

Степанов Павел Андреевич

СИСТЕМЫ АНАЛИЗА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

Статья посвящена решению проблем извлечения информации из текстов естественного языка. Описаны сложности, возникающие при анализе текстов, а также основные подходы к решению типичных задач этой области. Представлен общий подход к построению систем анализа текста, а также предложен разработанный в процессе исследований подход, основанный на применении альтернативного подхода к анализу текста на основе потоков интерпретаций.

Адрес статьи: www.gramota.net/materials/1/2013/6/49.html

Статья опубликована в авторской редакции и отражает точку зрения автора(ов) по рассматриваемому вопросу.

Источник

Альманах современной науки и образования

Тамбов: Грамота, 2013. № 6 (73). С. 159-161. ISSN 1993-5552.

Адрес журнала: www.gramota.net/editions/1.html

Содержание данного номера журнала: www.gramota.net/materials/1/2013/6/

© Издательство "Грамота"

Информация о возможности публикации статей в журнале размещена на Интернет сайте издательства: www.gramota.net

Вопросы, связанные с публикациями научных материалов, редакция просит направлять на адрес: almanac@gramota.net

- научить разбираться в разных точках зрения на одну проблему.

Метод проектов может включать разные формы организации учебной, поисковой, исследовательской и других видов деятельности, определяющих современные компетенции будущего специалиста.

Таким образом, метод проектов, с одной стороны, предоставляет обучающимся возможность самостоятельного приобретения знаний в процессе решения трудноразрешимых практических задач, требующих интеграции знаний из различных предметных областей, с другой стороны, метод проектов – это современная педагогическая технология, которая предполагает совокупность исследовательских, поисковых, проблемных методов, творческих по своей сути.

Список литературы

1. **Вербицкий А. А.** Активное обучение в высшей школе: контекстный подход. М., 1991.
2. **Ковалевская Е. А.** Мотивация студентов при обучении иностранному языку в неязыковых вузах и особенности его преподавания // Альманах современной науки и образования. Тамбов: Грамота, 2011. № 2 (45).
3. **Прохорец Е. К., Плеханова М. В., Хмелидзе И. Н.** Основы проектной работы на занятиях по немецкому языку в техническом вузе. Томск, 2003.
4. **Соколова Т. А.** К вопросу о повышении качества образования при обучении иностранному языку // Инновации в языковом образовании: обучение иноязычному общению в различных типах учебных заведений в контексте компетентностного подхода: материалы третьей всероссийской научно-практической конференции. Иркутск: ИГЛУ, 2012.
5. **Тарева Е. Г.** На пути к обеспечению нового качества лингвистического образования // Инновации в языковом образовании: обучение иноязычному общению в различных типах учебных заведений в контексте компетентностного подхода: материалы второй всероссийской научно-практической конференции. Иркутск: ИГЛУ, 2003.
6. **Улюмджиева В. Э.** К проблеме использования активных методов обучения // Иностранные языки: теория и практика. М.: Тезаурис, 2011. № 1 (12).

УДК 004.82; 004.912; 81.322.2

Технические науки

Статья посвящена решению проблем извлечения информации из текстов естественного языка. Описаны сложности, возникающие при анализе текстов, а также основные подходы к решению типичных задач этой области. Представлен общий подход к построению систем анализа текста, а также предложен разработанный в процессе исследований подход, основанный на применении альтернативного подхода к анализу текста на основе потоков интерпретаций.

Ключевые слова и фразы: обработка естественного языка; *natural language processing*; анализ текста; извлечение информации; поток интерпретаций.

Степанов Павел Андреевич

Новосибирский государственный университет

stefan.nsk@gmail.com

СИСТЕМЫ АНАЛИЗА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА[©]

Объем информации, доступной в сети Интернет, растет с каждым годом. Причем большая часть этой информации представляет собой тексты на естественном языке. В зависимости от области знаний, информация может быть представлена в виде статей, комментариев или сообщений на публичном форуме. Информация в сети Интернет дублируется, уточняется и пополняется ежедневно. Нетрудно понять, что имеющиеся в данный момент доступные ресурсы всемирной сети представляют собой колоссальную базу знаний, представленных в форме, сложно поддающейся компьютерной обработке – в виде текста на естественном языке.

Хотя информация в сети Интернет представлена множеством разных способов и в различных форматах, основным объектом внимания в данной статье будет текст на естественном языке. Если наука пока только «подбирается» к построению эффективных алгоритмов для распознавания визуальных образов, звука и видео, то текст на естественном языке, казалось бы, представлен в наиболее «удобоваримом» для машинной обработки виде. Тем не менее, извлечение информации из текстов является достаточно сложной задачей.

