоёмок), либо на разобранном тексте. Если текст подвергся одному, а лучше всем этапам анализа текста (морфологический даст БНФ, синтаксический — связи между словами, семантический — смысловую карту связей), можно получить более точную информацию о тексте, хоть это и более затратно [1, 2], к примеру, на основании деревьев синтаксического разбора (при наличии синтаксической разметки).

Важнейшим этапом в задаче извлечения ключевых фраз является расчет их весов информативности, который позволяет оценить их значимость по отношению друг к другу в документе. Для каждой из отобранных ключевых фраз рассчитываются признаки, которые позволяют судить о важности кандидата для данного документа. Набор отобранных ключевых фраз ранжируется по значениям признаков, например, в соответствии с их частотностью и весами информативности, рассчитанными по одной из методик (к примеру, количество «соседей» в графе связей между словами в тексте). После ранжирования производится отбор лучших ключевых фраз из этого списка или отбираются кандидаты, превышающие установленный минимальный порог значения признака.

Обзор аналогов

В настоящее время существует большое количество систем автоматического извлечения ключевых фраз из текста на естественном языке. Ключевыми факторами при

отборе аналогов в данной статье были рекомендации экспертов, несколько исследовательских работ, посвящённых аналогам, а также популярность соответствующих систем в современном IT–сообществе.

1. OpenCalais

OpenCalais — Web-сервис, предназначенный для автоматического извлечения семантических метаданных из текстов на естественном языке. Начиная с 2007 года, развитием и поддержкой сервиса занимается корпорация Thomson Reuters.
Семантические метаданные представлены в виде именованных сущностей (англ. named entity), а также связанных с ними фактов и событий. Именованные сущности, в свою очередь, могут рассматриваться как ключевые слова и фразы исходного текста.

Функционирование системы OpenCalais основано на методах обработки естественного языка, машинного обучения и других алгоритмах. Для извлечения семантических метаданных применяются предварительно подготовленные онтологии различных предметных областей в формате RDF. Исходный текст подвергается предварительной обработке (графематической и морфологической разметке), затем размеченные словосочетания проходят идентификацию при помощи обученной модели распознавания именованных сущностей, между которыми ведётся поиск семантических отношений. Полученный граф сущностей и отношений между ними преобразуется в набор RDF-троек.

Web-сервис OpenCalais бесплатен и доступен для некоммерческого и коммерческого использования, однако требует регистрации для получения APIключа. Сервис построен по архитектуре REST и использует формат XML для обмена данными с пользователями. На сегодняшний день Web-сервис OpenCalais не способен обрабатывать русскоязычные тексты.

2. Extractor

Extractor — система автоматического извлечения терминов, функционирующая с 2002 года и используемая многими организациями в собственных решениях по обработке естественного языка.

Работа системы Extractor основана на применении генетических алгоритмов в сочетании с методами машинного обучения и статистическими методами обработки естественного языка. Первоначальное обучение системы ведётся на основе размеченного корпуса текстов.

Ознакомиться с возможностями Extractor можно при помощи демонстрационного Web-приложения ExtractorLive.com. Для создания собственных решений на основе технологии Extractor требуется приобрести набор инструментов разработчика.
На сегодняшний день Extractor не способен обрабатывать русскоязычные тексты.

3. Yahoo! Term Extraction Web Service

Yahoo! Term Extraction Web Service — сервис, используемый в поисковой системе Yahoo! Search, предназначенный для извлечения ключевых фраз из текста на естественном языке. Документация Yahoo! Term Extraction Web Service использует закрытую технологию извлечения терминов. Обмен данными с пользователем осуществляется в форматах XML и JSON.

В данный момент обработка русскоязычных текстов при помощи Yahoo! Term Extraction Web Service невозможна.

4. TerMine

TerMine — Web-сервис извлечения терминов, разработанный в британском Национальном центре анализа текста (англ. The National Centre for Text Mining).

