А.И.Панченко

Материалы диссертационного исследования «Метод извлечения семантических отношений из разнородных источников текстовой информации»: краткое описание[1]

Введение

Существует множество типов семантических отношений между словами – синонимы, меронимы, антонимы и т.п. В рамках данной работы под семантическими отношениями понимаются синонимы, гиперонимы, ко-гиперонимы (слова имеющие общий гипероним) и ассоциации. Подобные отношения успешно применяются в задачах автоматической обработки текста (АОТ), таких как, расширение поискового запроса, классификация текстовых документов, разрешение омонимии, генерация изображений по тексту или создание вопросно-ответных систем. Построение систем АОТ приобрело особенную актуальность и важность в последнее время, когда количество текстов, представленных в электронном виде, стремительно умножается с каждым годом.

Семантические отношения заданные между терминами являются полезным ресурсом при построении систем АОТ, в силу пробела между лексической формой текста и его смыслом. Часто одно понятие, такое как “компьютер”, задается в тексте множеством эквивалентных по смыслу терминов, таких как “ЭВМ”, “электронно-вычислительная машина” или “машина”. Ес-ли система АОТ опирается только на лексическое представление текстов и не учитывает семантические связи между словами это зачастую приводит к субоптимальной производительности приложения.

Семантические отношения позволяют более точно представить смысл текстов в системах АОТ. Семантические отношения фиксируются в различных ресурсах, к числу которых относятся, прежде всего, тезаурусы, онтологии, терминологические классификаторы и словари синонимов. Огромное количество сил исредств было потрачено на создание ресурсов, таких как WordNet, RussNet,РуТез, Cyc, MeSH, EuroVOC или AgroVOC. Несмотря на это, существующие базы лексико-семантических знаний часто недоступны или недостаточны для конкретного приложения, предметной области или языка. При этом5ручное создание требуемых ресурсов – крайне дорогостоящий и трудоемкий процесс. В связи с этим, актуальной задачей является разработка методов автоматического извлечения семантических отношений.

Ранее было предложено множество методов извлечения отношений та-ких как дистрибутивный анализ, латентно-семантический анализ или методы основанные на лексико-синтаксических шаблонах. Однако качество из-влеченных отношений с помощью существующих подходов, в терминах точности и полноты, существенно ниже чем качество отношений построенных вручную. В силу этого, актуальной является задача разработки новых методов извлечения семантических отношений, превосходящих по характеристикам существующие подходы.

Цель работы.

Целью данной работы является разработка метода извлечения семантических отношений (синонимов, гиперонимов, ко-гиперонимов и ассоциаций), превосходящего по характеристикам существующие подходы, результаты работы которого были бы применимы в системах автоматической обработки текстов.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Произвести анализ существующих методов извлечения семантических отношений из текстов и других источников текстовой информации(словарей, онтологий и т.п.);

[1] В данном материале приводятся результаты исследований, выполненных при поддержке гранта РГНФ №12-04-12039в

2. Разработать подход к решению задачи извлечения семантических от-ношений из разнородных источников текстовой информации;

3. Произвести экспериментальную проверку предложенного метода на основании:

(a) сравнения с семантическими отношениями определенными в словарях;

(b) использования извлеченных отношений в модуле расширения информационно-поискового запроса;

(c) использования извлеченных отношений в модуле классификации имен файлов;

(d) использования извлеченных отношений в для измерения степени подобия коротких текстов;

(e) использования извлеченных отношений при построении лексико-семантической поисковой системы.

Научная новизна.

Научная новизна работы состоит в следующем:

1. Впервые произведен аналитический и экспериментальный сравнительный анализ 16 ключевых базовых методов извлечения семантических отношений, использующих 4 основных источника текстовой информации: корпус естественно-языковых текстов, Веб корпус естественно-языковых текстов, определения слов из словарей или энциклопедий и семантические сети. Показано что различные методы предоставляют взаимодополняющую информацию.

2. Впервые были предложены комбинированные метрики семантической близости, основанные на всех 4 ключевых источниках информации. При этом были предложены оригинальные подходы к комбинированию разнородных метрик, основанные на логистической регрессии, машине опорных векторов и факторизации разряженного тензора.

3. Предложен метод извлечения семантических отношений основанный на разработанных комбинированных метриках семантической близости и алгоритме взаимного ближайшего соседа. Показано что метод значительно превосходит по большинству характеристик 16 базовых и 7 альтернативных комбинированных методов извлечения отношений. Эффективность предложенного метода подтверждается успешным применением его при расширении поискового запроса, классификации имен файлов, нахождении семантической близости коротких текстов и при построении лексико-семантической поисковой системы.