А. Ф. Семеняка

ИЗВЛЕЧЕНИЕ ЗНАНИЙ ИЗ ИСТОРИЧЕСКИХ ТЕКСТОВ

Рассматриваются существующие проблемы в области извлечения знаний из исторических текстов. Предлагается онтологический подход с привлечением экспертов в области истории и лингвистики.

Введение

В реальной исторической действительности источники, являющиеся документами определенного исторического времени, теснейшим образом взаимосвязаны. Каждый отдельный документ – часть этой совокупности, ее элемент, что определяет некоторые моменты его содержания, новизны или повторяемости информации. Каждая историческая эпоха порождает разные типы источников. Большинство исторических источников - тексты. Однако заключенная в них полезная информация зачастую не- или слабо структурирована, а значит, ее невозможно обработать и проанализировать классическими вычислительными методами и средствами.

В настоящее время имеется всё возрастающая потребность в обработке такой текстовой информации, повышении качества и эффективности имеющихся методов обработки текстов, а также извлечении содержащейся в них знаний и, основанного на этих знаниях, логического вывода. Однако эта задача осложняется проблемами обработки такого рода текстов.

I. Проблемы извлечения знаний из исторических текстов

Обработка исторических текстов связана с решением следующих проблем: *Разное и противоречивое написание; различный словарный запас; разная и противоречивая морфология; более длинные предложения по сравнению с современными текстами; непоследовательное использование знаков препинания; различный синтаксис и непоследовательный порядок слов; переключение в тексте с одного языка на другой язык или диалект и обратно; существенные различия между текстами из разных периодов времени, жанров и авторов;*

II. Предлагаемый подход к извлечению знаний из исторических текстов

Решение вышеперечисленных проблем предлагается выполнить при помощи онтологического подхода, основанного на лингвистической и предметной базах знаний с привлечением экспертов в области истории и лингвистики. Схема их взаимодействия с компонентами системы извлечения знаний представлена на рисунке 1.

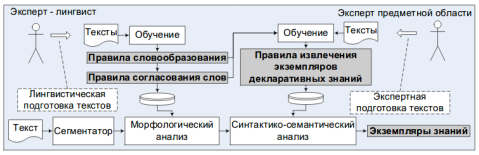

Рис. 1 – Схема взаимодействия экспертов с компонентами системы извлечения знаний

Темными блоками выделены знания, извлекаемые из текстов на каждом из трех этапов: формирование правил о языке, формирование правил извлечения и непосредственное извлечение экземпляров. На первом этапе эксперт-лингвист подбирает типовые тексты предметной области. Они пропускаются через процедуру автоматического обучения, которая формирует правила словообразования и согласования слов, учитывая предметные аспекты и вышеперечисленные проблемы. На втором этапе эксперт предметной области выполняет подготовку другого массива текстов. Это подразумевает явное выделение в текстах экземпляров декларативных знаний предметной области истории, которое выполняется в виде аннотирования. Размеченные тексты поступают на вход второй процедуре обучения, которая, используя полученные знания о языке, формирует правила извлечения. На третьем этапе другая выборка текстов, в которой отражены целевые экземпляры, подвергается морфологической и синтактико-семантической обработке с применением правил, созданных предыдущих этапах. Результатом этого являются извлеченные знания представленные в формальном виде.

III. Выводы

Предлагаемый подход позволяет построить систему, осуществляющую извлечение знаний из текстов по истории и использовать их в базах знаний интеллектуальных систем, проведения логического вывода с целью получения новых знаний, на основе уже существующих.

1. Симаков, К. В. Метод обучения модели извлечения знанийи зестественно-языковых текстов / А. М. Андреев, Д. В. Березкин // Вестник МГТУ. – 2007. – №3. – С. 75-94.

Семеняка Анатолий Фёдорович, магистрант кафедры интеллектуальных информационных технологий БГУИР, toli44777@gmail.com.

Научный руководитель: Захаров Владимир Владимирович, кандидат технических наук, доцент, доцент кафедры интеллектуальных информационных технологий БГУИР, zvv2064@mail.ru.