OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

ВОЗМОЖНОСТЬ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ АДРЕСАТА НА ОСНОВЕ СЕМАНТИКО-СИНТАКСИЧЕСКИХ ОСОБЕННОСТЕЙ ТЕКСТА

Глазкова А.В.

Тюменский государственный университет,
г. Тюмень, Россия
anna_glazkova@yahoo.com

Решение проблем организации эффективного доступа к информации становится все более актуальным в связи с широким использованием интернет-технологий. Вследствие этого много внимания привлекает к себе задача автоматической классификации текстовых документов. В работе обсуждается возможность автоматической классификации текстов в зависимости от того, какой аудитории они адресованы.
Ключевые слова: классификация текстов; информационные ресурсы; информационный поиск; искусственный интеллект.

Введение

Объем информации в мире постоянно растет, и в связи с этим одним из ключевых направлений современной компьютерной науки является разработка методов систематизации имеющихся данных.

Необходимость интенсивного развития средств информационного поиска – в том числе для работы с поисковыми системами и электронными сообщениями – включает в себя также потребность в улучшении методов и алгоритмов автоматической классификации текстов, что является необходимым условием улучшения результативности обработки текстовой информации.

1. Краткий обзор текущего состояния технологий автоматической обработки текстов

В середине XX века в рамках дисциплины «искусственный интеллект» выделилось направление, связанное с обработкой естественного языка (Natural Language Processing). Трудность задач данного направления заключается в том, что естественный язык представляет собой сложную и многоуровневую систему, постоянно изменяющуюся в процессе жизнедеятельности человека. Многообразие естественных языков, их лексического, морфологического и синтаксического составов, также порождает сложность в создании и реализации методов обработки текстов на естественном языке.

В основе решения задач обработки текстов на естественном языке лежит использование моделей документов, отражающих словообразовательные, грамматические и смысловые особенности текстов [Рафанов, 2010]. Создание этих моделей опирается соответственно на методы морфемного, синтаксического и семантического анализа, методы статистического анализа и морфологические характеристики лексем.

Результаты автоматизации морфологического анализа текста в настоящее время применяются в текстовых редакторах, поисковых системах, модулях проверки правописания. Входными данными морфологического анализа является текстовый документ, выходом – список предложений, каждое из которых представляет собой список слов. Каждое слово, в свою очередь, – список лексем, то есть групп словоформ, соответствующих одной нормальной (словарной) форме слова и одной части речи, а также стоящих в одном числе. На данном этапе помимо грамматической информации для каждого слова запоминается, как правило, его смещение от начала текста и следующий за словом разделитель [Осипов, 2013].

Задачи анализа синтаксического и семантического анализа в полной мере не решены, несмотря на богатую теорию в области возможных путей их автоматизации. Широкое применение находят лишь методы, основанные на статистических характеристиках анализируемого текста (количественной оценке частоты встречаемости тех или иных слов и

словосочетаний). Основной задачей синтаксического анализа является установление различных зависимостей между лексемами, выявленными на этапе морфологического анализа, – в частности, выделение именных групп. Понятие именной группы – словосочетания, в которое входит корневое существительное – используется в задачах классификации текстов, а также при сравнении поисковых образов документа и запроса. Помимо установления зависимостей, синтаксический анализ обнаруживает однородные члены предложения. Таким образом, на этом этапе между лексемами могут быть установлены два вида связей: управление и однородность. Входными данными синтаксического анализа являются, как правило, предложения на выходе морфологического анализа, выходными – предложения в виде списков вариантов, каждый из которых может представлять собой список деревьев зависимостей [Осипов, 2013].

При этом системы, реализующие методы синтаксического анализа документов, не предусматривают в полной мере возможности настройки процесса обработки текста и средств пополнения правил грамматики использующегося естественного языка. Область семантического анализа документов нуждается в разработке методов, аналогичных по своей сути методам, применяемым человеком при анализе информации, то есть в создании компьютерных систем, основанных на базах знаний и процедурах логического вывода и принятия решений, которые могли бы эффективно использоваться для задач автоматической обработки текстов [Большакова, 2011] [Мокроусов, 2010]. Разработка семантических моделей необходима для решения различных задач информационного поиска, в том числе и задачи классификации документов.

