Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.896+004.853

Семеняка
Анатолий Фёдорович

Средства извлечения знаний из исторических текстов с применением методов естественно-языковой обработки

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1-31 80 10 «Теоретические основы информатики»

Научный руководитель

Захаров Владимир Владимирович

Кандидат технических наук, доцент кафедры ИИТ

Минск 2020

КРАТКОЕ ВВЕДЕНИЕ

В предметной области истории источники теснейшим образом взаимосвязаны. Каждый отдельный источник – часть совокупности, что и определяет некоторые моменты их содержания, новизны или повторяемости информации. Большинство исторических источников - письменные, многие из которых доступны в электронном виде. Однако заключенная в них полезная информация зачастую не структурирована или структурирована слабо, а значит, ее невозможно обработать и проанализировать классическими вычислительными методами и средствами. Исследователям может быть сложно ориентироваться и искать необходимую информацию в таких больших объемах данных. В настоящее время, вместе с увеличением доступности исторических текстов в цифровой форме растет потребность в обработке такой текстовой информации с целью извлечения содержащихся в ней знаний, получении на её основе новых знаний, посредством логического вывода, а также повышении качества и эффективности имеющихся методов такой обработки.

Этим требованиям наиболее полно соответствуют интеллектуальные системы в области истории. Базы знаний таких систем требуют постоянного пополнения, особенно в сферах, апеллирующих к большим объёмам информации, таких как история.

При наполнении базы знаний инженерами по знаниям вручную, им приходится руководствоваться и по-своему интерпретировать документацию и принятые соглашения о представлении знаний в наполняемой базе знаний. Проводить соответствия между исследуемыми сущностями и их знаками в пополняемой БЗ приходится также вручную, инициируя соответствующие поисковые запросы с определёнными аргументами, для каждой предполагаемой сущности. Кроме того, инженер по знаниям всегда должен вручную верифицировать получившиеся формальные конструкции.

Такая работа зачастую шаблонна. Схожие естественно-языковые конструкции, встречающиеся в разных текстах, могут быть интерпретированы по одному и тому же принципу. То есть информация об интерпретации определённых фрагментов текстов, соответствующих определённым моделям, может быть формализована и использована многократно, тем самым снижая трудоёмкость процесса пополнения баз знаний в дальнейшем.

В связи с этим возникает необходимость разработки компонента автоматического извлечения знаний из исторических текстов и пополнения базы знаний.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью данной диссертации станет снижение трудоёмкости и сроков разработки и пополнения баз знаний интеллектуальных систем в области истории за счёт автоматизации этого процесса посредством автоматического извлечения знаний из исторических текстов с применением методов естественно языковой обработки.

Для достижения поставленной цели в рамках диссертации необходимо выполнить следующие задачи:

– анализ литературы по теме исследования в области применения методов естественно-языковой обработки к историческим текстам;

– анализ существующих подходов и средств к обработке естественно-языковых и исторических текстов;

– проектирование компонента автоматического извлечения знаний из исторических текстов и пополнения базы знаний;

– реализация компонента автоматического извлечения знаний из исторических текстов и пополнения базы знаний;

– оценка и верификация полученных результатов.

Объектом исследования являются исторические тексты. Предмет исследования - модели и методы их обработки.

Все результаты, приведенные в диссертации, получены соискателем самостоятельно на основе изучения литературы, моделей, средств и методов извлечения знаний из исторических текстов с применением методов естественно-языковой обработки. Вклад научного руководителя В. В. Захарова связан с постановкой цели, задач исследования, анализом возможных путей решения и оценкой результатов.

Результатом магистерской диссертации является реализованный компонент автоматического извлечения знаний из исторических текстов, интегрированный в прикладную интеллектуальную систему по истории города Минска, и использующийся для пополнения её базы знаний.

Положения диссертационной работы докладывались и были напечатаны в сборнике 55-ой научной конференции аспирантов, магистрантов и студентов учреждения образования БГУИР и сборнике научно-технической конференции OSTIS-2018.

