OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.8 + 004.9

СОЗДАНИЕ ХРАНИЛИЩА НАУЧНЫХ ПУБЛИКАЦИЙ С ПРИМЕНЕНИЕМ МЕТОДОВ КОРПУСНОЙ ЛИНГВИСТИКИ

Елохов Е.С., Югов А.С., Ланин В.В.

Национальный исследовательский университет «Высшая школа экономики» – Пермь, г. Пермь, Российская Федерация

eugene.yelokhov@gmail.com, yugovas@live.ru, lanin@perm.ru

В работе рассматривается подход к созданию удобного хранилища публикаций, основанный на методах поиска и обработке неструктурированных текстов (научных публикаций) с применением программного комплекса GATE.
Ключевые слова: корпус документов; выделение ключевых понятий; аннотирование текста; GATE.

Введение

Число доступных научных публикаций растѐт с каждым днем. Это объясняются прежде всего тем, что Интернет сделал доступными большое число работ, растѐт число и самих электронных библиотек (РИНЦ, Springer, ACM и т.д.).

В связи с этим все более и более возрастает сложность задачи поиска статей по интересующей исследователей тематике. Для увеличения скорости поиска информация должна быть систематизирована. Распространенной технологией упорядочивания информации является рубрикация документов, то есть описание содержания документа через элементы некоторого замкнутого списка тем – рубрикатора [Петровский, Глазкова, 2007]. В больших информационных коллекциях (от библиотеки статей факультета, которая становится достаточно объемной через несколько лет работы факультета, до крупных хранилищ, электронных библиотек), имеет смысл говорить только об автоматической рубрикации.

Примерами систем рубрикации электронных документов являются Softinform Search Server (http://www.softinform.ru) и модуль рубрикации текстов в системе asknet (http://asknet.ru/). В большинстве из них применяются статистические методы, методы сравнения с образцом и методы поиска по ключевым словам. Однако статистические подходы малоэффективны по временным показателям, а метод ключевых слов зачастую выдает неполные или, напротив, недостоверные результаты. Становится очевидным, что решение задачи рубрикации невозможно без использования метазнаний.

Особенно актуальными становятся задачи семантического поиска по корпусу документов и автоматической семантической разметки документов для дальнейшего использования этой разметки при поиске.

1. Описание задачи

В рамках проекта по созданию системы поиска и аналитической обработки публикаций по методам и средствам моделирования информационных систем возникла задача выбора инструментальных средств обработки неструктурированных текстов. В качестве требований к таким системам были сформулированы следующие требования: расширяемость, поддержка русского и английского языков, возможность работы с тезаурусами и другими онтологическими ресурсами. После анализа существующих систем выбор пал на программный комплекс GATE (General Architecture for Text Engineering, сайт проекта http://gate.ac.uk).

При анализе публикаций с использованием возможностей GATE планируется решить следующие задачи:

– организация аннотированного хранилища ресурсов (статей);

– реализация механизма автоматического выделения ключевых слов текста;

– реализация механизма автоматического структурного анализа публикаций в произвольных форматах (выделение основных структурных элементов статьи: аннотации, введения и т.д.);

– выявление разнообразных связей между публикациями.

Очевидно, что все перечисленные задачи не покрываются функциональными возможностями GATE, поэтому планируется реализовать их путем разработки собственных решений с использованием GATE API.

2. Инструментарий GATE

GATE – это пакет инструментов Java для разработки и развертывания компонентов программного обеспечения, предназначенных для обработки естественного языка. Данная система с открытым исходным кодом подходит для любых операций обработки текстов различных размеров.

Необходимо отметить, что в лингвистических ресурсах, которые используются при работе с GATE, содержится три типа данных: документы, корпуса и аннотации. Аннотации представлены в виде графов, которые моделируются как Java-наборы. Имеется возможность работать с различными форматами документов (XML, RTF, HTML, SGML, текстовый файл). В каждом случае анализируемый текст конвертируется в единую унифицированную модель аннотации. Формат аннотации является видоизмененным TIPSTER-форматом [Grishman, 1997], который, в свою очередь, совместим с Atlas-форматом [Bird and Liberman, 1999] и использует механизм «отдельно хранимой разметки». Документы, корпус документов и аннотации хранятся в базах данных, визуализируются с помощью инструментов среды разработки. Эти типы данных доступны на уровне кода в фреймворке.

Кроме того, GATE располагает интеллектуальной системой под названием ANNIE (A Nearly-New Information Extraction System), которая включает в себя набор ресурсов, обеспечивающих токенизацию, POS-тэггинг, разбивку на предложения, извлечение именованных сущностей и анализ кореферентности.

Таким образом, GATE предоставляет набор инструментов для первичной обработки текстов на естественном языке, которые позволяют структурировать текст и подготовить его к дальнейшему более сложному анализу.

