УДК 001.103:002

ПРАКТИЧЕСКИЕ АСПЕКТЫ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ

И. Е. Воронина

Воронежский государственный университет

Поступила в редакцию 16.10.2012 г.

Аннотация. Обсуждается применение результатов лингвистических исследований для создания специализированного языка моделирования для обучения решению юридических задач.
Ключевые слова: компьютерная лингвистика, обработка естественного языка выделение ключевых слов, язык моделирования, обучение.
Annotation. Application of results of linguistic researches for creation of specialized language of modeling for training to the solution of legal tasks is discussed.
Keywords: computer linguistics, key words selection, modeling language, machine learning.

ВВЕДЕНИЕ

Главная цель всех проводимых лингвистических исследований – разобраться в структуре языка. Все языковые уровни характеризуются наличием базовых элементов. Изучение языка может идти с двух позиций – анализа и синтеза: выявленные правила синтеза могут способствовать проведению анализа и наоборот. Для исследования и максимальной формализации каждой языковой подсистемы создаются программные средства, реализующие процесс изучения путем выявления и проверки правил анализа и синтеза. Фиксация правил анализа и синтеза приводит к созданию анализаторов и синтезаторов каждого уровня иерархии (рис.1).

Традиционно, любая формализация подразумевает наличие совокупности правил, позволяющих строить описание объекта на декларативном или функциональном уровне. По сути дела, эти правила позволяют ответить на вопрос «как можно» (построить, описать, сделать и т. д.). В [1]предлагался подход к формализации, основанный на системе правил «как нельзя». Правила вида «как нельзя» разбиваются на группы. Каждая группа правил определяет фильтр. Каждый фильтр – это подсистема запретов на наличие семантического отношения, связывающего структурные единицы. Вся проблема в том, что ни один из фильтров нельзя определить однозначно и сразу. Именно по этой причине весь разрабатываемый инструментарий должен быть ориентирован на использование опыта и интуиции исследователя, подкрепляемых использованием математических оценок для принятия решения [2–4] (рис. 2).

Предлагалось формировать правила в виде запретов на сочетаемость для каждого языкового уровня [1–4].

Представленный подход лег в основу создания инструментальных исследовательских средств моделирования русского словообразования, исследования сочетаемости слов в предложениях. Кроме того, были реализованы некоторые сопутствующие лингвистические задачи:

– Программные средства для снятия неоднозначности слов в тексте.

– Когнитивно-графическая модель лексико-семантической системы.

– Программа выделения тематически маркированной лексики.

– Система выделения ключевых слов в текстах на естественном языке.

Прикладные лингвистические задачи отличает их заказной характер. В большинстве своем они представляют собой тот или иной социальный заказ. Их реализация протекает в диалоге «заказчик–разработчик». Еще одной особенностью прикладных задач является их проверяемость, при этом проверяемость повторная, неоднократная и каждый раз на новом материале.

В условиях «сверхзадачи» по реализации полной цепочки исследований, отдельные подзадачи могут сыграть важную практическую роль. К таким может быть отнесена задача выделения ключевых слов.

