МЕТОДЫ АВТОМАТИЗИРОВАННОГО ДИСКУРСИВНОГО АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ В ЗАДАННОМ КОНТЕКСТЕ

Дмитрий Тонконогов, Ирина Артемиева

***Abstract**: в данной работе рассматриваются методы автоматизированного дискурсивного анализа и описывается структура проектируемого программного средства, предназначенного для мониторинга дискурса неструктурированных текстов на английском и русском языках в рамках заданного контекста. Во введении даётся краткое описание состояния в области автоматического извлечения знаний из неструктурированной текстовой информации. В первой части рассматриваются уже применяемые методы для дискурсивного анализа с выделением их преимуществ и недостатков. Также рассматривается их применимость к проектированию требуемого программного средства. Во второй части приводится описание проектируемого программного средства – описывается назначение модулей, их взаимоотношение, а также предполагаемые методы решения задач. Приведены предполагаемые для использования методы дискурсивного анализа. В заключении описывается текущее состояние проектируемого программного средства, указываются модули, требующие дополнительное исследование и проектирование.*

***Keywords**: discourse analysis, linguistics, intellectual system for discourse analysis.*

***ACM Classification Keywords**: H.3.3 Information Search and Retrieval*

Введение

Вопросы автоматического извлечения знаний из неструктурированной текстовой информации, бесспорно, являются актуальными в современном мире, что объясняется необходимостью решения практических задач мониторинга больших информационных потоков в сетевом дискурсе с целью их адаптивного агрегирования. В настоящее время происходит стабильно высокий рост количества информационного контента в мире, что привело к появлению такого понятия, как «информационное общество».

Одной из причин такого роста является повсеместное внедрение социальных средств общения в сети Интернет. Сама концепция развития сети Интернет, названная Web 2.0 [O'Reily, 2005], подразумевает глубокую социализацию сетевых ресурсов. Теперь каждый пользователь сети может высказать свое мнение относительно любого события, мнения, факта и т.д.

Если не учитывать некоторые искажающие факторы (например, сетевые тролли, провокаторы, агитаторы), то можно заметить, что социальная сетевая паутина может быть использована в политике для получения информации о некотором обществе – его настроении, требованиях, тенденциях. В настоящий момент такую работу выполняют лингвисты, психологи, социологи [Gee, 2006], и на волне информационного общество привело к развитию направления дискурсивного анализа(discourse analysis).

Очевидно, что за приемлемое время специалисты могут провести дискурсивный анализ только сравнительно небольшого сообщества, посему задача автоматизации для данного направления стоит

особо остро. Но автоматизацию усложняет тот факт, что анализ проводится по артефактам обычного человека из сети, то есть неструктурированной информации (запись из блога, твит, комментарий).

Самым простым промежуточным решением являются системы извлечения знаний. Существует ряд разработок - анализаторов текстов на естественном языке, которые способны обеспечивать процесс извлечения знаний из текстов на русском и английском языках. В силу лингвистической направленности системы обработки связных текстов называют лингвистическими процессорами [Шаров, 1997]. Однако существующие анализаторы ограничены в своих функциях из-за направленности исключительно на извлечение информации, при этом задача хранения знаний и поиска по ним остается нерешенной (для этих целей используются внешние модули). Для целей обработки неструктурированной текстовой информации, направленной на повышение эффективности использования текстов, прибегают к методам построения формальной объектной структуры [Тригуб, 2004].

Однако стоит заметить, что задача поиска и извлечения знаний должна происходить в каком-то определённом контексте. В случае ручного дискурсивного анализа такая задача выполняется на стадии поиска неструктурированной информации, но в случае автоматизированных систем это уже будет невозможно сделать на уровне, достаточном для игнорирования этого в последующем анализе.