Приведем ниже общее описание процесса анализа текста, а также список основных проблем. На вход любой системе для анализа текста, как правило, поступает текст в исходном формате, то есть в популярных форматах документов DOC, OAT, RTF, PDF, HTML. Такие документы кроме самого текста содержат также форматирование, сноски, «лишние» участки текста (реклама). Первой сложностью при анализе текста является извлечение текста из документов, поданных на вход. Далее, полученный текст следует пропустить через графематический анализатор, задача которого – определить границы слов, предложений, абзацев, а также отметить числа, знаки препинания и прочие символы и последовательности символов, не являющиеся словами, но

играющие важную роль в предложении. Сложность данного этапа состоит в том, что некоторые последовательности символов часто бывает невозможно идентифицировать на этом этапе анализа. Если принять во внимание необходимость обработки опечаток, проблема идентификации слов становится непростой задачей.

Затем, когда предложения размечены, и слова локализованы, начинает работу морфологический анализатор. Применяя имеющийся словарь и набор словоизменительных правил, этот анализатор дополняет каждое слово набором грамматических характеристик. Среди них известные всем падежи, рода и числа, а также некоторые менее известные грамматические значения. На этапе морфологического анализа также возникает ряд сложностей с обработкой опечаток, особенно если они допущены в окончании слова, играющем, как известно, ключевое значение в определении набора грамматических характеристик. Ряд сложностей возникает при анализе неизвестных слов, а также заимствованных из других языков.

Следующий этап анализа – синтаксический. Наиболее популярным подходом к этому анализу является построение дерева синтаксического разбора предложения. В процессе построения дерева исходное предложение перестраивается в древовидную структуру, представляющую собой аналог дерева синтаксического разбора некоторой формальной грамматики. К сожалению, зачастую доступная в сети информация представлена в виде разговорной речи и, как следствие, слабо поддается полному синтаксическому разбору, так как даже опытный лингвист не сможет построить дерево синтаксического разбора для подобных предложений. В качестве решения этой проблемы существует альтернативный подход к синтаксическому анализу, а именно – частичный синтаксический анализ. При частичном синтаксическом анализе полное дерево разбора не строится, а вместо этого анализатор сосредотачивается на поиске заранее определенных синтаксических конструкций в тексте.

Такой подход вполне приемлем, так как естественный язык не имеет грамматики в привычном математикам смысле. Естественный язык не имеет терминальных и нетерминальных символов, правил вывода и не может быть отнесен ни к одному типу в иерархии Хомского. Естественные языки, как правило, имеют достаточно вольные «правила» построения грамматических конструкций, слабо поддающиеся формализации. В достаточной степени формализуемы лишь некоторые синтаксические отношения:

1. Отношения согласования и зависимости при построении словосочетаний.
2. Валентности глаголов и отглагольных причастий.
3. Ссылки на контекст (различные типы местоимений).
4. Общие правила построения предложений (простое предложение, сложносочиненное, сложноподчиненное предложение и другие сложные предложения).

Заключительным этапом анализа текста является семантический анализ результатов, полученных на этапе синтаксического анализа. Для данного этапа не существует устоявшихся моделей и подходов. В большинстве систем роль семантического анализатора играет эвристически реализованный модуль, осуществляющий поставленную перед ним задачу. Ясно, что даже при наличии дерева полного синтаксического разбора текста, все равно сложно реализовать алгоритм, извлекающий полезную информацию на основе полученного дерева.

Далее будет описан предлагаемый подход к анализу текста.

В рамках данной работы проводился ряд исследований, направленных на извлечение информации из текстов естественного языка для автоматизированного построения онтологий. Одним из результатов является новый подход к анализу текста: «поток интерпретаций». Поток интерпретаций является множеством пар (участок текста) + (некоторая информация). При этом не ставится никаких ограничений на тип информации и ее формат. Таким образом, основное отличие предлагаемого подхода от традиционного состоит в том, что, если традиционно принято делить процесс анализа на этапы, жестко разделяя при этом порядок исполнения этапов анализа, а также результат каждого этапа, то в подходе, основанном на потоках интерпретаций, предлагается фиксировать лишь формат результатов анализа.

Приведем пример потока интерпретаций для некоторого текста. Рассмотрим следующий текст:

«Аналитическая геометрия (раздел геометрии) исследует геометрические фигуры средствами алгебры».

Для данного текста составляются следующие интерпретации:

1. Интерпретации, соответствующие всем словам и знакам препинания.
2. Интерпретации, соответствующие результатам морфологического анализа слов.
3. Интерпретация, рассматривающая уточнение в скобках как пробельный символ.
4. Интерпретации, соответствующие терминам «аналитическая геометрия» и «геометрические фигуры».
5. Интерпретация, несущая информацию о разрешении валентности глагола «исследует»: субъект = аналитическая геометрия, объект = геометрические фигуры.
6. Любые другие виды информации об участках текста.

Анализаторы, действующие как преобразователи данных в традиционной модели анализа текста, заменяются обработчиками потока интерпретаций, основная задача которых – дополнять поток интерпретаций новой информацией на основе существующей. При этом не накладывается никаких ограничений на количество альтернативных интерпретаций одного и того же участка текста.