3. Реализация ИС для хранения и обработки научной литературы

3.1. Создание хранилища документов

Первым этапом обработки корпуса является создание единого хранилища документов и наполнение этого хранилища. На этом этапе необходимо обеспечить удобное и эффективное ведение (хранение и добавление) исходных документов.

Для применения GATE в дальнейшем на этом этапе требуется создать отдельный каталог для хранения рубрицированных заранее экспертами документов.

3.2. Лингвистическая разметка документов корпуса

Лингвистическая разметка – одно из основных понятий корпусной лингвистики. Разметка даèт возможность идентифицировать тексты по различным параметрам, позволяя осуществлять осмысленный поиск по корпусу. Разметка текста заключается в приписывании текстам и их компонентам дополнительной информации (метаданных).

Метаданные должны удовлетворять ряду требований, «семи максимам» Д. Лича [Leech, 1993] (Leech's seven maxims of annotation). Разметка должна быть независима от текста: должна быть возможность убрать разметку и просмотреть текст без неè и, наоборот, вычленить только разметку. Принципы разметки и их разработчики должны быть известны конечному пользователю. Пользователь должен быть поставлен в известность о том, что разметка не является безошибочной, а представляет собой лишь потенциально полезный инструмент. В основу разметки должны быть положены общепринятые и, по возможности, теоретически нейтральные лингвистические принципы. И, наконец, ни одна разметка не может априорно считаться стандартом.

Реализацию базовой разметки предоставляет готовый функционал комплекса GATE: токенизация и выделение абзацев, предложений, слов на ее основе; морфо-синтаксический анализ (определение части речи).

Более сложную разметку (библиографические данные, сведения об авторе и т.п.) можно реализовать при помощи доработки инструментов GATE.

3.3. Выделения ключевых понятий в документе

Набор ключевых слов формирует представление о работе в первом приближении и является важной характеристикой, показывающей, стоит ли данную конкретную работу рассматривать более детально или можно сразу сделать вывод о ее нерелевантности. Поэтому важно, чтобы каждый научный текст был охарактеризован таким образом.

Выделение ключевых понятий не входит в базовый набор функционала программной системы GATE, поэтому планируется реализовать дополнительный модуль с использованием GATE API.

На данный момент для поиска ключевых слов используются закономерности, открытые Ципфом в чистом виде (TF-IDF), либо алгоритмы LSI (latent semantic indexing). В рамках текущего исследования остановимся на первом варианте, как на наиболее простом для реализации исследовательского прототипа.

Чаще всего в качестве ключевых слов выступают устойчивые словосочетания [Белоногов и др., 2002],

поэтому расширим выбранный метод выделения ключевых слов на основе частотных характеристик при помощи добавления алгоритма выделения словосочетаний, в основе которого лежит последовательное вычисление частотных характеристик ($L = 2, 3, \ldots, L\_max$) и фильтрация повторяющихся L-грамм по критерию устойчивости [Гусев и Саломатина, 2004].

3.4. Применение GATE для автоматического аннотирования текстов

Как правило, процессы анализа текста состоят из одних и тех же этапов. Ключевым моментом является постановка задачи извлечения, причем аннотаторам (в роли которых выступают люди) следует решить эту задачу подобным образом. Кроме того, необходимо создать высококачественный образец, который задаст направление разработки и измерения результатов автоматизированного анализа. Довольно распространено использование двойных или тройных аннотаций, когда несколько человек занимаются задачей аннотирования независимо друг от друга, после чего измеряется уровень сходства (так называемое «меж-анотаторское сходство») с целью оценки и контроля качества предоставленных данных.

Рассмотрим основные этапы процесса анализа текста, а также инструменты GATE, которые используются на каждом этапе.

На первом этапе необходимо собрать множество текстов, на основе которых будут созданы аннотации. Примерами таких текстов могут являться научные статьи, истории болезней, технические задания, отчеты по клиническим испытаниям, электронные письма, сообщения из социальных сетей, парламентские акты и т.д.

Далее составляется структурированное описание интересующих пользователя тем текста. Например, описание может быть представлено в виде корпоративной телефонной книги или списка названий препаратов. Таким образом, будет сформирована «черновая» онтология.

После этого нужно точно поставить задачу аннотирования и удостовериться в правильности этой постановки. Рекомендуется использовать инструмент GATE Teamware (или GATE Developer для небольших проектов) для того, чтобы задать «золотой стандарт» набора аннотаций корпуса, относящихся к данной онтологии.

Кроме того, необходимо разработать прототип цепочки процессов анализа текста. Рекомендуется использовать инструмент GATE Developer для построения цепочки процессов с целью автоматизированного аннотирования и оценки результата работы, сопоставляя результат с образцовым набором аннотаций.

Следующим шагом является развертка и удостоверение в работоспособности системы анализа. Рекомендуется применить цепочку процессов анализа текста к анализируемому корпусу с использованием GATE Cloud или внедрить ее в систему, используя GATE Embedded.

