Мельничук С.Ф. *— рецензент Гемба О.В.*
УНК "ИПСА" НТУУ "КПИ"

Представление лингвистической информации в базе знаний системы автоматизации построения онтологий

Задача автоматизации построения онтологии может быть решена с использованием методов обработки текстов на естественном языке, реализация которых требует предоставления лингвистической информации – знаний о языке обрабатываемых текстов. В работе рассматриваются способы построения лингвистических баз знаний, используемых в различных системах автоматической обработки текстов для оценки возможности применения их в процессе построения онтологий.

Процесс обработки текстов на естественном языке может быть представлен как последовательность нескольких этапов: лексического, морфологического, синтаксического и семантического, алгоритмы каждого из которых требуют предоставления лингвистической информации.

Первым этапом является лексический анализ, цель которого состоит в начальной разбивке текста на отдельные слова (лексемы) и группы (фразы, предложения) для последующих этапов. Для реализации лексического анализатора база знаний должна содержать информацию о разделителях слов и предложений (пробельных символах и символах пунктуации), неизменяемых оборотах, а также регулярные выражения, позволяющие распознать определенные элементы текста: адреса электронной почты, фамилии с инициалами, имена собственные, аббревиатуры и т. п.

Морфологический анализ выполняет определение начальной формы каждой лексемы (леммы), части речи и парадигмы, без чего невозможно выявить взаимосвязи между элементами текста. Реализация морфологического анализа русскоязычных текстов строится на использовании грамматического словаря А.А. Зализняка и морфологических словарях. Множество лемм, частей речи и типы парадигм заносятся в таблицы реляционной базы данных, дополняемые множеством правил морфологического разбора и правил словообразования для лемматизации слов, не найденных в словаре.

Входными данными синтаксического анализатора являются результаты морфологического анализа, результатом – дерево разбора каждого предложения. Для выполнения грамматического разбора база знаний содержит словарь моделей управления глаголов, а также множество правил (грамматику) разбора. Также существует подход к реализации синтаксического анализа на основе множества образцов разобранных предложений (базы данных деревьев зависимости) и информации о вероятности применения того или иного дерева зависимости из базы.

На этапе семантического анализа выявляются смысловые отношения между элементами текста на основе тезауруса языка, задающего бинарные отношения на множестве слов (синонимия, антонимия, гипо- и гиперонимия и т. п.). База знаний поддержки данного этапа должна включать соответствующие словари синонимов, омонимов и т. п.

Результатом работы является структура лингвистической базы знаний системы автоматизированного построения онтологий, включающей реляционную базу знаний с информацией различных типов словарей (морфологических, синонимов и т. п.), а также множество правил грамматики и множество регулярных выражений, формирующих анализаторы различных уровней.

Литература

1. Карпов В.А. Язык как система. – Минск: Высшэйшая школа, 1992. – 302 с.
2. Lucja M. Iwanska, Stuart C. Shapiro – Natural Language Processing and Knowledge Representation. – AAAI Press, 2000. – 480 p.