Диссертация состоит из введения, трёх глав, заключения, библиографического списка. Общий объем работы составляет 84 страницы, из которых основного текста 68 страниц, 32 иллюстрации на 12 страницах, 3 таблицы, библиографический список из 42 наименований на 4 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В работе рассматриваются преимущества использования онтологического подхода, совместно с подходом, основанным на правилах, для извлечения знаний из текстов предметной области истории и последующей их интеграции в прикладную систему по истории города Минска, основанной на технологии OSTIS (Open Semantic Technologies for Intelligent Systems = Открытые семантические технологии проектирования интеллектуальных систем).

Во введении определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы.

В первой главе рассмотрены существующие исследования в области автоматической обработки исторических текстов. Выполнен анализ существующих подходов и средств к обработке естественно-языковых текстов и изучаются особенности их применения к историческим текстам. Произведена постановка задачи по созданию компонента автоматического извлечения знаний из исторических текстов и пополнения базы знаний. Предложен онтологический подход, совместно с подходом, основанным на правилах и описаны преимущества его использования для решения поставленной задачи.

Во второй главе рассмотрена система, для которой проектируется компонент, описана структура компонента, описан алгоритм его работы, все его модули и их взаимодействие друг с другом, а также взаимодействие с инженером по знаниям и экспертами.

В третьей главе описана реализация спроектированного компонента. База знаний прикладной системы, для которой проектировался компонент, построена по технологии OSTIS. Компонент реализован в среде и средствами данной технологии. Согласно предложенному подходу, для реализации спроектированного компонента, в рамках машины обработки знаний пополняемой прикладной системы, созданы программные агенты, реализующие алгоритм работы спроектированного компонента. В базу знаний системы добавлены шаблоны извлечения, разработана лингвистическая часть базы знаний, включающая морфологические, синтаксические и лексические описания, связывающие естественно-языковую и предметную часть. Реализованный компонент подвергся апробации и верификации.

ЗАКЛЮЧЕНИЕ

Разработанный компонент автоматического извлечения знаний из исторических текстов и пополнения базы знаний осуществляет свою работу, последовательно обрабатывая исходный текст на естественном языке различными модулями лингвистического анализатора, за счёт использования онтологических правил и шаблонов извлечения. На всех этапах извлечения, верификации и интеграции компонент учитывает временной, пространственный контекст и существующие знания в пополняемой базе знаний.

Все этапы работы компонента происходят в одной среде, что облегчает налаживание процесса взаимодействия между этими этапами.

Благодаря предложенному подходу к планированию событий появления недостающих фактов в памяти компонента, он может разрешать неоднозначные текстовые конструкции даже после того, как автоматическая обработка источника завершилась (например при обработке других источников и выявления недостающей ранее информации из них).

За счёт пополнения базы знаний, в среде которой работает компонент, новыми знаниями, расширяются и знания о предметной области и лингвистическая составляющая.

Система, построенная по предложенному подходу, хорошо масштабируется и в рамках обработки новых языков и для обработки текстов новых предметных областей.

Разработанный компонент, за счёт использования шаблонов и правил, хорошо подходит для обработки исторических текстов со специфичной лексикой или частично структурированных источников. Кроме того, с его помощью можно извлекать знания из источников на разных языках, или мультиязычных источников.

Данный подход пригоден для построения такого рода компонентов в любых интеллектуальных системах, использующих модель общей семантической памяти и многоагентном подходе работы с ней.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Семеняка, А. Ф. Извлечение знаний из исторических текстов / А. Ф. Семеняка // Информационные технологии и управление: 55-я научная конференция аспирантов, магистрантов и студентов учреждения образования «Белорусский государственный университет информатики и радиоэлектроники» (Минск, 22 -26 апреля 2019 г.) / редкол.: Л. Ю. Шилин [и др.]. –Минск: БГУИР, 2019. –с. 10

2. Бойко, И. М. Приобретение знаний на основе текстов естественного языка / И. М. Бойко, А. Н. Гордей, А. В. Губаревич, А. Ф. Семеняка // Открытые семантические технологии проектирования интеллектуальных систем : материалы междунар. науч.-техн. конф., Минск, 15–17 февр. 2018 г. / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: В. В. Голенков (отв. ред.) [и др.]. — Минск, 2018. — Вып. 1. — с. 199–206.