НЕКОТОРЫЕ АСПЕКТЫ ОСУЩЕСТВЛЕНИЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВЕ РАСПОЗНАВАНИЯ ИХ АДРЕСАТОВ

Глазкова А.В.
Тюменский государственный университет, г. Тюмень, ул. Семакова, д. 10, +79091826371, anya_kr@aol.com

1. Введение

В связи с активным развитием технологий разработки интеллектуальных систем в настоящее время увеличивается потребность в исследованиях, направленных на улучшение механизмов информационного поиска. Использование поисковых систем, электронных библиотек, спам-фильтров немыслимо без применения инструментов обработки текстовой информации.

Данная работа направлена на рассмотрение вопроса автоматического определения адресата текста – в частности, возможности классификации текстов в зависимости от того, какой возрастной аудитории они адресованы. Мы хотим обрисовать определенный минимум, необходимый для создания системы, реализующей заявленную функцию. В дальнейшем планируется определить состав набора признаков для классификации, сделать предложения по созданию базы знаний и обучающего корпуса, методам классификации и лингвистического анализа.

2. Задача классификации и ее формальная постановка

Задача классификации текстов заключается в определении принадлежности текста одному или нескольким классам. Для каждого документа-объекта при этом выделяются наборы признаков – слов и их взаимозависимых наборов. Для формирования наборов этих признаков для каждого документа используются лингвистические и статистические методы [1]. Численные значения, принимаемые объектами того или иного класса, вычисляются в процессе обучения классификатора. По завершении обучения принадлежность текста к классу определяется при помощи проведения анализа признаков текста с учетом полученных весовых значений.

Автоматическая классификация может применяться в таких областях информационного поиска:

- поиск в электронных библиотеках и сети Интернет;
- фильтрация почтового спама;
- составление интернет-каталогов;
- подбор контекстной рекламы;
- снятие неоднозначности при автоматическом переводе текстов и др.

Приведем формальную постановку задачи классификации. Пусть дано множество категорий и множество документов . Целевая функция , которая для каждой пары <документ, категория> определяет, соответствуют ли они друг другу, неизвестна. Задача состоит в построении классификатора , максимального близкого к функции [2]. Основными подходами к решению данной

задачи являются наивный байесовский подход, метод k ближайших соседей, построение деревьев решений, использование метода опорных векторов и создание нейронных сетей [3].

3. Выявление характеристик адресата текста

В рамках разработки методов и алгоритмов автоматической классификации текстов рассматриваются вопросы распределения текстов по жанрам, времени написания, автоматического распознавания автора и языка. Любой текст, как известно, явно или не явно предназначается конкретному читателю как в широком смысле – например, группе людей, говорящих на определенном языке, так и в узком – например, представителям одной возрастной категории [4]. При этом текст как бы включает в себя образ «своей» идеальной аудитории, аудитория – «своего» текста [5].

В рамках своей коммуникативной деятельности автор составляет текст, имея установку на максимально полное доведение до адресата своего замысла для того, чтобы адресат его (автора) понял. Речь должна быть ориентирована на слушателя, и естественным следствием такой установки является намерение автора использовать такие содержание и структуру прогнозируемого текста, а также такие средства языка для их выражения, которые в своей совокупности были бы доступны пониманию реципиента, которому адресован текст. В работе Каменской О.Л. [6] рассматривается понятие коммуникативного портрета адресата текста и в связи с этим выделяются основные «слагаемые» личности реципиента, необходимые для понимания адресованного ему текста. К ним относятся:

- индивидуальное знание адресата в той области, в рамках которой будет протекать коммуникативный акт (то есть непрерывно кон-струируемая и модифицируемая динамическая система данных, которыми располагает индивид);
- наличие специальных знаний в области, которой посвящен текст;
- объем активного тезауруса личности в данной области знаний (под этим термином понимается организованное знание, которым обладает субъект о словах и других вербальных символах).

Таким образом, к составу набора признаков для автоматического распознавания адресата текста можно отнести данные, полученные на основе словарного состава документа – подобной характеристикой может быть, например, отношение количества терминов к общему числу слов.

С рассматриваемой задачей тесно связаны исследования удобочитаемости или читабельности документов (Readability), опирающиеся на анализ синтаксиса и словаря текста. Основными критериями, оказывающими воздействие на значение показателя удобочитаемости, считаются количество слов в предложении, количество терминов в тексте, число символов в слове [7]. Число и состав критериев может меняться в зависимости от жанра, коммуникативной задачи и языка текста [8]. В качестве иных особенностей, влияющих на классификацию, можно указать количество сложносочиненных и сложноподчиненных предложений, количество обособлений, причастных и деепричастных оборотов. Одной из наиболее часто применяемых мер определения сложности восприятия текста читателем, адаптированных для русского языка, является индекс Флэша.

Он вычисляется, исходя из количества слов в тексте и в предложении, а также слогов в словах. Полученное значение находится в диапазоне от 0 (очень низкий уровень удобочитаемости) до 100 (очень высокий) [9]. Предлагается считать, что значение индекса от 90 до 100 соответствует уровню образования пятиклассника, а, например, значение от 0 до 30 – уровню студента вуза.

Процесс идентификации потенциального адресата текста подразумевает обращение к некому набору «эталонов» – базе знаний, отражающей характерные черты текстов, предназначенных для той или иной категории читателей (как с точки зрения словаря, так и с точки зрения синтаксиса) [10]. Для текста с неизвестной категорией будет требоваться определить его наиболее вероятный класс, то есть соотнести с одним из известных классов или с несколькими из них.

4. Заключение

В работе вкратце рассмотрена проблема реализации классификации текстов по категориям реципиентов и существующие научные направления, затрагивающие решение этой задачи на основе лексических и синтаксических характеристик текста. Предложено реализовать систему классификации текстов на основе распознавания их потенциальных адресатов.

Список литературы

1. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

2. Sebastiani F. Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47. – 2002.

3. Du R., RSafavi-Naini R., Susilo W. Web Filtering Using Text Classification, Proceedings of the 11th IEEE International Conference on Network (ICON 2003), pp. 325-330 – 2003.

4. Lipka N. Modeling Non-Standard Text Classification Tasks. – Weimar, Germany: Bauhaus-Universität Weimar, 2013. – 158 p.

5. Лотман Ю.М. Внутри мыслящих миров. – С.-Петербург: Языки русской культуры, 2000. – 464 с.

6. Каменская О.Л. Текст и коммуникация. – М.: Высшая школа, 1990. – 78 с.

7. Stephens C. All About Readability: [Электронный ресурс] // Plain Language. 2007-2010. URL: http://plainlanguage.com/newreadability.html (Дата обраще-ния: 19.03.2012).

8. DuBay W. The Principles of Readability: [Электронный ресурс] // Plain Language at Work Newsletter. 2013-2014. URL: http://www.impact-information.com/impactinfo/readability02.pdf (Дата обраще-ния: 19.03.2012).

9. Оборнева И. В. Автоматизация оценки качества восприятия текста. Вестник Московского городского педагогического университета, 2(5). – 2005.

10. Глазкова А.В. Возможность автоматического определения адресата на основе семантико-синтаксических особенностей текста / Открытые семантиче-ские технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2014): материалы IV междунар. науч.-техн. конф. – Минск: БГУИР, 2014. – 576 с.