OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822

СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТА ДЛЯ РАСПОЗНАВАНИЯ ЭЛЕМЕНТОВ ВНЕШНЕГО ВИДА ЧЕЛОВЕКА

Долбин А.В.[*], Розалиев В.Л.[*], Орлова Ю.А.[*], Заболеева-Зотова А.В.[**]

[*]*Волгоградский государственный технический университет, г. Волгоград, Россия*

sizeof.void34@gmail.com

vladimir.rozaliev@gmail.com

yulia.orlova@gmail.com

[**]*Российский фонд фундаментальных исследований, г. Москва, Россия*

zabzot@gmail.com

В работе приводится описание различных методов обработки теста на естественном языке. Большинство описанных методов относятся к методам, основанным на машинном обучении. Рассматривается метод распознавания именованных сущностей с разрешением кореференции местоимений в третьей лице. Затем приводится использованием латентного размещения Дирихле и латентно-семантического анализа.
Ключевые слова:метод опорных векторов; ЛСА; ЛДА; распознавание именованных сущностей.

Введение

Распознавание элементов внешнего вида человека относится к категории задач информационного поиска.В данной статье рассматривается задачи распознавания именованных сущностей и извлечения фактов и неструктурированного текста, составленного на естественном языке.

Задача семантической обработки текста появилась относительно недавно. Но несмотря на это уже является чрезвычайно актуальной.

Результаты семантического анализа текста могут быть использованы в чрезвычайно большом спектре научных областей. Несмотря на свою востребованность, семантический анализ текста является одной из сложнейших математических задач. Это заключение вытекает из следующих факторов:

- естественный язык не формализован и этим обуславливается целый ряд сложностей в понимании текста;
- возможно неоднозначное толкование один и тех же слов. Так как одно и тоже слово может менять свой смысл в зависимости от контекста, то и программа «понять», как интерпретировать данное слово;
- производительность алгоритмов анализа текста.

Главная цель статьи – извлечение из текстов фактов о внешнем виде человека. Однако используемые методы извлечения фактов из текста могут быть использованы и для других областей.

1. Поиск именованной сущности «человек» в тексте.

1.1. Описание метода поиска личности.

Извлечение объектов и фактов из текста является задачей NLP и непосредственно Textmining. В данном разделе описывается способ нахождения личностей в тексте на русском языке. Данный способ также можно адаптировать под большинство современных языков.

Основная идея для поиска сущностей в тексте – использование контекстных правил и регулярных выражений. Данный метод относится к категории обучения с учителем. Обучение с учителем – один из способов машинного обучения, в ходе которого система обучается с использованием заранее составленной выборки. На основе этой выборки требуется установить зависимость между данными и на выходе получить точный ответ.

В качестве обучающей выборки используется обычный текст, составленный на русском языке. Программе требуется считать входные данные и из каждого образца обучающих данных (абзац, предложение) составить регулярное выражение особого типа.

1.2. Применение контекстных правил для поиска личности.

Первое, что требуется сделать – заменить искомую сущность (в данном случае упоминание о личности) на специальный символ (например, {P}). В рамках решения задачи нахождения человека в тексте это единственная необходимая ручная операция.

Следующим этапом является применение над обучающей выборкой алгоритмов графематического анализа для разделения текста на отдельные предложения и слова [Солошенко и др., 2014]. Как только все слова разделены, то при помощи доступного словаря или корпуса русского языка нужно получить часть речи слова.

Основная часть формирования регулярных выражений – часть речи слова. К примеру, рассмотрим предложение:

«У {P}, сидящей напротив, очень выразительный взгляд».

Из данного предложения можно сформировать следующее правило:

«PREP? PERSON PRTF ADVB+ выразительный взгляд».

Знак вопроса означает, что в данном случае предлог можно опустить, т.к. он употребляется на первом месте в предложении. Слово PERSON – потенциальная личность в тексте. Знак «+» означает, что слово с данной частью речи употребляется один или более раз подряд. Ключевые слова, которые относятся к тематике поиска (в данном случае рассматривается внешний вид человека) заносятся в отдельный список и никак не интерпретируются. Дополнительно, ключевые слова могут сопровождаться в выражении логической операцией или «|» Под специальной группой символов и находится искомая сущность. Очевидным плюсом данного подхода является то, что он не зависит от каких-либо граммем искомой сущности. В приведенном выше примере под личностью может подразумеваться как и слово «девушка», так и чье-либо имя. [Mikheev, 1999]

Для решения задачи данным способом требуется решить две основные проблемы:

- необходим достаточно большой объем обучающих данных, иначе система не сможет составить достаточное количество регулярных выражений;
- обучающая выборка должна быть обработана вручную.

1.3. Использование метода опорных векторов для разрешения кореференции в третьем лице.

В рамках данной работы рассматривается только разрешение кореференции местоимений в третьем лице, т.к. это один из наиболее простых случаев. Кореференция – связь нескольких отсылок в тексте к одному объекту. Для разрешения кореференции применяется метод опорных векторов. Метод опорных векторов относится к методам обучения с учителем. Следует рассматривать задачу бинарной классификации, т.к. пространство можно разделить на 2 класса: «является кореференцией»\«не является кореференцией».