К основным методам классификации текста на основе обучающей выборки относятся [Маннинг, 2011]:

- вероятностные методы (байесовские – например, наивный байесовский метод на основе мультиномиальной модели или модели Бернулли, методы, оценивающие распределения вероятностей – например, метод максимальной энтропии);
- нечисленные методы (деревья и правила принятия решений);
- методы на основе векторной модели (классификация Роккио и метод k ближайших соседей);
- геометрические методы (метод опорных векторов и его модификации) и другие.

2. Задача классификации текстовых документов

2.1. Постановка задачи классификации

Для представления задачи классификации в общем виде рассмотрим некоторое конечное множество категорий C и конечное множество документов D:

$$C = \{c_1, c_2, ..., c_{|C|}\}. \quad (1)$$

$$D = \{d_1, d_2, ..., d_{|D|}\}. \quad (2)$$

Целевая функция Φ, которая для каждой пары <документ, категория> определяет, соответствуют ли они друг другу, неизвестна:

$$\Phi : D \times C \to \{0,1\}. \quad (3)$$

Необходимо найти классификатор Φ', то есть функцию, максимально близкую к функции Φ [Sebastiani, 2002].

Таким образом, целью классификации текстовых документов является разделение документов из имеющегося множества по классам. Для этого используется обучающая выборка документов и алгоритм обучения, при помощи которого можно получить классификатор, или функцию классификации, отображающую документы в классы. В процессе выполнения классификации необходимо обеспечить достижение высокой точности не только на обучающей выборке, но и на новых данных.

2.2. Применение систем классификации текстов

Системы автоматической классификации текстов используются для решения многих прикладных задач информационного поиска, требующих сопоставления различных типов документов одной или нескольким выделенным категориям на основании содержания анализируемых текстов. К таким задачам относятся:

- поиск в электронных библиотеках и сети Интернет;
- фильтрация почтового спама;
- распределение текстов (например, новостных) по тематическим категориям;
- составление интернет-каталогов;
- подбор контекстной рекламы;
- автоматическое аннотирование и реферирование текстов;
- снятие неоднозначности при автоматическом переводе текстов;
- ограничение области поиска в поисковых системах;
- определение кодировки и языка текста;
- идентификация автора текста;
- жанровая классификация текстов и другие.

3. Возможность определения адресата текста

Одной из задач классификации документов является идентификация автора текста. Развитие вычислительной техники в настоящее время дает возможность не только для реализации таких

традиционных функций, как определение средней длины слов в тексте, подсчет числа слов в предложении, предложений в абзаце и так далее, но и для вычисления более сложных показателей – как на основе статистических данных, так и на основе семантико-синтаксических моделей лексико-грамматических структур документов.

В различных работах предпринимались попытки на основе анализа естественно-языкового текста сделать выводы о его языке, эпохе написания, литературном направлении – в случае, если текст художественный, типе (проза или поэзия), формате (роман, повесть, очерк, эссе), жанре и, наконец, авторе [Орлов, 2012]. В данной работе рассматривается возможность классификации текстов на естественном языке в зависимости от того, кому он адресован, – в частности, в зависимости от возраста аудитории. Решение этой задачи может быть использовано для усовершенствования механизмов информационного поиска, в частности поиска в сети Интернет и библиотеках электронных документов.

Процесс идентификации потенциального адресата текста подразумевает обращение к некому набору «эталонов» – базе знаний, отражающей характерных черты текстов, предназначенных для той или иной категории читателей. Для текста с неизвестной категорией будет требоваться определить его наиболее вероятный класс, то есть соотнести предложенный текст с одним из известных классов или с несколькими из них.

