Применение ЛСА и ЛДА для выявления элементов внешнего вида человека в тексте на естественном языке

А. В. Долбин, В. Л. Розалиев, Ю. А. Орлова
(Волгоградский государственный технический университет)

Данная работа посвящена семантическому анализу текста, составленному на естественном языке, и распознаванию именованных сущностей. Основной целью исследования было сравнение латентного размещения Дирихле и латентно-семантического анализа для выявления элементов внешности человека в тексте. В качестве сравнительной характеристики методов была выбрана полнота поиска информации.

Ключевые слова: распознавание именованных сущностей, ЛСА, ЛДА, разрешение кореференции, распознавание внешности человека.

Введение

Распознавание именованных сущностей относится к категории задач информационного поиска. На данный момент существует большое число методов для извлечения знаний из текста. Несмотря на то, что подобная задача появилась относительно недавно, она уже является одним из самых важных направлений в сфере компьютерных технологий. Благодаря извлеченным знаниям возможно получение дополнительной информации об объекте исследования [1].

Эта статья представляет собой результат исследования точности методов латентно-семантического анализа (ЛСА) и латентного размещения Дирихле (ЛДА) для распознавания элементов внешнего вида человека из текста. Мы использовали данные методы для анализа текстов на русском языке, однако они могут быть легко перенесены на другие языки.

Информационная модель человека

Была разработана информационная модель представления внешнего вида человека. В качестве модели представления знаний был выбрал фрейм. Описание фрейма «Внешний вид», имеющего слоты рост, телосложение, голова, волосы, лицо, лоб, брови, ресницы, нос, губы, подбородок, зубы, шея, плечи, грудь, спина, ноги, руки было выполнено на языке FRL [4].

Латентно-семантический анализ и латентное размещение Дирихле

Латентно-семантический анализ — это метод обработки текстовой информации, который анализирует связь между заданной коллекцией терминов и документов. Главная цель данного метода — найти документы, векторное пространство которых максимально близко к векторному пространству поискового слова [3].

Латентное размещение Дирихле используется для автоматической идентификации одной или более тем, которые содержат документы. Данный метод не учитывает семантику предложения, а просто работает с «мешком слов». С одной стороны, латентное размещение Дирихле дает информацию о том, с какой вероятностью каждое ключевое слово может относиться к каждой из потенциальных тем. С другой стороны, на выходе также получаем вероятность того, насколько документ может относиться к одной из тем [3].

Поиск именованной сущности в тексте

Одним из способов поиска сущностей является использование контекстных правил. Основная идея данного метода заключается в том, что заранее составленный набор обучающей выборки обрабатывается специальным образом, в следствие чего и формируются контекстные правила. Обучающая выборка представляет из себя отдельные предложения, никак не связанные друг с другом, в которых все слова заменены на условные обозначения, а стоп-слова просто опущены [5].

Для личности в тексте нами были составлены следующие правила обработки обучающей выборки:

- Любое упоминание о человеке в тексте заменяется на специальное обозначение PERSON;

- Вместо всех остальных слов используются их части речи, которые определяются по корпусу или словарю;
- Если слово обучающей выборки потенциально может быть использовано для извлечения знаний о внешнем виде, то оно указывается в своей начальной форме;
- Любое ключевое может быть отмечено знаками '?', '+', '|'. Знак '?' означает, что данную позицию при поиске совпадения можно опустить. Знак '+' — данное слово может повторяться более 1 раза подряд. Знак '|' представляет из себя логическое «или»;
- После каждого предложения опционально могут быть указаны словоформы, употребление которых в предложении исключает возможность определить личность [5].

Для разрешения кореференции местоимений в третьей лице допускается применение метода опорных векторов. В качестве обучающих данных нужно вручную разметить текст, в котором анафора (ссылка на объект при помощи одного из значений синонимического ряда кореференции) и антецедент (определяющий объект, в данном случае личность, упоминается в тексте до анафоры) заменены специальными символами. На выходе метод опорных векторов выдает вероятность, с которой анафора соотносится с антецедентом [2].

