DOI: 10.34220/MAMSP_174-178
УДК 004.9

АНАЛИЗ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ АЛГОРИТМА ЧАСТОТНО-КОНТЕКСТНОЙ КЛАССИФИКАЦИИ К ТЕКСТАМ РАЗНЫХ СТИЛЕЙ

М.О. Чайка[1], И.А. Бунеев[1], В.А. Величко[1]

[1]ФГБОУ ВО «Воронежский государственный лесотехнический университет имени Г.Ф. Морозова»

В данной статье рассматривается применение алгоритма частотно-контекстной классификации к текстам различных стилей. Выделяются основные особенности различных стилей, влияющие на эффективность алгоритма. Доказано, что метод выделения тематики текста при помощи алгоритма частотно-контекстной классификации лучше всего работает применительно к научным и юридическим документам и, в текущем виде практически неприменим для художественных текстов. Это делает актуальной задачу модификации алгоритма для определения тематики художественных текстов.

Ключевые слова: частотно-контекстная классификация, тематика, тексты, стили текстов, алгоритм.

ANALYSIS OF THE EFFECTIVENESS OF APPLYING THE FREQUENCY-CONTEXT CLASSIFICATION ALGORITHM TO TEXTS OF DIFFERENT STYLES

M.O. Chaika[1], I.A. Buneev[1], V.A. Velichko[1]

[1]Voronezh State University of Forestry and Technologies named after G.F. Morozov

This article discusses the application of the frequency-context classification algorithm to texts of various styles. The main features of different styles that affect the efficiency of the algorithm are highlighted. It is proved that the method of selecting the subject of the text using the frequency-context classification algorithm works best in relation to scientific and legal documents and, in its current form, is practically inapplicable for literary texts. This makes the task of modifying the algorithm to determine the subject of literary texts relevant.

Keywords: frequency-context classification, subject matter, texts, text styles.

Накопленные к настоящему времени колоссальные объемы информации, в совокупности с непрерывно увеличивающимися темпами ее роста определяют актуальность и значимость исследований в области информационного поиска [1, 2]. Одним из методов поиска документов и текстов является определение тематически близких текстов при помощи алгоритма частотно-контекстной классификации (далее алгоритм). В его основе лежит выборка частотно значимых слов дополненных связанными с ними контекстом словами при помощи чего достигается более точное определение тематики текста.

Эффективность применения данного метода может различаться в зависимости стиля текстов, к которым он применяется. Таким образом, становится актуальной задача определения эффективности алгоритма к текстам различных стилей.

Классификация функциональных стилей еще не вполне устоялась и проходит в настоящее время стадию изменений, а именно рассматривается определение таких новых стилей, как религиозный и ораторский. В рамки данной статьи исследование новых стилей не входит, поэтому было принято решение рассматривать исторически устоявшиеся. Одним из подходов к классификации основан на том, что каждой сфере деятельности человека соответствует определённый стиль. Так для сферы массовой информации используется публицистический стиль, для деловой и политической – официально-деловой, для научной сферы – научный стиль, а для обиходно-разговорной – разговорный. Представленные стили Д.Э. Розенталь организовал в структуру [3], разделённую на две подкатегории и пять разновидностей стилей (рисунок 1) в соответствии с излагаемой ими информацией. Так информация, изложенная разговорным стилем (первая категория), используется людьми для общения. Категорию книжных стилей целесообразно поделить по предназначению на четыре основных вида: научный, художественный, официально-деловой и публицистический (рисунок 1).

К разговорному стилю (РС) относятся: заметки, поговорки, пословицы, рассказы и беседы. К научному стилю (НС) – рецензии, рефераты, доклады, словари, учебники. Художественный стиль (ХС) включает повести, романы, баллады поэмы и т.п. В свою очередь официально-деловой стиль (ОДС) – постановления, указы и расписки. А к публицистическому стилю (ПС) следует отнести репортажи, эссе, ораторские речи, статьи, фельетоны, листовки.

Рассмотрим основные особенности каждого из стилей более подробно.

