OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

ПРОВЕРКА ИНФОРМАТИВНОСТИ КЛАССИФИКАЦИОННЫХ ПРИЗНАКОВ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Глазкова А.В.

Тюменский государственный университет,
г. Тюмень, Россия

anya_kr@aol.com

В статье рассматриваются базовые единицы автоматической обработки текстовой информации на примере автоматической классификации русскоязычной художественной литературы по ее возрастной аудитории. Экспериментально проверяется информативность ряда морфологических признаков, оцениваются лексические и некоторые количественные характеристики текстов различных категорий. Постановка данной задачи описывалась авторами в ходе выступления на конференции OSTIS-2014.
Ключевые слова: извлечение знаний; классификация; обработка естественного языка; семантический анализ.

Введение

Работа посвящена выявлению базовых признаков для автоматической обработки неструктурированной текстовой информации на русском языке на примере автоматической классификации художественной литературы по ее возрастной аудитории [Глазкова, 2014]. Эта задача актуальна в первую очередь для решения проблемы оптимизации информационного поиска в сети Интернет и хранилищах электронных документов.

Визуально определить возраст потенциального адресата текста обычно несложно, трудность заключается в поиске автоматического алгоритма решения задачи. В настоящий момент универсального набора признаков, позволяющего провести классификацию по заданному основанию, не выявлено, однако в работе ряда российских и зарубежных исследователей предлагаются отдельные признаки, которые могли бы быть положены в основание такой классификации. Часть признаков предложена авторами исследования. В работе проводится экспериментальная проверка информативности выявленных признаков.

В ходе исследования тексты делились на две категории: детские и взрослые. Это обусловлено наличием только этих двух категорий в выборке текстов, используемой для эксперимента. В дальнейшем планируется увеличить число классификационных категорий.

1. Корпус текстов

Для проведения эксперимента использовалась база Национального корпуса русского языка [НКРЯ, 2014]. Корпус состоит из заведомо качественных и максимально разнообразных текстов на русском языке с известным жанром. Авторы работали с двумя выборками – художественными текстами различных жанров (историческая проза, приключения, документальная проза и т.д., кроме детской литературы, – всего 5 902 документа, 9 332 659 предложений, 94 538 056 слов) и детской литературой (всего 632 документа, 547 735 предложений, 4 742 627 слов).

Также в работе использовалась [«База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)», 2014], состоящая из 510 текстов, ориентированных на читателей детского возраста.

2. Уточнение границ категорий

Поскольку в своем исследовании мы делим тексты на взрослые и детские, необходимо уточнить, что включают в себя данные категории.

В [ФЭБ, 2014] говорится, что детская литература включает в себя художественные, научно-художественные и научно-популярные тексты, написанные специально для детей, однако далее уточняется, что «обычно в понятие «детская литература» включается также широкий круг произведений для взрослых, прочно вошедших в

обиход детского чтения, — прежде всего, произв. народного творчества и классиков». Исходя из этого, можно сделать вывод о том, что границы категорий в задаче классификации текстов по возрасту их адресатов являются весьма размытыми. В данной работе под текстами, относящимися к детским, понимались те из них, которые:

- являются понятными детям с точки зрения лексики;
- соответствуют уровню их коммуникативного развития, являются информативными и представляют интерес для детской аудитории.

3. Краткий обзор литературы и выбор классификационных признаков

В ряде работ [Алексеева, 2009; Здир, 2009] в качестве отличительной характеристики текстов, предназначенных для младшей возрастной аудитории, указывается ограниченное использование автором функционально-смыслового типа речи описание. Под типом речи [Культура письменной речи, 2014] понимается способ изложения, избираемый автором и ориентированный на одну из задач: статически изобразить действительность, описать ее; динамически отразить действительность, рассказать о ней; отразить причинно-следственные связи явлений действительности. В соответствии с этими целями коммуникации выделяются три основных типа речи: описание, повествование, рассуждение.

В данном исследовании мы оценили возможность использования морфологических характеристик описания в качестве основания для классификации текстов по их возрастной аудитории.

Функционально-смысловой тип речи описание [Культура письменной речи, 2014] отличается от других типов речи тем, что он дает представление о каком-либо предмете или явлении перечислением их признаков и свойств. Ведущую роль в описании играют прилагательные, обеспечивающие выразительность и наглядность изображения. Описание как способ изложения противоположно повествованию и направлено на статическое отображение действительности, что выражается в меньшей в сравнении с другими типами речи частоте использования глаголов и глагольных форм.

В связи с этим в выборках проводился поиск следующих характеристик и их последующее сравнение:

- число глаголов;
- число особых глагольных форм (деепричастия, причастия);
- число прилагательных.