В методе опорных векторов необходимо выбрать прямую, максимально удаленную от группы точек. Расстояние от этой прямой до каждой точки – максимально. Если такая прямая существует, то ее называют гиперплоскостью. Опорные вектора – это точки, расстояние до которых от гиперплоскости.

$$\frac{1}{\| w \|} \quad (1)$$

Метод опорных векторов строит классифицирующую функцию:

$$F(x) = sign(< w, x > + b). \quad (2)$$

w – нормальный вектор к разделяющей гиперплоскости, b – вспомогательный параметр, треугольные скобки – скалярное произведение. Необходимо выбрать такое w и b, которые максимизируют расстояние до каждого класса. Таким образом, необходимо решить задачи оптимизации.

Для реализации SVM также требуется обучающая выборка, размеченная вручную. Для выборки необходимо специальными символами разметить антецедент и потенциальную анафору. А также, к какому из двух классов относится каждый обучающий набор данных.

Были выделены следующие параметры для метода опорных векторов:

- количество предложений, разделяющих анафору и антецедент;
- стоит ли антецедент в именительном падеже;
- расположение анафоры в предложении (ближе к началу или концу предложения);
- расположение антецедента в предложении (ближе к началу или концу предложения);
- количество существительных и местоимений, расположенных в предложениях;
- совпадает ли падеж анафоры и антецедента;
- совпадает ли род анафоры и антецедента;
- совпадает ли число анафоры и антецедента.[Толпегин, 2006]

2. Извлечение фактов о внешнем виде человека.

2.1. Латентно-семантический анализ

Латентно-семантический анализ – метод обработки текстовой информации, анализирующий взаимосвязь между коллекцией терминов и документов. Основная задача данного метода – нахождение документов, которые наиболее близки в векторном пространстве к поисковому слову. ЛСА применяется для индексирования текста на естественном языке. Особенность ЛСА – частичное снятие омонимии с индексируемых слов.[Орлова и др., 2015]

Алгоритм латентно-семантического анализа:

- составить частотную матрицу термины на документы;
- стоит ли антецедент в именительном падеже;
- над частотной матрицей применить метод оценки релевантности TF-IDF для получения более правдоподобных результатов; [Маннинг и др., 2011]
- использование сингулярного разложения над частотной матрицей на матрицы U, S, Vt;
- Сократить количество строк в матрице Vt до 2. Для матрицы Uсократить количество столбцов до 2;
- По матрицам Vt, U определить координаты ключевого параметра.

Входной текст для латентно-семантического анализа:
«У Ольги светлые волосы и голубые глаза (1). Ногти у нее покрашены красным лаком (2). Она носит туфли на высоком каблуке (3). В её сумке всегда найдется шоколадка (4). У нее есть любимый кот по кличке Порш (5). В понедельник утром ей снова на работу (6).». Ключевое слово для поиска – «волосы».

Таблица 1 – Результат работы латентно-семантического анализа

Номерпредложения	Координаты	Индекс
0	(-1; 0)	0.0
2	(0; 0)	1.0
3	(0; 0)	1.0
4	(0; 0)	1.0
5	(0; 0)	1.0
1	(0; -1)	1.4142

Согласно таблице 1, ЛСА точно нашел искомый документ, максимально релевантный к заданному запросу.

2.2. Латентное размещение Дирихле

Латентное размещение Дирихле – порождающая модель, позволяющая объяснять результаты обработки данных с помощью неявных групп. В ЛДА каждый документ рассматривается как набор различных тематик. Количество тематик является один из выходных параметров данного метода.

Входной текст для модели латентного размещения Дирихле в текущем примере тот же, что и при рассмотрении ЛСА. Набор ключевых терминов: «светлые, волосы, покрашены, красным, лаком, высоком, найдется, шоколадка, глаза, носит, голубые, кот, суббота, работа, дел, идти».

В итоге на выходе метод определяет четыре тематики по входному тексту:

- кот, шоколадка, лаком, красным;
- найдется, волосы, идти, дел;
- голубые, глаза, светлые, волосы;
- идти, носит, высоком, дел.

В рамках одного предложения или абзаца может быть упомянуто несколько тем. Но методы кластеризации документов по темам это не могут учесть. В связи с этим и применяется метода латентного размещения Дирихле. В отличии от обычной кластеризации, в методе для каждого заданного слова по распределению выбирается тема. Латентное размещение Дирихле относится к методам машинного обучения.

В результате обучения модели получаются векторы, отображающие как распределены в каждом документе заданные темы и векторы, отображающие, какие ключевые термины наиболее вероятны в той или иной теме. В итоге можно получить информацию о темах в рамках документа и о списке слов, характерных для данной темы.

