Гемба О.В., Середа А.А.
УНК "ИПСА" НТУУ "КПИ"

Первичный анализ текста на естественном языке в системе автоматизации построения онтологий

Автоматизация процесса построения онтологий включает нескольких этапов, на первом из которых происходит формирование первоначального перечня терминов предметной области и структуры их взаимосвязей на основе лингвистического анализа коллекции текстов, что требует включения в систему определенных модулей лингвистического анализа. В работе рассматриваются основные шаги первичного (предсемантического) анализа текстов на естественном языке (лексический, морфологический и синтаксический), модели и алгоритмы, используемые для их реализации.

Задача лексического анализа состоит в разделении входного текста на отдельные слова и предложения с некоторой первичной обработкой (выделением неизменяемых слов и оборотов, имен собственных, электронных адресов и т. п.). База знаний содержит информацию о разделителях слов и предложений (пробельных символах и символах пунктуации), неизменяемых оборотах, а также регулярные выражения, позволяющие распознать, например, имена собственные. Задача данного модуля анализа решается путем построения детерминированного конечного автомата, формирующего из входного текста аннотированный список лексем.

Морфологический анализ выполняет определение начальной формы каждой лексемы (леммы), части речи и парадигмы, без чего невозможно выявить взаимосвязи между элементами текста. Для поддержки данного модуля база знаний включает данные о правилах словообразования, а также о существующих морфемах (корнях, префиксах, суффиксах) языка и их сочетаемости. Для каждой поступающей на вход лексемы выполняется ее поиск в базе словоформ. Если словоформа не найдена, выполняется разделение лексемы на морфемы и на основе правил из базы знаний делается предположение о лемме и парадигме. Набор правил может задаваться явно (экспертом) либо строиться в рамках самообучаемой системы на основе индуктивной логики.

На этапе синтаксического анализа определяются взаимосвязи между отдельными словами и частями предложений. Результатом является граф, узлами которого выступают отдельные слова. Если два слова в предложении связаны каким-либо образом, то соответствующие им вершины графа также связаны дугами с определенной окраской (например, вопросами, задаваемыми от одного слова к другому). Для выполнения данного этапа анализа необходимо сформулировать правила разбора в виде контекстно-свободной грамматики, отражающей правила согласования слов в именных и глагольных группах, свертки однородных членов предложения, согласования подлежащего и сказуемого и т. п. Порядок применения правил определяется алгоритмом разбора, который в случае неудачи применения очередного правила выполняет откат. В результате может быть получено несколько деревьев разбора, из которых выбирается полносвязное дерево с единственной вершиной.

Результаты перечисленных этапов анализа текста на естественном языке являются входными данными для модуля семантического анализа, а также могут быть использованы в поисковой системе, основанной на статистической обработке информации для повышения релевантности результатов поиска.

12th International conference on science and technology "System analysis and information technologies" SAIT 2010
Institute for Applied System Analysis of National Technical University of Ukraine "KPI", Kyiv, Ukraine, May 25–29, 2010
ISBN 978-966-2153-42-2, revision 1.1 (2010.05.30), available at http://sait.org.ua/books/sait2010.ebook.pdf