В соответствии с развернутым определением [Культура письменной речи, 2014] было выдвинуто предположение о том, что в текстах выборки $V_1 = \{T_1^{V_1}, T_2^{V_1}, ..., T_n^{V_1}\}$, где $T_k^{V_1}, k = \overline{1,n}$ – тексты художественной литературы, кроме детской, доля глаголов и особых глагольных форм должна быть выше, что в текстах выборки $V_2 = \{T_1^{V_2}, T_2^{V_2}, ..., T_l^{V_2}\}$, где $T_k^{V_2}, k = \overline{1,l}$ – тексты, относящиеся к детской литературе. И напротив – доля прилагательных в текстах выборки V_2 будет превышать соответствующий показатель в выборке V_1.

Также в качестве классификационных признаков были рассмотрены количественные признаки, основой для выбора которых послужили работы, посвященные оценке удобочитаемости (readability) текстов [Flesch, 1948; Оборнева, 2005; Шпаковский, 2012]. В качестве примера в данной работе оценены:

- среднее количество слов в предложении;
- средняя длина слов текста;
- процент многосложных слов в тексте (более трех слогов).

Кроме того, множества слов текстов из выборки V_1 и выборки V_2 были представлены в качестве моделей bag-of-words [EECS, 2014], далее было проведено сравнение полученных множеств.

4. Эксперимент

4.1. Морфологические характеристики

Значения итоговых показателей представлены в таблице 1, где для каждой выборки и морфологической характеристики приводится частота рассматриваемых частей речи:

$$F_{V_i} = \frac{N'_{V_i}}{N_{V_i}}, \quad (1)$$

где N_{V_i} – общее число слов в выборке, N'_{V_i} – число вхождений части речи в соответствии с результатами поиска.

Таблица 1 – Морфологические характеристики, полученные на основе выборок

Выборка	Частота (F_{V_i})		
	Глаголы	Особые формы глагола	Прилагательные
V_1	0,2	0,03	0,11
V_2	0,21	0,025	0,1

4.2. Количественные признаки

В таблице 2 отражены значения количественных признаков текстов двух выборок.

Таблица 2 – Количественные характеристики, полученные на основе выборок

Признак	Выборка	
	V_1	V_2
Среднее количество слов в предложении	11	6
Средняя длина слов текста (для слов, состоящих больше, чем из двух букв)	7	6
Процент многосложных слов в тексте (более трех слогов)	22,95	13,9

4.3. Лексические признаки

Для оценки лексического состава слова текстов каждой из рассматриваемых выборок были представлены в виде множества лексем, объединяющих в себе словоформы каждого встречающегося в тексте слова, и соответствующих им частотностей:

$$T_k^{V_i} = \{L_j, c_j\}, \quad (2)$$

где L_j – лексема, c_j – частотность данной лексемы.

Таким образом, на основе выборок и (2) были организованы модели bag-of-words:

$$M^{V_i} = \{L_j^{T_k^{V_i}}, c_j^{T_k^{V_i}}\}. \quad (3)$$

В дальнейшем были получены разности множеств, построенных на основе V_1 и V_2, включающие в себя только те лексемы, которые входят в одну из выборок:

$$M^{V_1} \setminus M^{V_2}, \quad M^{V_2} \setminus M^{V_1}. \quad (3)$$

Стоит отметить, что стоп-слова при построении множеств не исключались, поскольку они по определению оказались бы исключенными при вычитании множеств.

Пример выборки случайных лексем, входящих в разности множеств, представлен в таблице 3.

Таблица 3 – Примеры лексем, входящих в разности множеств

$M^{V_1} \setminus M^{V_2}$ (лексемы, отсутствующие в текстах детской литературы)	$M^{V_2} \setminus M^{V_1}$ (лексемы, присутствующие только в текстах детской литературы)
siemens	буквочка
гуманный	булочный
директива	бульканье
иконопись	булькин
исхлестать	капустница
навытяжку	колоннада
ноющий	накостылять
обобрать	нисколечко
обустройство	остроухий
окопаться	пернатое
синергетик	пилка
столыпинский	примерещиться
сторублевка	таращиться
татуировать	трусоватый
третьекурсник	уголек

5. Обсуждение результатов

Значения, полученные для морфологических признаков текстов (таблица 1), позволяют увидеть, что предположение, выдвинутое в пункте 3 и касающееся преобладающих типов речи, является верным. В соответствии с итоговыми показателями частота глаголов в выборке 1 составляет 0,2 (то есть 20%), особых глагольных форм – 0,03, прилагательных – 0,11. При этом частоты для выборки 2 – 0,21, 0,025 и 0,1 соответственно.

Несмотря на полученное подтверждение, различие в полученных значениях невелико, что не позволяет использовать данные характеристики в качестве классификационного признака.

Значения количественных признаков текстов (таблица 2) демонстрируют более существенные различия между текстами двух выборок. Предполагается, что полученный список информативных признаков можно расширить, добавив в него характеристики, связанные с синтаксической сложностью предложения (число грамматических основ, количество и вид придаточных предложений и т.д.).

Лексические признаки также могут послужить основанием для классификации. Безусловно, размер рассматриваемых выборок не является достаточным для формирования разностей множеств, исчерпывающе характеризующих словарный состав текстов, адресованный разным возрастным категориям читателей, однако даже на данном этапе работы полученные данные являются довольно информативными.