Сервис TerMine работает на основе метода C-value и применяет анализатор TreeTagger для предварительной морфологической разметки текста.

Демонстрационный Web-интерфейс TerMine позволяет обрабатывать тексты исключительно на английском языке.

5. Maui

Maui — система тематической классификации текстовых документов, работающая на основе методов обработки естественного языка и машинного обучения.

Схема функционирования системы состоит из двух этапов работы: этапа первоначального построения и обучения модели и этапа применения обученной модели к решению задачи тематической классификации текста.

Результаты тематической классификации, полученные при помощи Maui, могут рассматриваться в качестве тегов (меток) исходного текста. Без использования обученной модели Maui функционирует как система автоматического извлечения ключевых фраз.
Система Maui является свободным кроссплатформенным программным обеспечением и распространяется на условиях лицензии GNU General Public License Version 3. При

помощи Web-приложения maui-indexer можно ознакомиться с основными возможностями системы.

В настоящий момент Maui не способна обрабатывать русскоязычные тексты.

6. TextAnalyst

TextAnalyst — инструмент для поиска информации и анализа содержания текстов, имеющий возможность выделения ключевых слов.

Функционирование TextAnalyst основано на применении методов обработки естественного языка в сочетании с методами машинного обучения.

Пакет TextAnalyst доступен для ознакомительного использования в виде приложения для семейства операционных систем Windows и способен обрабатывать тексты на русском языке.

7. AOT

AOT — проект, направленный на создание системы автоматического перевода «ДИАЛИНГ». В рамках проекта AOT разработан комплекс инструментов автоматической обработки текста, в том числе графематический, морфологический и синтаксический анализаторы русского языка.

Все инструменты, разработанные в рамках проекта AOT (в том числе и синтаксический анализатор) являются свободным кроссплатформенным программным обеспечением и распространяются на условиях лицензии GNU Lesser General Public License Version 2.1.

8. ContentAnalyzer

ContentAnalyzer — инструмент для анализа содержания тематических Web-страниц в реальном времени, выделения списков ключевых слов и словосочетаний, построения автореферата текста документа.

Функционирование ContentAnalyzer обеспечивается за счёт вычисления характеристик текста, указанных ниже.

— частота термина/словосочетания в документе;

— отношение частоты к числу слов документа;

— вес термина в документе (с учётом частоты и весовых коэффициентов);

— вес термина к числу слов документа.

Пакет ContentAnalyzer распространяется бесплатно, доступен для использования в виде приложения для семейства операционных систем Windows и способен обрабатывать тексты как на русском, так и на английском языках.

9. Семантическое зеркало

Семантическое зеркало — система тематической классификации Webстраниц, разработанная компанией «Ашманов и партнёры».

Сервис «Семантическое зеркало» обрабатывает текст Web-страницы и определяет её тему: анализирует слова, семантические связи между ними, выделяет самые важные термины. Темы определяются по рубрикатору, где к каждой рубрике приписано некоторое множество терминов.

Результат тематической классификации можно использовать в качестве списка ключевых фраз исходного текста или в качестве набора тегов (меток). Эту информацию можно использовать для показа контекстной рекламы и новостей на актуальную тему.

На сайте компании «Ашманов и партнёры» доступна демонстрационная версия «Семантического зеркала», имеющая лимит в 128 обращений с одного IP-адреса в сутки. Сервис обрабатывает тексты как на русском, так и на английском языках.

Сравнение аналогов

Вышеуказанные системы автоматического извлечения ключевых фраз из текста на естественном языке оцениваются по следующим критериям:

1) поддержка русского языка (“Р”):

0 — поддержка отсутствует;

1 — поддержка присутствует;

2) качество результата по итогам экспертной оценки (“К”):

0.0 — минимальная оценка;

1.0 — максимальная оценка;

3) доступность аналога (“Д”):

0.0 — использование аналога требует приобретения платной лицензии или временной подписки;

0.5 — существуют полноценные бесплатные версии аналога, бета-версии или специальные версии для академических исследований;

1.0 — аналог распространяется как свободное программное обеспечение;

4) независимость аналога от наличия онтологии заданной области знаний или специализированного тезауруса в процессе извлечения ключевых фраз (“О”):

0 — аналог спроектирован с целью использования специализированного тезауруса или онтологии области знаний в процессе выделения терминов;

1 — результат выделения терминов не зависит от наличия специализированного тезауруса или онтологии области знаний.