Заключение

Таким образом, на данном этапе все описанные методы используется для общей цели – распознавания элементов внешнего вида человека по тексту на естественном языке. В качестве основного механизма будет использоваться латентно-семантический анализ, т.к. он обладает более точными результатами, по сравнению с ЛДА [Коляда и др., 2014]. Преимущество латентного размещения Дирихле заключается в том, что с его помощью можно определить неявные элементы именованной сущности. К примеру, для распознавания внешнего вида возможно нахождение фразеологизмов. Метод разрешения кореференции в дальнейшем планируется расширить и не ограничиться только местоимениями в третьей лице.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов №15-07-05440,15-07-07519, 15-37-70014,16-07-00453.

Библиографическийсписок

[Mikheev, 1999]Named Entity Recognition Without Gazetteers / Andrei Mikheev [and etc.]; - 9th Conference on the European Chapter of the Association for Computational Linguistics, Stroudsburg, PA, 1999, pp. 1-8.

[Толпегин, 2006]Толпегин, П. В. Алгоритм автоматизированного разрешения анафоры местоимений

третьего лица на основе методов машинного обучения. [Электронный ресурс] // Режим доступа : http://www.dialog-21.ru/digests/dialog2006/materials/html/Tolpegin.html, свободный. — Загл. с экрана. (18.06.2014).

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

[Маннинг и др., 2011]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. - Москва.: И. Д. Вильямс, 2011. — 528 с.

[Коляда и др., 2014]Коляда, А. С. ПрименениелатентногоразмещенияДирихледляанализапубликацийизнаукометрическихбазданных / А. С. Коляда, В. А. Яковенко, В. Д. Гогунский // Одесский национальный политехнический университет. – 2014. – Вып. 1. – с 186 – 191.

[Солошенкоидр., 2014] Thematic Clustering Methods Applied to News Texts Analysis / СолошенкоА.Н., ОрловаЮ.А., РозалиевВ.Л., ЗаболеевоЗотоваА.В. // Knowledge-Based Software Engineering : Proceedings of 11th Joint Conference, JCKBSE 2014 (Volgograd, Russia, September 17-20, 2014) / ed. by A. Kravets, M. Shcherbakov, M. Kultsova, Tadashi Iijima ; Volgograd State Technical University [etc.]. – [Б/м] : Springer International Publishing, 2014. – P. 294-310. – (Series: Communications in Computer and Information Science ; Vol. 466).

SEMANTICANALYSIS OF TEXT FORRECOGNITION OF THE ELEMENTS OF HUMAN APPEARANCE

Dolbin A.V. [*], Rozaliev V.L. [*], Orlova Y.A. [*],Zaboleeva-Zotova A.V. [**]

**Volgograd State Technical University, Volgograd, Russia*

sizeof.void34@gmail.com

vladimir.rozaliev@gmail.com

yulia.orlova@gmail.com

*** Russian Foundation for Basic Research, Moscow, Russian Federation*

zabzot@gmail.com

The paper describes the various methods of processing natural language. Most of these methods are based on machine learning. The method of recognition of named entities with coreference resolution of pronouns in the third person is described. Then, given the use of the latent Dirichlet allocation and latent semantic analysis.

Introduction

Recognition ofelements ofthe appearance ofa manfalls into the categoryof information search. This article discusses theproblem ofnamed entityrecognitionandfact extraction from unstructured textin a natural language.

The problem of semantic text processing is a relatively new. But in spite of this is an extremely urgent.

The results of the semantic analysis of text can be used in a very large range of scientific fields. Despite its relevance, semantic analysis of the text is one of the most complex mathematical problems. This conclusion stems from the following factors:

- natural language is not formalized, and this caused a number of difficulties in understanding the text;
- possibly ambiguous the same words. Since the same word can change its meaning depending on the context, and the program should know how to "understand" the word;
- performance issue.

The main purpose of the article - the extraction of facts about the appearance of a person. However, methods that were used forfacts extraction from the text can be used for any other theme.

Main Part

This article describes the mathematical methods for the identifying the elements of human appearance. Before processing natural language with these methods, every word must be reduced to the normal form. An alternative solution is the use of the Porter stemmer.

First, you must determine the persons for further processing. To this end, based on the training sample generated contextual rules are then applied to the processed text. This approach has a number of positive sides. The disadvantage of this method is that it is impossible to distinguish a person from any other entity.

Then you need to allow the coreference pronouns. This problem was solved by using the method of support vector machine. Support Vector Machines is classified as machine learning method. In feature extraction space is divided into 2 parts by hyper-plane. Then the model can conclude whether anaphora and antecedent are referenced to the same person.

The final step is the search for elements of the appearance of a man with the key words using the method of latent semantic analysis. After the LSA was used, the latent Dirichlet allocation should be usedto determine possible phraseology, that can describe some aspect of human appearance.

Conclusion

Thus, at this stage, all methods used for the common goal - recognition of elements of the appearance of the humanin natural language. As a main mechanism latent semantic analysis should be used, as it has more accurate results, compared with the LDA [Kolyada et al., 2014]. The advantage of placing the latent Dirichletallocation is that it can be used to determine the implicit elements of a named entity. For example, to detect the the possible presence of phraseology in human appearance. Coreference resolution method in the future is planned to expand and not be limited only in the third person pronouns.