ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ ПОНЯТИЙ ИЗ ТЕКСТОВЫХ ДОКУМЕНТОВ ПРИ ГЕНЕРАЦИИ ПРЕСС-ПОРТРЕТА

Гостева Е. С.
Национальный технический университет
«Харьковский политехнический институт»
г. Харьков, ул. Пушкинская, 79/2, тел.707-63-60
e-mail: lisa.gosteva@ gmail.com

Одним из вариантов решения проблемы идентификации фактов в текстах и извлечения их характеристик является использование набора образцов. Образцами могут служить возможные лингвистические варианты фактов, которые можно поместить в интегрированный ресурс знаний (РЗ), объединяющий базу предметных знаний и словарь. Такой подход позволяет представить найденные ключевые понятия, представленные событиями и отношениями, в виде структур, которые в том числе можно хранить в базах данных [1-4].

Лингвистическая составляющая ресурса знаний — словарь. Словарь связан с базой предметных знаний посредством ссылок от дескрипторов к элементам знаний: дескрипторы словаря базовой лексики ссылаются на концепты, а дескрипторы словаря собственных имен — на априори известные экземпляры концептов из базы фактов.

Словарь базовой лексики и словарь собственных имен имеют схожее устройство — это дескрипторные словари (дескриптор представляет множество синонимичных выражений). В отличие от тезауруса, дескрипторы в словаре базовой предметной лексики не связаны друг с другом никакими парадигматическими отношениями (роль последних выполняют отношения между соответствующими элементами базы предметных знаний). В словаре собственных имен словарным входам приписаны довольно общие категории типа «имя лица», «название организации» (такие категориальные метки удобно использовать на этапе извлечения первичных текстовых фактов).

Сложность извлечения фактов с помощью образцов связана с тем, что на практике их нельзя представить в виде простой последовательности слов. Поэтому для идентификации различных уровней компонентов и отношений требуется предварительная обработка естественного языка на разных уровнях: первичная фильтрация документа; лингвистическая обработка: графематика, морфология, синтаксис; выделение простейших семантических структур; собственно извлечение информации и объединение построенных структур или интеграция фактов [2].

На стадии интеграции найденные в документах факты, исследуются и комбинируются. Это выполняется с учетом отношений, которые определяются местоимениями или описанием одинаковых событий. Также на этой стадии делаются выводы из ранее установленных фактов.

Компонент генерации фактов решает две основные задачи: генерацию (порождение) смысла будущей единицы пресс-портрета и лингвистический синтез самого высказывания по порожденному смыслу [3, с. 49-51]. Первый этап генерации фактов включает: определение информации, которая будет формировать пресс-портрет, построение семантической сети (графа), определение последовательности выдаваемой пользователю информации в соответствии с порядком фраз в выходном тексте и определение лексем, которые будут замещать позиции семантической сети конечного текста.

Второй этап процесса генерации конечного текста пресс-портрета связан с построением фраз на естественном языке. Для этого необходимо найти решение для следующих задач: построение синтаксической структуры будущей фразы; определение морфологической информации для входящих в составные части фразы слов; морфологический синтез всех словоформ фразы на естественном языке. Данная информация синтезируется из результатов лексико-синтаксической обработки текста.

Объединение лингвистических и предметных знаний в одном ресурсе, во-первых, облегчает первичное наполнение и последующую поддержку, а во-вторых, дает возможность использовать предметные знания уже на этапе первичной обработки текста правилами извлечения информации. Такой подход позволяет разработать специальный язык запросов к РЗ, при этом правила могут не ограничиваться словарной информацией, а обращаться к семантической сети (или онтологии) и базе фактов для проверки различных условий, требующих навигации по отношениям.

Список литературы

1. *Александровский Д.А., Кормалев Д.А., Куршев Е.П., Сулейманова Е.А, Трофимов Е.В.* Реализация ресурса знаний в системе извлечения информации из текста.// Сборник докладов. — М.: МАКС Пресс, 2007.

2. *Барсегян А. А., Куприянов М С., Степаненко В.В.*Технологии анализа данных: DaraMiningVisualMiningTextMining, OLAP. – СПб.: БХВ-Петербург, 2007. —384 с.

3. *Зубов А.В., Зубова И.И.* Основы искусственного интеллекта для лингвистов: Учеб.пособие. – М.: Университетская книга; Логос, 2007 – 320 с.

4. *Селезнев К.* Обработка текстов на естественном языке.//«Открытые системы», № 12, 2003.