Цель настоящей работы – изучение задачи и рассмотрение подходов к созданию интеллектуальной программной системы, предназначенной для мониторинга больших информационных потоков в сетевом публицистическом дискурсе на английском и русском языках для их анализа, извлечения требуемой информации и обобщения этой информации в соответствии с заданными контекстами.

Обзор методов автоматизированного дискурсивного анализа

Общей особенностью большинства реализованных систем является использования ими электронного тезауруса для английского языка WordNet. Этот тезаурус является одним из самых полных для английского языка, также в нем присутствует множество различных отношений между элементами.

В работе, выполненной в ВШЭ [Градосельская, 2011], алгоритм основан на работе с зерновыми концептами. Утверждается, что алгоритм структурирования экспертных текстов при помощи зерновых концептов можно разбить на следующие этапы:

- задаем зерновые концепты статьи - основные смысловые термины;
- проводим предварительный дискурс-анализ, где основными «центрами напряжения» становятся зерновые концепты;
- остальные концепты добавляются только в том случае, если они обеспечивают связь (посредничество) между зерновыми концептами;
- укрупнение посредничающих концептов в «смысловые гнезда».

Автоматизация алгоритма была проведена с помощью пакета программ Matlab. В данной работе никак не рассмотрена возможность накопления зерновых концептов, для последующего проведения поиска по ним. Также для анализа используются структурированные экспертные тексты. Рассматривать работу с точки зрения ограничения по контексту не совсем корректно, поскольку контекст был задан соответствующим фильтром экспертных текстов, проведенным до исполнения алгоритма.

В ряде зарубежных источников данная задача ставится как задача классификации с использованием соответствующей математической модели Precision-Recall [Stede, 2008]. На выходе такого классификатора будет ряд текстов с коэффициентом корреляции текста относительно заданного дискурса. При такой постановке задачи нет необходимости уделять внимание особенностями

лексического, синтаксического анализа языков, но основная сложность перемещается в другую область – необходимо выделить максимально правдоподобные признаки, по которым можно классифицировать текст [Theijssen, 2007].

Признаки, при этом, организуют в следующую устоявшуюся структуру:

- поверхностные признаки;
- синтаксические признаки;
- лексические признаки;
- ссылочные признаки;
- дискурсивные признаки.

Также в настоящее время существуют так называемые комбинированные работы – в них присутствует, как и элементы лингвистики (корпусный анализ, анализ лингвистических концептов), так и подходы, имеющие в своей основе алгоритмы машинного обучения [Hilbert et al., 2010].

В качестве примера можно привести пакет программ GATE. Он представляет собой мощную систему для всевозможного анализа текстов [Gate, 2011]. В данном случае проблемой является его недискурсивная направленность – с помощью данного пакета можно провести свой сколько угодно глубокий разбор, вплоть до использования внутренних хранилищ и лингвистических онтологий, но для получения статистической информации необходимо создавать свой модуль с достаточно сложной семантикой.

Можно сделать вывод, что общей нерешенной к настоящему времени проблемой для всех рассмотренных подходов является отсутствие строго заданного и изменяемого контекста, в рамках которого происходит дискурсивный анализ. Также не развиты подходы к накоплению и поиску информации – большинство подходов предполагают использование только неструктурированных текстов на входе. Тем самым, создание программной системы, предназначенной для мониторинга больших информационных потоков в сетевом публицистическом дискурсе на английском и русском языках для их анализа, извлечения требуемой информации и обобщения этой информации в соответствии с заданными контекстами, является актуальной задачей.

Основные компоненты программной системы

Рассмотрим теперь архитектуру программного средства, построенного на базе методов дискурсивного анализа, предназначенного для обработки неструктурированных текстов на английском и русском языка.

Основными модулями проектируемой системы являются модули обработки неструктурированных текстов, извлечения данных и семантическое хранилище. Модуль обработки текстов с помощью вспомогательной функции оценки анализирует тексты и преобразует их в вид, пригодный для хранения в семантическом хранилище, с сохранением оценки. Модуль извлечения данных позволяет, основываясь на заданном контексте предметной области составить запрос к семантическому хранилищу. Поиск производится по всему множеству документов, и результаты возвращаются для дальнейшего анализа в модуль извлечения.