Важным этапом является наполнение сервера индексного поиска информации. Рекомендуется использовать GATE Mímir для хранения аннотаций, относящихся к онтологии в мультипарадигматическом сервере индексного поиска информации.

В заключении необходимо предоставить результаты конечным пользователям одним из перечисленных ниже способов: выгрузить данные для анализа в статистические пакеты, базы данных и т.д.; создать интерфейс пользователя, специфический для этой предметной области, с целью эффективного использования GATE Mímir, или заняться интеграцией в существующую фронтальную систему через веб API со стилем построения REST.

Очень часто используется итеративная разработка (некоторые этапы или группы этапов повторяются) и интеграционное тестирование в рамках agile-методов [Bizer, Heath, and Berners-Lee, 2009].

При использовании предложенного подхода возможен поиск по корпусу документов с использованием аннотаций или онтологий. Кроме того, имеет место процесс поддержки системы и предусмотрена вероятность изменения информационных потребностей пользователя. В каждом из случаев используется справочник или полуавтоматическое аннотирование, а также автоматизированные измерения и регрессионное тестирование (для проверки стабильности существующих результатов анализа или для структурирования проведения анализов в будущем).

Заключение

В данной работе был описан подход к реализации информационной системы, реализующей функции хранилища научных публикаций. Кроме непосредственно функции хранения документов ИС реализует функции формирования метаданных, служащих для повышения скорости и точности поиска как информации по документам, так и самих документов. В качестве метаинформации выступает набор ключевых понятий и аннотация.

В дальнейшем планируется реализация системы для автоматического построения активных ссылок на другие документы (если они тоже находятся в системе) и построение карты публикаций.

Работа выполнена при поддержке Научного фонда НИУ ВШЭ по программе софинансирования грантов РФФИ (проект № 13-09-0143).

Библиографический список

[Белоногов и др., 2002] Белоногов, Г.Г. и др. Автоматический концептуальный анализ текстов // НТИ, 2002. сер. 2. № 10. С. 26–32.

[Гусев и Саломатина, 2004] Гусев В.Д., Саломатина Н.В. Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Труды международной конференции Диалог'2004. М.: Наука, 2004. С. 530 – 535.

[Петровский, Глазкова, 2007] Петровский М., Глазкова В. Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов // Вычислительные методы и программирование. — 2007. — № 8. — С. 57–69.

[Bird and Liberman, 1999] Bird S., Liberman. M. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, 1999. http://xxx.lanl.gov/abs/cs.CL/9903003.

[Bizer, Heath, and Berners-Lee, 2009] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22. doi:10.4018/jswis.2009081901

[Grishman, 1997] Grishman. R. TIPSTER Architecture Design Document Version 2.3. Technical report, DARPA, 1997. http://www.itl.nist.gov/div894/894.02/related_projects/tipster/.

[Leech, 1993] Leech, G. 1993. Corpus annotation schemes. Literary and Linguistic Computing, 8(4):275–281.

PUBLICATIONS REPOSITORY CREATION BASED ON CORPUS LINGUISTICS TECHNIQUES

Elokhov E.S., Yugov A.S., Lanin V.V.

National Research University Higher School of Economics

eugene.yelokhov@gmail.com, yugovas@live.ru, lanin@perm.ru

The purpose of this work is to analyze an approach to publications repository creation based on search methods & processing of unstructured data (scientific papers) using GATE, a Java suite of tools.

Introduction

A problem of complicated search for specific scientific papers arises impacting people that intend to use accumulated knowledge. In order to increase data processing speed the data is to be systematized. The common way to achieve it is to implement the process of categorization, so text content is to be grouped into categories. It is worth bearing in mind that an automated process of categorization is needed in large data collections that range from faculty publications to works in Russian Science Citation Index. At present there's a deficit in software tools of e-documents categorization. Most of them use statistical methods, pattern matching techniques and methods of searching by keywords. Nevertheless statistical methods are ineffective in terms of time, and the results of searching by keywords implementation are often incomplete or excessive. It can be clearly seen that the only cure for this problem is using metaknowledge.

Main Part

E-documents search and analytic processing system creation comprises a number of tasks. The principal objectives to be reached are as follows. Firstly, publications warehouse should be created. Secondly, keywords automatic extraction technique is to be implemented. Moreover, publications structural elements extraction technique should be created. In addition, relations between publications must be detected.

In order to achieve this ultimate aim it was decided to use GATE (a Java suite of tools for all sorts of natural language processing tasks, including information extraction). Particularly semantic annotations are going to be used. Keywords extraction is going to be based on Zipf's law extended with collocations extraction. Furthermore, it will be possible to order publications by subject based on keywords extraction and annotations.

Conclusion

The main conclusion to be drawn is that the information system that performs a function of publications repository was described. It also performs functions of metadata collection. Metadata comprises a set of keywords and an annotation. It should be noted that metadata helps to increase a search speed and improve search accuracy. In the future in the context of the system it will be possible to automatically build active links to other documents and to establish a publication scheme.