Классификация ресурсов знаний в системе извлечения информации из текста

Сулейманова Е. А.
yes@helen.botik.ru
Переславль-Залесский, ИПС РАН

Ядром интеллектуальной системы анализа текстов является сплав знаний различной природы. Соотношение между компонентами системы, которые ведают разными знаниями, зачастую далеко не очевидно [3, 4].

В статье описан подход к систематизации знаний системы интеллектуального анализа текста в контексте задачи извлечения информации [2].

Основание классификации

Знания систематизируются по трем измерениям, которым поставлены в соответствие бинарные дифференциальные признаки со значениями: «предметные» — «лингвистические», «о классах» — «об индивидах», «априорные» — «из текстов».

Признаки делят пространство знаний на 8 секторов, каждому из которых соответствует свой набор значений признаков, Рис. 1. Рассмотрим, что представляют собой эти сектора (нумерация секторов произвольная), и какой компонент ресурса знаний соответствует каждому из них.

Типы знаний и компоненты системы

1. *«предметные», «о классах», «априорные»*
   Этому набору признаков отвечает *онтология* — общие знания системы об устройстве мира и предметной области в терминах концептов (классов сущностей) и их свойств. Концепты, атрибуты, признаки, отношения для удобства будем называть *элементами онтологии*. Элементы онтологии организованы в иерархические структуры.
2. *«предметные», «об индивидах», «априорные»*
   Наряду с общими знаниями о концептах интеллектуальная система должна располагать сведениями и о свойствах некоторых конкретных индивидов — экземпляров концептов. Эти сведения содержатся в *базе априорных фактов*.
3. *«предметные», «об индивидах», «из текстов»*
   Это целевые знания — знания о свойствах конкретных объектов, извлекаемые системой из текстов. Из них формируется *база текстовых фактов*[1].

[1] В соответствии с предложенным ранее [1] разделением подходов к извлечению информации на извлечение информации в «слабом» и «сильном» смысле, упомянутые здесь текстовые факты — это факты в «сильном» смысле, то есть результат переработки первичных текстовых фактов (извлеченных в «слабом» смысле) в знания.

4. *«предметные», «о классах», «из текстов»*
   В контексте перспективной задачи автоматизации пополнения онтологии на основе текстов в этот сектор должны попасть обнаруженные системой в текстах *знания о ранее не известных* (не отраженных в онтологии) *концептах и/или новых свойствах известных концептов.*
5. *«лингвистические», «о классах», «априорные»*
   Сюда можно отнести все лексикографические источники (за исключением упомянутых в последующих пунктах) и лингвистические модели. Отдельно стоит сказать о *словаре базовой предметной лексики системы.* Он организован как дескрипторный словарь: дескриптор представляет множество синонимических выражений. В отличие от тезауруса, дескрипторы в словаре не связаны друг с другом никакими парадигматическими отношениями, а содержат лишь ссылки на элементы онтологии. Таким образом, лингвистические знания отделены от экстралингвистических и отпадает необходимость в тезаурусе как некоторой переходной форме.
6. *«лингвистические», «об индивидах», «априорные»*
   В системе это *словарь собственных имен.* Его словарные входы тоже содержат ссылки на предметные знания, но это ссылки на конкретные объекты (экземпляры концептов) из базы априорных фактов. Кроме того, словарным входам приписаны довольно общие категории типа «ФИО», «название организации» (такие категориальные метки удобно использовать на этапе извлечения первичных текстовых фактов).
7. *«лингвистические», «об индивидах», «из текстов»*
   В процессе анализа текста и извлечения фактов строится динамический *словарь новых собственных имен* — в него включаются имена обнаруженных в тексте объектов. После проверки динамический словарь может служить источником пополнения словаря собственных имен системы.
8. *«лингвистические», «о классах», «из текстов»*
   Последний сектор возвращает нас к теме автоматизированного пополнения онтологии: здесь речь идет о соответствующем *пополнении словаря базовой предметной лексики системы.*

Заключение

(Пополняемая) онтология и базы априорных и текстовых фактов (верхние сектора) образуют *базу предметных знаний системы.* Нижние сектора — это база *лингвистических знаний.* Отметим, что границу между знаниями о классах и экземплярах (передняя и задняя половины «куба знаний») можно считать незыблемой, тогда как грань между

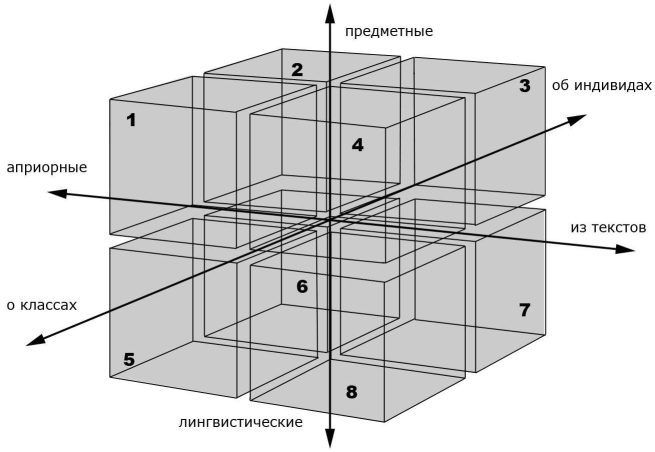

Рис. 1. Куб знаний.

знаниями априорными и извлеченными из текстов (левая и правая половины) периодически стирается.

Работа выполнена при поддержке РФФИ, проект № 05-01-00442а.

Литература

[1] *Куршев Е. П., Сулейманова Е. А.* Представление предметных знаний в системах интеллектуального анализа текста // Междунар. конф. «Программные системы: теория и приложения», ИПС РАН, Переславль-Залесский, октябрь 2006 г. — Т. 1 — М.: Физматлит, 2006. — С. 379–390.

[2] *Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В.* Исследование методов извлечения информации из текстов с использованием автоматического обучения и реализация исследовательского прототипа системы извлечения информации // ММРО-13 (наст. сб.) — 2007. — С. **??**–**??**.

[3] *Нариньяни А. С.* Кентавр по имени ТЕОН: тезаурус+онтология // Междунар. семинар Диалог'2001 по компьютерной лингвистике и ее приложениям. — Т. 1. — Аксаково, 2001. — С. 184–188.

[4] *Нариньяни А. С.* ТЕОН-2: от тезауруса к онтологии и обратно // Междунар. семинар Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». — Т. 1. — М.: Наука, 2002. — С. 307–313.