Список параметров для метода опорных векторов, составленный для разрешения кореференции местоимений в третьем лице:

- Количество предложений, разделяющих анафору и антецедент;
- Стоит ли антецедент в именительном падеже;
- Расположение анафоры в предложении;
- Расположение антецедента в предложении;
- Количество местоимений и существительных, расположенных в предложениях с анафорой и антецедентом;
- Совпадает ли падеж анафоры и антецедента;
- Совпадает ли род анафоры и антецедента;
- Совпадает ли число анафоры и антецедента [2].

Сравнение методов семантического анализа

Была реализована программа на языке Python 3 с использованием библиотеки Pymorphy2 для выявления внешнего вида человека. Pymorphy2

использует корпус русского языка «OpenCorpora», который насчитывает около полутора миллионов словоупотреблений. Обучающие выборки для метода разрешения кореференции и метода контекстных правил состояли из 500 позиций. Для применения латентного размещения Дирихле и латентно-семантического анализа данные для обучения не требуются.

Полнота информационного поиска оценивалась как отношение число найденных элементов к общее число элементов внешнего вида человека в документе. Результаты проведенного эксперимента представлены в таблице 1. Тексты на русском языке, используемые в данном эксперименте, были взяты из следующих областей: художественная литература, блоги, юридические тексты.

Таблица 1. Результаты сравнения методов семантического анализа текста.

Кол-во документов	Число слов в документе	ЛСА без распознавания личности	ЛДА без распознавания личности	ЛСА с распознаванием личности	ЛДА с распознаванием личности
1	100	0.7	0.68	0.85	0.83
2	100	0,69	0,67	0,84	0,82
5	200	0,69	0,67	0,84	0,82
7	200	0,67	0,65	0,83	0,81
10	500	0,65	0,62	0,81	0,79
12	500	0,64	0,62	0,80	0,78
15	500	0,64	0,61	0,79	0,78

Таблица 1 показывает, что если поиск выполняется только на отрывках текста с упоминание человека, то точность распознавания значительно увеличивается. В свою очередь, латентно-семантический анализ показывает более высокую точность по сравнению с латентным размещением Дирихле, хотя прирост является незначительным.

Заключение

Было проведено исследование на предмет извлечения элементов внешнего вида человека из текста на естественном языке с использованием методов латентно-семантического анализа и латентного размещения Дирихле. В качестве критерия эффективности для сравнения указанных

методов была выбрана полнота информационного поиска. Оба метода показали хорошие результаты по результатам эксперимента. Однако стоит отметить, что ЛСА показал более высокий показатель полноты информационного поиска по сравнению с ЛДА. Но в свою очередь, предварительная обработка текста, которая заключается в поиске ссылок на упоминание личности в тексте, дает значительный прирост эффективности информационного поиска для обоих методов семантического анализа.

Список литературы

[1] Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце; пер. с англ. под ред. П. И. Браславского, Д. А. Клюшина, И. В. Сегаловича. — М.: Вильямс, 2011.

[2] Толпегин П. В. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения. [Эл. ресурс] // Режим доступа : http://www.dialog-21.ru/digests/dialog2006/materials/html/Tolpegin.html, свободный. — Загл. с экрана. (10.11.2016).

[3] Долбин А. В. Анализ текста с использованием математических методов для распознавания элементов внешнего вида человека / А. В. Долбин, Ю. А. Орлова, В. Л. Розалиев // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. — Волгоград, 2015. — № 14 (178). — С. 56–60.

[4] Автоматизация составления портретных изображений по естественно-языковому описанию / Ю. А. Орлова, А. В. Долбин, Е. В. Кипаева, В. Л. Розалиев // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. — Волгоград, 2015. — № 2 (157). — С. 71–76.

[5] Named entity recognition without gazetteers / A. Mikheev [and etc.] // 9th Conference on the European Chapter of the Association for Computational Linguistics. — Stroudsburg, PA, 1999. — P. 1–8.