ТЕХНОЛОГИЯ ПОСТРОЕНИЯ МНОГОЯЗЫЧНОГО БАЗОВОГО ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА

Чеусов А. В.
кафедра информационных систем управления факультета прикладной математики Белорусского государственного университета, 220030, г.Минск, ул. Ленинградская, 8,
e-mail: cheusov@tut.by

Современное развитие информационных технологий обострило актуальность автоматической обработки текстовой информации. К традиционным задачам информационного поиска, машинного перевода, классического аннотирования и реферирования добавились задачи автоматической классификации текстовых документов, автоматизации инженерии знаний, естественно-языкового интерфейса и cross-language функциональности, прагматического анализа (sentiment analyses или opinion mining) и т.д. Используемые при их решении средства лингвистической обработки, как правило, являются проблемно-зависимыми и, в силу этого, сильно ограниченными по своей функциональности, что тормозит их эффективное развитие и использование.

Диссертационное исследование «Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора» [1] посвящено актуальной тематике создания лингвистических технологий обработки естественно-языковых текстов на основе построения универсального лингвистического процессора. Актуальность темы подтверждается ее соответствием важнейшим государственным направлениям развития науки и техники, в соответствии с перечнем приоритетных направлений научных исследований Республики Беларусь.

Целью диссертационных исследований является разработка эффективных методов, алгоритмов и программных средств так называемого базового лингвистического анализа текстовых документов в многоязычной информационной среде. При достижении поставленной цели были решены несколько задач, основными из которых можно считать следующие.

Во-первых, была сформулирована концепция базового лингвистического процессора (БЛП), и выделены его унифицированные признаки. Во-вторых, определены характеристики лингвистической базы знаний БЛП и создан универсальный язык анализа текстов, что позволило построить алгоритмы лексического, синтаксического и семантического видов анализа текстов (и их композиции). Практическим результатом применения теоретических разработок явилось алгоритмическое, технологическое и программное обеспечение промышленного многоязычного БЛП.

В исследовании рассмотрен круг задач, связанных с разработкой лингвистической базы знаний (ЛБЗ) БЛП, определен ее состав и количественные характеристики: лексико-грамматический, синтаксический,

семантический классификаторы свойств естественного языка, базовый корпус текстов (БКТ), базовый словарь (БС) и базу данных лингвистических правил анализа текста.

Разработана принципиальная схема и соответствующая ей технология автоматизации разработки тегированного БКТ, основанная на ручном аннотировании лексико-грамматическими классами слов его относительно небольшой по объему части и последующем использовании получаемых при этом статистических данных с целью предварительного автоматического понижения степени лексико-грамматической многозначности слов из планируемых к аннотированию текстов создаваемого БКТ.

Учитывая высокую трудоемкость создания базовых словарей, объем которых составляет миллионы словоформ, предложено эффективное по скорости и точности решение задачи автоматизации их разработки на этапе построения полных парадигм для добавляемых в БС слов. Оно основано на создаваемой автоматически базе правил морфологического преобразования слов. Дано обоснование соответствующего алгоритма, определены минимально необходимые объемы словарей для различных естественных языков, гарантирующие оптимальное построение указанных баз правил.

Разработан язык расширенных регулярных выражений WRE для формального описания лингвистических правил, максимально соотнесенный с требованием доступности при использовании экспертами, возможностью обобщения разрабатываемых правил и построения эффективного алгоритмического обеспечения БЛП. Этот язык является обобщением языка традиционных регулярных выражений, в нем вместо терминальных символов используются предикаты и введены операции отрицания, вычитания и пересечения. Показано, что базовой процедурой алгоритмического обеспечения БЛП является процедура сопоставления цепочки знаков входного текста, с множеством представленных на языке WRE правил. Для ее реализации построен алгоритм сопоставления, основанный на теории конечных детерминированных автоматов, что обеспечило эффективное по скорости и точности решение задачи лингвистического анализа текстов.

Список литературы

1. Чеусов, А.В. Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора. – Автореф. канд. дисс. Белорусский гос. ун-т, Минск, 2013, - 16 с.

2. Чеусов, А.В. Принципиальная схема алгоритма сопоставления текстового входа для задачи его автоматического лингвистического анализа / А.В. Чеусов // Информатизация образования. – 2012. – № 1. – С.74-85.