Перед началом использования системы пользователи-эксперты должны зафиксировать контекст на основе заданной предметной области. Это может быть сделано с помощью онтологии либо графа, в котором будут указаны связи между терминами. Исходя из заданного контекста, формируется два промежуточных представления, предназначенных для использования в модуле поиска неструктурированной информации и модуле извлечения знаний. Использование контекста при поиске текстов позволяет снизить нагрузку на модуль извлечения данных – не потребуется реализовывать

дополнительную логику вычисления подмножества текстов, на основе которых будет производиться анализ дискурса.

Модуль обработки текстов должен использовать методы компьютерной лингвистики для извлечения семантики документов с учётом последующего переноса в хранилище. Одним из вариантов такого извлечения можно считать метод разметки корпусов – лемматизация, тем более что в настоящее время уже достаточно сильно развиты программные средства с полнотекстовым поиском. В то же время для модуля извлечения знаний можно применить методы машинного обучения для определения релевантности текста заданному контексту. После этого происходит дискурсивный анализ документов, результаты которого возвращаются в модуль извлечения данных.

Данная структура позволяет не фиксировать визуальное отображение результатов дискурсивного анализа – вместо этого предлагается использовать один из модулей для отображения данных в подходящем виде. На вход данного модуля будут передана полученная информация о дискурсивном анализе текстов, на выходе ее можно представить как облако тегов, дерево, граф отношений и т.д.

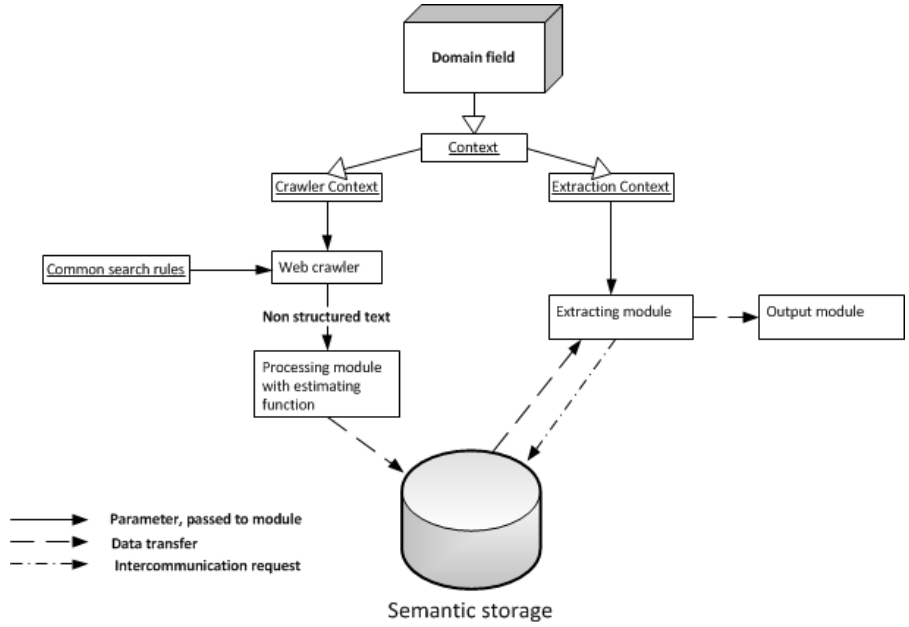

Рисунок 1. Структура интеллектуальной системы

Другой особенностью является возможность накопления информации об анализируемых текстах, что дает возможность получать информацию без анализа всех текстов, а также производить анализ новых текстов без прохождения повторного анализа для уже проанализированных текстов.