В результате обзора современных систем извлечения ключевых фраз из текста на естественном языке было найдено 9 аналогов. Для того, чтобы перейти к выбору прототипа системы извлечения ключевых слов, необходимо оценить каждый из аналогов в соответствии с вышеуказанными критериями. Результаты оценки сведены в табл. 1.

Таблица 1

Сравнение существующих аналогов по выбранным критериям.

№	Название аналога	Оценка по критериям				
		Р	К	Д	О	Σ
1	OpenCalais	0	0.8	0.5	0	1.3
2	Extractor	0	0.7	0.0	1	1.7
3	Yahoo! Term Extraction Web Service	0	0.6	0.5	1	2.1
4	TerMine	0	0.7	0.5	1	2.2
5	Maui	0	0.6	1.0	1	2.6
6	TextAnalyst	1	0.3	0.5	1	2.8
7	AOT	1	0.4	1.0	1	3.4
8	ContentAnalyzer	1	0.6	0.5	1	3.1
9	Семантическое зеркало	1	0.5	0.5	1	3.0

Согласно приведённым в табл. 1 результатам, русский язык поддерживают далеко не все системы, а качество работы рассмотренных систем не превышает 80%, и этот показатель может быть превзойдён при использовании инструментов анализа текста (синтаксический, далее — семантический разбор). Наиболее полно удовлетворяют запрашиваемым критериям системы 7, 8 и 9.

Переход к концептуальной модели

Поскольку лучшее качество достигается при аналитическом подходе к тексту или же при комбинации его со статистическим подходом, систему автоматического

извлечения ключевых фраз из текста на естественном языке следует разрабатывать с использованием графематических шаблонов, морфологического словаря (лексикона) и синтаксических правил. Эти данные определяются предварительно и хранятся в базе данных. Текст подлежит обработке графематическим анализатором, который вырабатывает информацию о разделении текста на абзацы, предложения и отдельные слова, необходимую для дальнейшей обработки. Каждое слово, выделенное графематическим анализатором, подвергается морфологическому анализу с целью построения морфологической интерпретации (часть речи, форма и т. д.), определения основы слова и формирования леммы. На основе имеющейся графематической и морфологической интерпретации текста выполняется построение и наполнение синтаксических групп и выявление отношений между ними. Ключевые фразы выделяются, к примеру, из именных групп, сформированных синтаксическим анализатором при помощи метода C-value. При выборе ключевых слов стоит уделить внимание стоп-словам (фильтрация по длине слова, словарь стоп-слов и пр.), а также частям речи, поскольку вероятность релевантности как ключевого слова, к примеру, существительного и наречия будет различаться.

Выводы

В данной работе данной работы проведено исследование методов автоматического извлечения ключевых фраз из текстов на русском языке, включая сравнение по эффективности, на основании которого сделано заключение о направлении разработки собственного метода. Качество можно повысить, сделав акцент на использовании синтаксического анализа и сдвига фокуса от статистических методов в сторону аналитических методов.

Список литературы

1. Волкова Л.Л. Приложения теории тесного мира в компьютерной лингвистике // 3-я Международная научно-практическая конференция «Модель подготовки специалистов новой формации, адаптированных к инновационному развитию отраслей»: сборник трудов. Душанбе, 2012. С. 172-175.
2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.

3. Белоусов А.И., Ткачев С.Б. Дискретная математика: учеб. для вузов / под ред. В.С. Зарубина, А.П. Крищенко. 4-е изд., исправл. М.: Изд-во МГТУ им. Н.Э. Баумана, 2006. 744 с.