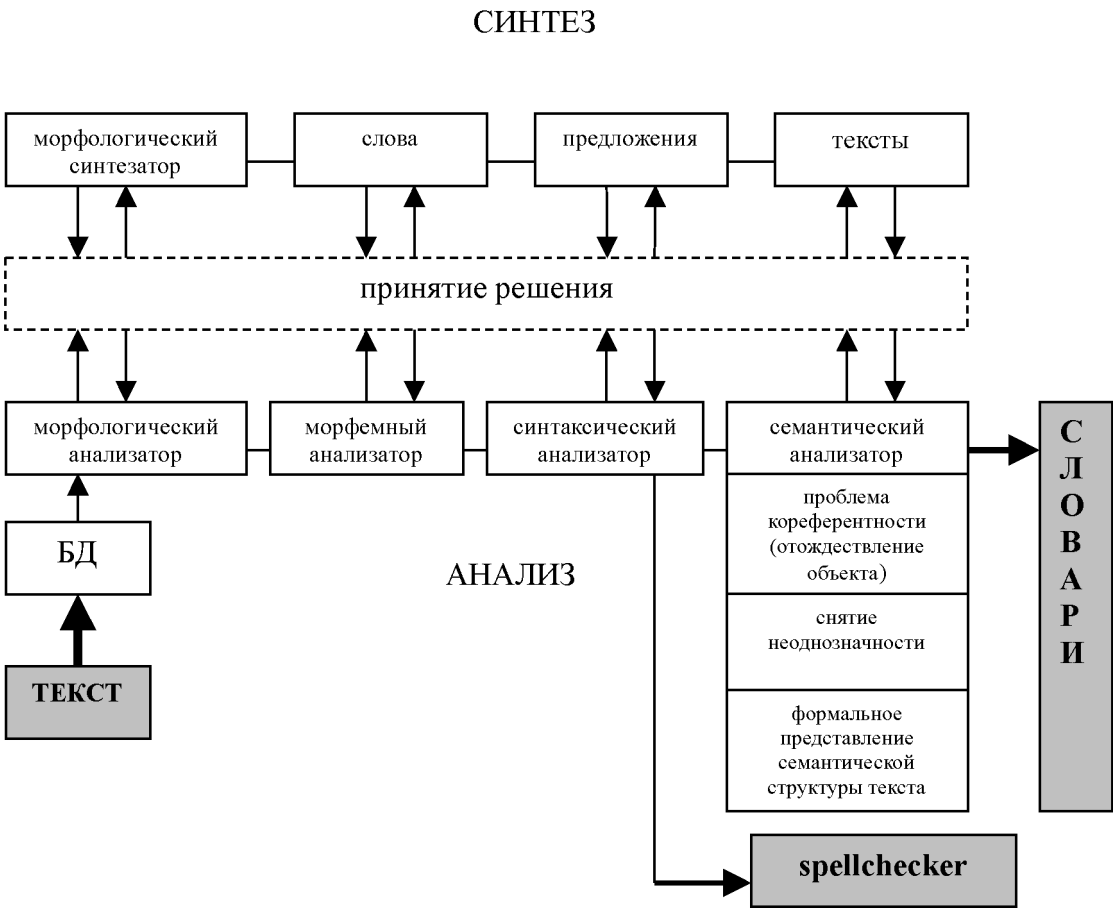

Рис. 1. Схема проведения исследований (Научно-методический центр компьютерной лингвистики ВГУ)

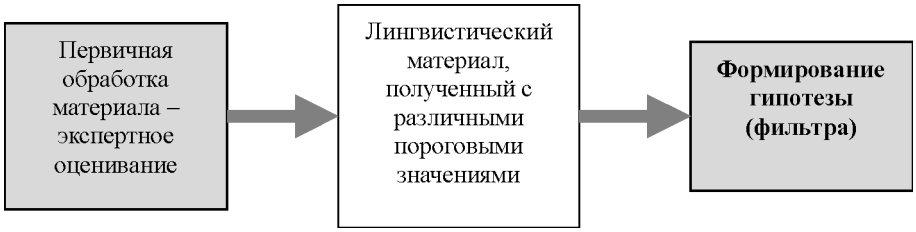

Рис. 2. Формирование гипотезы

ЛИНГВИСТИЧЕСКИЕ ПРОБЛЕМЫ В ЗАДАЧАХ ОБУЧЕНИЯ РЕШЕНИЮ ЮРИДИЧЕСКИХ ЗАДАЧ

Актуальность задачи автоматизации выделения ключевых понятий и словосочетаний в текстовых источниках неоднократно обсуждалась [5–7].

В контексте задач обучения – это использование ключевых слов и выражений при формировании макроструктуры текста, а затем и построения поля знаний – условного, неформального описания основных понятий и взаимосвязей между понятиями предметной области. Поле знаний – основа для создания формализованного представления знаний, позволяющего верифицировать ответ обучаемого, отслеживать ход его рассуждений.

А. А. Кретовым [8] был предложен метод формального выделения тематически маркированной лексики статистическим посредством «взвешивания» слов по функциональным параметрам, который был проверен на списке ключевых элементов, полученным при ручной обработке и предложен поход к выделению

ключевых слов с использованием алгоритма Гинзбурга и определения силы связи между словами [6].

В рамках исследования проблемы разработки программных средств обучения решению юридических задач приходится иметь дело с анализом текста развёрнутого ответа обучаемого, выявлением ключевых слов, рассмотрением и детальным поэтапным сравнением с верным ответом. Поскольку речь идет о неструктурированной информации, с одной стороны, могут быть два подхода к решению данного рода задач. Первый подход подразумевает обработку естественного языка со всеми вытекающими отсюда проблемами. Вторым вариантом может быть создание и использование специализированного языка моделирования. Но с другой стороны, оба подхода тесно связаны между собой.