Часть признаков, лежащих в основе классификации, может быть получена, исходя из лексического (словарного) состава текста – подобной характеристикой может быть, например, отношение количества терминов к общему количеству слов. Остальные признаки должны определяться в зависимости от уровня синтаксической и семантической сложности документа. К синтаксическим особенностям, влияющим на классификацию, можно отнести:

- количество сложносочиненных и сложноподчиненных предложений;
- длина предложений;
- наличие обособлений, причастных и деепричастных оборотов;
- длина слов.

Также следует оценить сложность текста, подлежащего классификации, с помощью математических методов, основанных на понятии информационной энтропии [Кутузов, 2010]. Одной из классических работ, посвященных этой проблеме, является статья А. Н. Колмогорова, где описаны комбинаторный, вероятностный и алгоритмический подход к определению понятия «количество информации». На их основе выведена взаимосвязь между сложностью текста и его энтропией [Колмогоров, 1965]. С понятием сложности текста связано вычисление индекса удобочитаемости, определение степени неоднозначности. Таким образом, для решения задачи потребуется выявление не только лингвистических, но и логических отношений на языковых объектах, то есть семантических связей между ними.

Заключение

В настоящей статье рассмотрена возможность реализации классификации текстов на основе категорий их адресатов. Основанием для классификации могут послужить как данные о лексическом составе текста, так и информация о его грамматической и семантической сложности.

Совершенствование алгоритмов и методов автоматической обработки текстовых документов является актуальной задачей в связи с широким использованием средств обработки информации.

Библиографический список

[Sebastiani, 2002] Sebastiani, F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys. - 2002. - Vol. 34, No. 1.

[Jurafsky, 2002] Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing , Computational Linguistics, and Speech Recognition / D. Jurafsky, J. H. Martin. // New Jersey: Pearson, 2009. – p. 988.

[Yang, 1999] Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213-3702, USA, 1999. – p. 8.

[Большакова, 2011] Большакова, Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие/ Е. И. Большакова, Э. С. Клышинский, Д. В. Ландэ, А. А. Носков, О.В. Пескова, Е.В. Ягунова // М.: МИЭМ, 2011. – 272 с.

[Колмогоров, 1965] Колмогоров А.Н., Три подхода к определению понятия «количество информации» / А. Н. Колмогоров // Проблемы передачи информации, 1965, 1:1, С. 3-11.

[Кутузов, 2010] Кутузов, А. Б. Методики определения сложности текста в рамках переводческого анализа / А. Б. Кутузов // Вестник Нижегородского государственного лингвистического университета им. Н.А. Добролюбова. Вып.4. Лингвистика и межкультурная коммуникация, 2009.

[Маннинг, 2011] Маннинг К. Введение в информационный поиск. : Пер. с англ. /, К. Маннинг, П. Рагхаван, Х. Шютце // М.: ООО «И.Д. Вильямс», 2011. – 528 с.

[Мокроусов, 2010] Мокроусов, М.Н. Разработка и исследование методов и системы семантического анализа естественно-языковых текстов: автореф. дис. … канд. техн. наук: / М. Н. Мокроусов // Ижевск, 2010. – 24 с.

[Орлов, 2012] Орлов, Ю. Н. Методы статистического анализа литературных текстов / Ю. Н. Орлов, К. П. Осминин// М.: Книжный дом «ЛИБРОКОМ», 2012. – 312 с.

[Осипов, 2013] Осипов, Г. С. Лекции по искусственному интеллекту. Изд. 2-е, испр. и доп. / Г. С. Осипов // М.: Книжный дом «ЛИБРОКОМ», 2013. – 272 с.

[Рафанов, 2010] Рафанов, С. В. К проблеме классификации текстов в машинном переводе / С. В. Рафанов // Вестник Московского государственного областного университета. Серия: Лингвистика, 2010, № 3, С. 36-42.

[Семенов, 2002] Семенов, А.Л. Математика текстов / А. Л. Семенов // М.: МЦНМО, 2002. – 16 с.

[Электронный ресурс] Автоматическая обработка текстов [Электронный ресурс]. – Режим доступа: http://www.aot.ru/index.html.