Для практического применения в качестве системы мониторинга необходимо реализовать подсистему поиска неструктурированных текстов в сети Интернет – «паука». Основываясь на относительно несложных правилах поиска, паук автоматически добавляет новые документы в семантическое хранилище. В таком случае достигается автономность системы – пользователь задает только общие правила поиска неструктурированных текстов и контекст предметной области.

Выделяется несколько групп пользователей:

- администратор;
- эксперты (лингвисты);
- инженеры знаний;
- прикладные пользователи.

Администратор отвечает за регистрацию пользователей в системе, определяя их «роли». Эксперты (лингвисты) отвечают за формирование множества лексем для определения контекста. В их задачу также входит совместное с инженерами знаний формирование лингвистических моделей классов и определение наборов ментальных признаков и их значений. Прикладными пользователями являются различные аналитики текстов, формулирующие свои задачи для обработки текстов.

Заключение

Как показал анализ уже реализованных систем, в настоящее время не существует систем для дискурсивного анализа неструктурированных текстов в заданном дискурсе. Наиболее близким к искомому оказался программный пакет GATE Университета Шеффилда, но недостатки (отсутствие семантического извлечения информации, отсутствие задаваемого контекста) делают его использование в данной области невозможным.

В работе рассмотрен подход к созданию интеллектуальной программной системы, предназначенной для мониторинга дискурса неструктурированных текстов на английском и русском языках в рамках заданного дискурса. Выделены основные задачи, которые должны быть решены программной системой, а также ее основные информационные и программные компоненты.

Однако для реализации данной программной системы необходимо провести ряд исследований для описания каждого модуля. Особенно остро стоит вопрос выбора методов для модуля обработки текстов и извлечения данных. В настоящей работе приведены только общие методы. В процессе дальнейшего проектирования они могут и должны быть пересмотрены. Но в любом случае для выбранных методов должны быть предоставлены экспериментальные доказательства применимости в указанной программной системе.

Список литературы

[Gate, 2011] Gate General Architeture overview - The University of Sheffield [электронный ресурс] // Режим доступа: http://gate.ac.uk/overview.html

[Gee, 2006] Gee, J. P. An Introduction to Discourse Analysis: Theory and Method. London: Routledge. UK, MPG book ltd, 2006

[Hilbert et al., 2010] Hilbert M., Lobin H., Barenfanger M., Lungen H., Puska C. A Text-technological Approach to Automatic Discourse Analysis of Complex Texts - Institut fur Germanistik Arbeitsbereich Angewandte Sprachwissenschaft und Computerlinguistik Justus-Liebig-Universitat Gießen, 2010 г. – 4 с.

[O'Reily, 2005] Tim O'Reily. What Is Web [электронный ресурс] // Режим доступа: http://oreilly.com/web2/archive/what-is-web-20.html

[Stede, 2008] Stede M. Local Coherence Analysis in a Multi-Level Approach to Automatic Text Analysis - Lexical-Semantic Resources in Automated Discourse Analysis, Volume 23, Number 2, 2008 - ISSN 0175-1336

[Theijssen, 2007] Theijssen D. Features for automatic discourse analysis of paragraphs - Department of Linguistics Radboud University Nijmegen, 2007 – 125c.

[Градосельская, 2011] Градосельская Г.В. Сетевой анализ постсоветского информационного пространства: перспективы разработки методологии - Сборник статей памяти А. Крыштановского. НИУ ВШЭ, РОС, ИС РАН. М.: НИУ ВШЭ, 2011. — VIII, 557 с. НИУ ВШЭ

[Тригуб, 2004] Тригуб Н.А. Система обработки неструктурированной текстовой информации на основе объектного подхода для повышения эффективности информационного поиска: автореферат дис…кандидата технических наук. – М., 2004. -189 с.

[Шаров, 1997] Шаров С. А. Инструментальная система для разработки лингвистических процессоров: Автореф. … дисс. к. ф.-м. наук. – М.: 1997. -27 с.