Попытка сформировать модель предметной области (Уголовное право) для учебных задач квалификации преступлений с использованием методологии проведения исследований лингвистических процессов, выявить сопутствующие проблемы и предложить возможные способы их решения, явно обозначила ряд вопросов.

Сначала была разработана базовая юридическая онтология для российской системы права, отрасли «Уголовное право» и исследована возможность использования онтологического моделирования для формализации принятия решений в уголовном праве [9]. На этом этапе выявлены проблемы использования стандартных средств онтологического моделирования. Оказалось, что онтологии, дающие возможность ввести термины, типы, и соотношения (аксиомы) для описания фрагментов знаний, не решают проблему достаточной степени формализации, поскольку существует ряд ограничений на пути к полноценному представлению знаний. Эти ограничения оказались непреодолимыми, по крайней мере, в рамках стандартных средств разработки онтологий, поскольку включают в себя огромное количество неточностей (недостаточная определенность ключевых понятий, существенно затруднённое установление смыслового соответствия между понятиями ввиду использования для их описания профессионального жаргона наряду со специальной терминологией и т.д.).

Язык моделирования мог бы стать как языком описания объектов и явлений, так и основой теории соглашений об однозначном понимании вещей и явлений. Он основа интеллектуального интерфейса при формировании учебных заданий и ввода развернутого ответа обучаемого. Но структура такого языка настолько неочевидна, что одним из подходов к его «выявлению» может быть сравнение макроструктур текстов заданий и ответов, который в первом приближении должен происходить «вручную», а потом, с использованием расширенного количества источников – в автоматизированном режиме с применением вышеупомянутых алгоритмов выделения ключевых слов и исследования сочетаемости.

Обсуждаемая задача настолько сложна и неочевида, что трудно прогнозировать степень глубины и адекватности получаемых поэтапно результатов. Но такие исследования ведутся, и если даже задача будет решена в сильно усеченном варианте, это может быть полезным вкладом в создание обучающих программ.

СПИСОК ЛИТЕРАТУРЫ

1. *Воронина И. Е.* Оценки сочетаемости структурных единиц в задачах формализации естественного языка / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2006. – № 1. – С. 51–57.

2. *Воронина И. Е.* Задачи словообразования как составная часть проведения исследований в области естественно-языкового общения / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2006. – № 2. – С. 135–141.

3. *Воронина И. Е.* Программные средства моделирования словообразования / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2007. – № 1. – С. 75–80.

4. *Воронина И. Е.* Задачи словообразования как составная часть проведения исследований в области естественно-языкового общения / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2006. – № 2. – С. 135–141.

5. *Титова О. С.* Программные средства выявления семантического поля слов/ И. Е. Воронина, А. А. Кретов, О.С. Титова // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2008. – № 2. – С. 111–122.

6. *Воронина И. Е.* Функциональный подход к выделению ключевых слов: методика и реализация / И. Е. Воронина, А.А. Кретов, И.В. Попова, Л.В. Дудкина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2009. – № 1. – С. 68–72.

7. *Воронина И. Е.* Алгоритмы определения семантической близости ключевых слов по их окру-

жению в тексте / И. В. Попова, И. Е. Воронина, А. А. Кретов // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2010. – № 1. – С. 148–153.

8. *Кретов А. А.* Метод формального выделения тематически нейтральной лексики (на примере старославянских текстов) // Вестник Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2007. – № 1. С. 81–90.

9. *Воронина И. Е.* Создание базовой онтологии для Российской системы права на основе онтологии LKIF_CORE / И. Е. Воронина, Е. А. Пигалкова // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2010. – № 1. – С. 154–159.

Воронина Ирина Евгеньевна – кандидат технических наук, доцент кафедры программного обеспечения и администрирования информационных систем факультета ПММ, Воронежский государственный университет. E-mail: irina.voronina@gmail.com

Voronina Irina Ye. – Associated Professor of Software & Information System Administering Chair, Department of Applied Mathematics, Computer Science & Mechanics, Voronezh State University. E-mail: irina.voronina@gmail.com