[Электронный ресурс] Лившиц, Ю. Классификация текстов. Алгоритмы для Интернета / yury.name/internet [Электронный ресурс]. – Режим доступа: http://yury.name/internet/.

[Ягунова, 2012] Ягунова, Е. В. Вариативность структуры нарратива и разнообразие стратегий понимания / Е. В. Ягунова // Человек и язык, 2012, С. 65-83.

THE CAPABILITY OF AUTOMATIC TEXT ADDRESSEE RECOGNITION BASED ON SYNTACTIC AND SEMANTIC FEATURES

Glazkova A.V.

Tyumen State University, Tyumen, Russia

anna_glazkova@yahoo.com

The article deals with the task of natural language text procesing. The paper discusses the capability of automatic text categorization based on characteristics of the text addressee.

Introduction

Currently, it is necessary to handle a large amount of information. Therefore, one of the important directions of modern computer science is to develop methods to systematize the available information.

Consequently, there emerged a need to improve the methods and algorithms of information retrieval, including automatic text categorization.

Document categorization methods are used to solve a variety of word processing tasks: filtering and spam recognition, automatic annotation and summarization, news distribution, and other categories.

Main Part

Natural language processing is a branch of science related both to artificial intelligence and linguistics that became a separate discipline in the middle-of the XX century. Natural language is a complex and evolving system. Diversity of natural languages, their lexical, morphological and syntactic structures also creates difficulty in developing and implementing methods for processing natural language texts. The basis of solving natural language processing is methods of morphemic, syntactic and semantic analysis, statistical analysis and morphological characteristics of lexical items. Current document analysis systems and methods of analysis don't give the full opportunity to set up text processing tools and renew grammar rules. Therefore, it's necessary to continue improving of information retrieval systems; this also applies to systems of text categorization.

Text categorization is the task of assigning a Boolean value to each pair <document, category>, where D is a domain of documents and C is a set of predefined categories. More formally, the task is to approximate the unknown target function Φ' called the classifier by means of a function Φ: $\Phi : D \times C \to \{0,1\}$.

Document categorization is a problem in library science, information science and computer science. The main difficulty of text categorization is complexity and ambiguity of natural language. Automatic document classification techniques include Naive Bayes classifier, k-nearest neighbour algorithms, method of selected points, decision trees, decision rules and others. Classification techniques have been applied to:

- spam filtering, a process which tries to discern E-mail spam messages from legitimate emails;
- email routing, sending an email sent to a general address to a specific address or mailbox depending on topic;
- language identification, automatically determining the language of a text;
- genre classification, automatically determining the genre of a text;
- readability assessment, automatically determining the degree of readability of a text, either to find suitable materials for different age groups or reader types or as part of a larger text simplification system.

This paper deals with the classification of natural language text, depending on the addressee of the text- in particular, depending on the age of the audience. The solution to this problem can be used to improve information retrieval mechanisms. The process of identifying a potential addressee of the text means an appeal to a set of "standards" – the knowledgebase that reflects the characteristic features of texts intended for a particular category of readers. It is required to assign the text with an unknown category to the most probable class. Some features for text categorization may be derived from the dictionary of the text (the ratio of the number of terms to the total number of words, etc.). Other features should be determined depending on the syntactic and semantic complexity of the document. Syntactic features affecting the classification include:

- complexity of sentences;
- average length of words;
- length of sentences;
- presence of participial constructions and adverb phrases.

Complexity of the text can also be evaluated using mathematical methods based on the concept of information entropy. One of the classic works devoted to this problem is the article of Kolmogorov A. N., which describes a combinatorial, probabilistic and algorithmic approach to the definition of "quantities of information". In this paper is derived the relation between the complexity of the text and its entropy. The notion of complexity associated the text readability index calculation and determination of the degree of ambiguity.

Conclusion

In this paper we consider the capability of realization of the text classification based on categories of recipients. The basis for classification may be data about the lexical composition of text and information about its grammatical and semantic complexity.

Development of algorithms and methods for automatic processing of text documents is an important task of artificial intelligence due to the wide use of information processing facilities.