УДК 004.89:004.93

Ф

СТРУКТУРА СИСТЕМЫ ИЗВЛЕЧЕНИЯ ИМЕННОЙ ГРУППЫ И АССОЦИИРОВАННЫХ С НЕЙ ДАННЫХ

Звенигородский Александр Сергеевич, Чернышова Валерия Николаевна[1]

Современные достижения в области информационных технологий позволили за короткий промежуток времени скопить в хранилищах данных различных организаций большие объемы информации, которая содержит скрытые пласты информации в виде знаний. В качестве такой информации может выступать описание какого-либо объекта, события, действующих лиц, локализации в пространстве и времени и т.д. [1]. Значительная часть информации в этих источниках представлена в виде естественных текстов, процесс аналитической обработки которых требует создания моделей, методик и систем интеллектуального анализа информации. В системах извлечения знаний из текстов также решается задача извлечения имен собственных.

Лингвистическая особенность и сложность выделения имен собственных субъекта, в частности, русских фамилий заключается в следующем[2]:

– фамилия является именем существительным, склоняется по падежам;
– в русском языке одно и то же слово может представлять собой имя субъекта (человека), наименование географического объекта или быть объектом лингвистики – имя прилагательное или существительное, которое не имеет отношения к субъекту.

Фамилия это один из атрибутов единственного в своем роде человека (субъекта). Конкретный субъект объективно может быть только один, но иметь некоторые значения атрибутов, совпадающие по значению с атрибутами других людей, к тому же значения атрибутов могут меняться со временем.

Для определения, обозначает ли лексема в тексте фамилию конкретного человека, необходимы не только морфологические знания, но и дополнительные знания, которые позволили бы повысить достоверность соотнесения лексемы с фамилией конкретного человека.

Дополнительные знания можно извлечь из перечня атрибутов, соотносимых с человеком. Например, перечень атрибутов, соотносимых с названием города, не совпадают с атрибутами человека или неодушевленного предмета.

Информация о части этих атрибутов и их значении может находиться в тексте[3,4], о части атрибутов можно делать предположение, например, по тематике текста, предыдущего информационного и смыслового содержания текста.

При разработке структуры системы мы вводим ограничение: анализ текста проводится с предположением, что назначение текста – передача информации о состоянии некоторой проблемной области.

Объекты ПрО определяются через атрибуты и значения этих атрибутов.

Процесс извлечения ИГС заключается в том, что требуется найти в тексте лексемы из лингвистического множества ИГС, семантически связанные с априорно заданным множеством атрибутов субъекта, определить значения этих атрибутов, по данным,

[1] Институт информатики и искусственного интеллекта ГВУЗ «Донецкий национальный технический университет», г. Донецк, Украина, zas@suiai.edu.ua, valeriyach@mail.ru

содержащимся в тексте, и оценить интегральный коэффициент достоверности.

На основании вышесказанного разработана структура системы представленная на рис. 1.

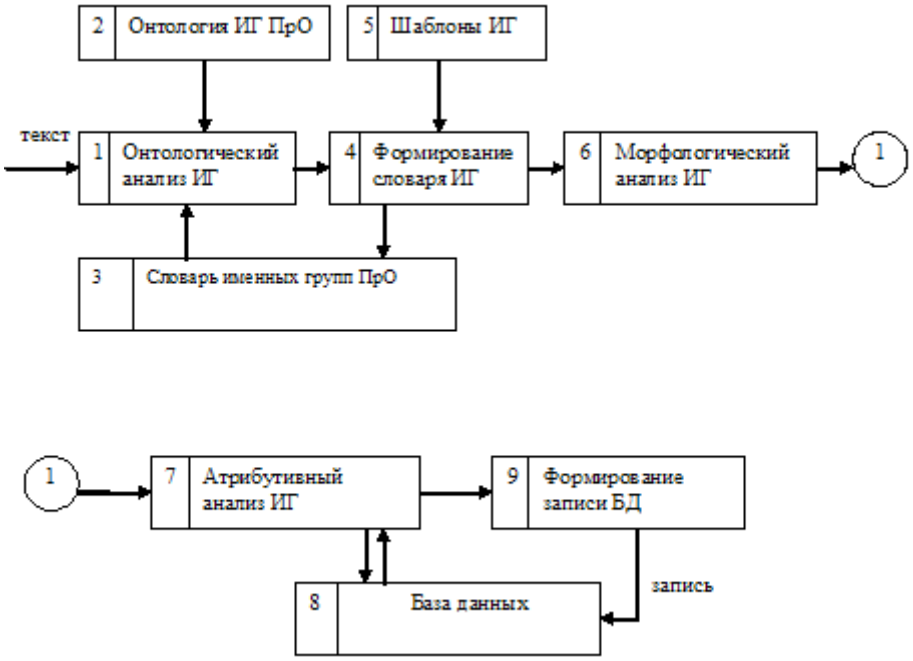

Рисунок 1. Структура системы извлечения именной группы и ассоциированных с ней данных

Блоки, представленные на рис. 1, выполняют следующие функции:

1. Онтологический анализ ИГ.

Онтология ИГ – это структурная спецификация ИГ. В основе онтологического анализа ИГ лежит глубокий структурный анализ предметной области.

2. Формирование словаря ИГ – блок выполняет обновление словаря именных групп на основании шаблонов ИГ.
3. Шаблоны ИГ - логико-морфологическая модель ИГ.
4. Морфологический анализ ИГ – анализ слов в тексте.

Морфологический анализатор, осуществляющий морфологический анализ всех слов текста на основе логико-морфологической модели. Морфологический анализ, проводимый в пределах слова, не может обеспечить стопроцентного однозначного определения его морфологических характеристик. Поэтому в системе МА предусмотрено корректирование результатов с помощью анализа грамматического контекста, т.е. снятие грамматической омонимии существительных.

1. Прагматический анализ - анализ состояния ситуации в текущем тексте с учетом контекста на основе собственной базы знаний. Прагматический анализ ограничен условием, что текст несет только лишь информацию ИГ. Так же в данном блоке выполняется проверка уникальности извлекаемых данных, путем обращения к БД (выполняется поиск в БД по ИГ).
2. БД – база данных.
3. Формирование записи БД – блок, отвечающий за приведение полученных

результатов, к форме представления записи в базе данных. И сохранение результата в БД. Функционирование данного блока основано на атрибутивной модели человека. Согласно этой модели и выполняется формирование записи БД.

Такая структура позволяет учитывать атрибутивную модель «субъекта» при анализе данных, а также при анализе учитываются языковые особенности фамилий. За счет этого словарь ИГ может обновлятся в процессе работы системы.

Список источников

1. Дудецкий В.Н. «Организация данных в системе понимания текста на ЕЯ», 1999 г.
2. Берков В.П. Правила употребления: Русские имена, отчества и фамилии. Высшая школа – 2005
3. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2004. – Москва, Наука, 2004. – С. 282-285. (http://www.rco.ru/article.asp?ob_no=629)
4. Кузнецов И.П., Ефимов Д.А.. Особенности извлечения знаний из текстов семантико-ориенированным лингвистическим процессором Semantix. // ИПИ РАН - М.: Наука, 2007.