Заключение

Авторы благодарят НП "Национальный корпус русского языка" за предоставление базы текстов для проведения экспериментов, а также «Фонд Михаила Прохорова», финансировавший тревел-грант в рамках Открытого благотворительного конкурса «Академическая мобильность», программа «Образование как социальный институт», блок «Наука, образование, просвещение».

Библиографический список

[Flesch, 1948] Flesch R. A new readability yardstick / R. Flesch // Journal of Applied Psychology, 1948, №32, С. 221-233.

[EECS, 2014] Bag-of-words representation of text / EECS. [Электронный ресурс]. –2014. – Режим доступа: https://inst.eecs.berkeley.edu. – Дата доступа: 27.11.2014.

[Алексеева, 2009] Редактирование детской литературы / Алексеева М.И.; – М.: Московский государственный университет, 2009.

[«База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)», 2014] «База данных метатекстовой разметки Национального корпуса русского языка» (коллекция детской литературы)». [Электронный ресурс]. –2014.

[Глазкова, 2014] Глазкова А.В. Возможность автоматического определения адресата на основе семантико-синтаксических особенностей текста / А.В. Глазкова // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2014): материалы конференции, 2014, С. 509-513.

[Здир, 2009] Здир В.В. Возраст читателя и специфика произведения для детей/ В.В. Здир // Практическая психология и логопедия, 2009, № 4, С. 44-47.

[Культура письменной речи, 2014] Культура письменной речи. [Электронный ресурс]. –2014. – Режим доступа: gramma.ru. – Дата доступа: 07.05.2014.

[НКРЯ, 2014] Национальный корпус русского языка. [Электронный ресурс]. –2014. – Режим доступа: ruscorpora.ru. – Дата доступа: 28.05.2014.

[Оборнева, 2005] Оборнева И.В. Автоматизация оценки качества восприятия текста/ И.В. Оборнева // Вестник Московского городского педагогического университета, 2005, №5, С. 86-91.

[ФЭБ, 2014] Фундаментальная электронная библиотека «Русская литература и фольклор». [Электронный ресурс]. –2014. – Режим доступа: http://feb-web.ru.. – Дата доступа: 18.11.2014.

[Шпаковский, 2012] Шпаковский Ю.Ф. Оценка трудности восприятия текста / Ю.Ф. Шпаковский // Труды БГТУ. Издательское дело и полиграфия, 2012, №9, С. 72-75.

CLASSIFICATION FEATURES INFORMATIONAL CONTENT TESTING FOR AUTOMATIC NATURAL TEXTS CLASSIFICATION TASK

Glazkova A.V.

Tyumen State University, Tyumen, Russia

anya_kr@aol.com

The article deals with the basic units of automatic text processing on the example of automatic classification of the Russian fiction for its age audience. Experimentally the author verify informational content of several morphological characteristics, estimate lexical and some quantitative characteristics of texts of different categories. The statement of this problem is described by the authors in a report at the conference OSTIS-2014.

Introduction

The work is devoted to identifying the basic features for automatic processing of unstructured text information in Russian on the example of the automatic classification of fiction for its age audience. This problem is important for solving the optimization problem of information retrieval on the Internet and storages of electronic documents.

Visual determination of the age of a potential recipient of the text is not difficult, but the difficulty is in finding an automatic algorithm for solving the problem.

Currently, universal set of features allowing classification under the specified base, is not revealed, but in several Russian and foreign researchers offered some features that could be used as a basis of this classification. Some features proposed by the authors of the study. The paper deals with experimental verification of the informational content of the identified features.

In the research all texts were divided into two categories: children and adults. It's planning to increase the count of categories in the future.

Main Part

This research estimated three types of features. We call these types are: morphological, quantitative and lexical.

Morphological features include:

- number of verbs;
- number of special verbs forms (transgressives, participles);
- number of adjectives.

Quantitative features are:

- average count of words in a sentence;
- average length of words in a text;
- percent of polysyllable words in a text (more than three).

For estimation of lexical features of the text the author suggests creating bag-of-words model for both children and adults texts. Then by the subtraction sets of words the author offer to make a list of words that are present in the same category and absent in the other.

It should be noted that the stop words in the construction of the sets were not excluded, since they are by definition would have been excluded by subtracting sets.

Conclusion

Values obtained for the morphological features allow confirming the differences for different age categories. Despite the received acknowledgment, the difference in the values obtained is small, which makes use of characteristic data as a classification feature.

The values of quantitative traits texts show a significant difference between the texts of the two samples. It is assumed that the list of informative features can be expanded by adding the characteristics associated with the syntactic complexity of sentences (the number of grammatical foundations, the number and type of subordinate clauses, etc.).

Lexical features may also serve as a basis for classification. Of course, the dimensions of the samples is not sufficient to form the set difference exhaustively characterize the vocabulary of the text addressed to different age groups of readers, but even at this stage of the findings are quite informative.