Замена преобразователей информации обработчиками потока позволяет произвольным образом комбинировать этапы анализа, в том числе возвращаться на уже заверченный этап.

Возможность добавления альтернативных интерпретаций одного участка текста позволяет легко порождать альтернативные ветки анализа. Например, при добавлении двух разных интерпретаций одного участка текста следующий обработчик будет вынужден разветвить процесс анализа на две ветки: по одной на каждую альтернативную интерпретацию.

Существует возможность отмечать любой участок предложения как незначительный, добавив в качестве его интерпретации пробельный символ. Тогда все обработчики на одной из веток анализа будут считать, что вместо данного участка в тексте стоит пробельный символ.

Подобным образом легко реализуются различные автоматические «исправители» опечаток. Все возможные варианты просто добавляются в качестве альтернативных интерпретаций.

В модели потоков интерпретаций легко реализуется возврат на предыдущие этапы анализа и даже оценка достоверности полученных результатов. Обработчики потока интерпретаций независимы друг от друга, и поэтому анализ с их помощью представляет собой легко масштабируемый, прозрачный процесс.

Разумеется, подход, основанный на потоках интерпретаций, не лишен недостатков. Основной недостаток данного подхода – значительно большая сложность реализации обработчиков по сравнению с анализаторами. Увеличение сложности связано со сложностями обработки потока интерпретаций, который может содержать альтернативные интерпретации одного участка текста, неинтерпретированные участки и др. Кроме того, структура обработчиков, напрямую независимых друг от друга, безусловно, сложнее планарной структуры анализаторов. Тем не менее, получаемые удобства стоят затраченных усилий на разработку сложных алгоритмов обработчиков потоков интерпретаций.

В статье был описан разработанный подход к анализу текстов на естественном языке при помощи потоков интерпретаций. Данный подход является естественным усовершенствованием традиционного поэтапного подхода к анализу. Имея достаточно приемлемый список недостатков, потоки интерпретаций представляют собой удобный инструмент, позволяющий эффективно вмешиваться в любой этап процесса анализа. Вместо набора преобразователей предлагается использовать обработчики потока интерпретаций, каждый из которых работает независимо и имеет своей целью дополнение потока новой информацией. Отсутствие ограничений на количество альтернативных интерпретаций одного и того же участка текста позволяет легко порождать альтернативные ветки анализа, сложно реализуемые в традиционном подходе.

Список литературы

1. Власов Д. Ю., Пальчунов Д. Е., Степанов П. А. Извлечение отношений между понятиями из текстов на естественном языке // Вестник НГУ. 2010. Т. 6. № 3. С. 23-33.
2. Зализняк А. А. Грамматический словарь русского языка. Словоизменение. Около 100000 слов. М.: Русский язык, 1997.
3. Пальчунов Д. Е. Решение задачи поиска информации на основе онтологий // Бизнес-информатика. 2008. № 1. С. 3-13.
4. Современный русский язык: учеб. для филол. спец. высших учебных заведений / В. А. Белошапкова, Е. А. Брызгунова, Е. А. Земская и др. 3-е изд., испр. и доп. М.: Азбуковник, 1997.

УДК 947

Исторические науки и археология

Статья раскрывает процесс взаимоотношений между российской властью в лице местной оренбургской администрации и казахами Младшего Жуза, принявшими российское подданство в первой половине XVIII века. Основное внимание в работе автор уделяет военно-политическому аспекту утверждения российской власти на территориях казахских племен Младшего Жуза в первой четверти XIX века.

Ключевые слова и фразы: казахи; оренбургская администрация; пограничная линия; баранта; казачество; военные экспедиции.

Сыровегин Игорь Эдуардович

Магнитогорский государственный технический университет им. Г. И. Носова
isyrovieghin@mail.ru

ОРЕНБУРГСКИЕ ВЛАСТИ И КАЗАХИ МЛАДШЕГО ЖУЗА (ПЕРВАЯ ЧЕТВЕРТЬ XIX ВЕКА)[©]

К началу XIX века внешнеполитические позиции России на юго-восточном направлении можно характеризовать как достаточно прочные и опиравшиеся на уже сложившуюся систему военного и гражданского управления. Основа русского присутствия на юго-восточных рубежах была заложена в 30-е годы XVIII века во время работы «Оренбургской экспедиции» тайного советника И. К. Кириллова. Основным достижением миссии Кириллова стала закладка Оренбургской пограничной линии, вокруг которой стала складываться вся система управления краем на весь последующий период, вплоть до середины XIX века. Освоение и колонизация Оренбургского края продолжались на протяжении большей части XVIII и первой половины XIX века. 15 марта 1744 года императрица Елизавета Петровна издала указ об образовании Оренбургской губернии, с чего начинается систематическая работа русских властей по освоению этого огромного края [8, с. 9].