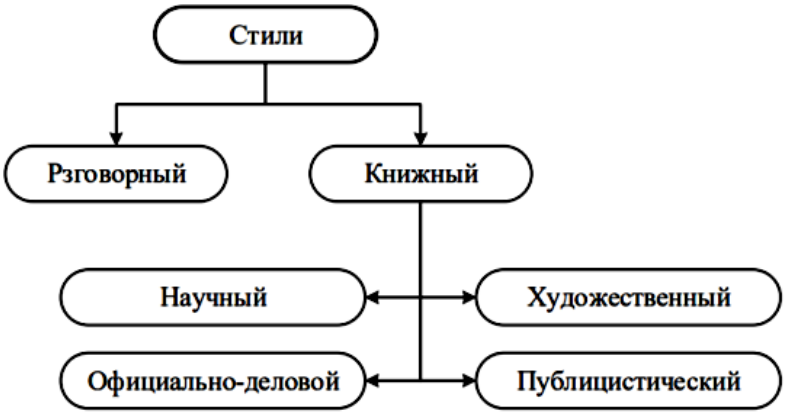

Рисунок 1 – Классификация стилей

Параметр, определяющий качество работы алгоритма – это соотношение частоты слов, имеющих прямое отношение к тематике текста к частоте остальных слов. Если общие слова встречаются чаще, чем определяющие тему – определить тему текста при помощи алгоритма не представляется возможным [4-9].

Разнообразие слов и словесных форм аналогично негативно влияет на качество выделения тематики текста.

Таким образом, в текстах научного стиля, в которых многократно повторяется одно или несколько ключевых понятий (терминов), применение алгоритма позволяет с наибольшей точностью определить тематику текста.

То же верно и для юридических документов официально-делового стиля. Здесь помимо тематически важных слов и словосочетаний повторяются так же многие не ключевые термины и понятия, однако это достаточно легко исправляется при помощи применения словаря, содержащего наиболее распространенные слова и выражения, не имеющие отношения к тематике текста.

В художественном стиле, для удобства читателя, авторы предпочитают избегать повторения слов, используют метафоры, описания и размытые формулировки. Тематика зачастую формируется не несколькими словами или словосочетаниями, а целыми абзацами текста. Наиболее часто в художественных текстах повторяются распространенные, не несущие тематической нагрузки слова, зачастую глаголы (Например «говорит» или «сказал»). Все вышеперечисленное существенно осложняет применение алгоритма для определения тематики текста и делает его практически неэффективным.

Хотя для публицистического стиля характерны многие черты художественных произведений необходимость четко определить основную идею и донести ее до слушателей вынуждает автора конкретно сформулировать основные тезисы и несколько раз использовать их в одной формулировке. Точность алго-

ритма применительно к таким текстам несколько ниже, однако, является допустимой.

На основании вышесказанного можно сделать вывод: метод выделения тематики текста при помощи алгоритма частотно-контекстной классификации лучше всего работает применительно к научным и юридическим документам и, в текущем виде практически неприменим для художественных текстов. Это делает актуальной задачу модификации алгоритма для определения тематики художественных текстов.

Список литературы

1. Лавлинский, В.В. Информационные системы для извлечения данных из неструктурированного текста с использованием онтологий / В.В. Лавлинский, Ю.О. Зольникова // Моделирование систем и процессов. – 2018. – Т. 11, № 3. – С. 30-34.

2. Лавлинский, В.В. Правила формирования сложных связей из неструктурированного текста / В.В. Лавлинский, Ю.О. Зольникова // Моделирование систем и процессов. – 2018. – Т. 11, № 3. – С. 34-39.

3. Розенталь Д.Э. Справочник по русскому языку. Практическая стилистика / Д.Э. Розенталь //Издательский дом «ОНИКС 21 век»: Мир и образование. – 2001. – 381 с.

4. Хазеев, Д.Р. Приложение нейронных сетей к определению стиля текста / Д.Р. Хазеев // Новые информационные технологии в автоматизированных системах. – 2019. – № 22. – С. 121-124.

5. Гращенко, Л.А. Библиографические ссылки как фактор квалиметрии научных текстов / Л.А. Гращенко, Г.В. Романишин // Новые информационные технологии в автоматизированных системах. – 2017. – №20. – С. 89-94.

6. Стилистический энциклопедический словарь русского языка ; под ред. М.Н. Кожиной. – М. : Флинта: Наука, 2011. – 696 с.

7. Чубур, Т.А. Дискурсивная объективация вербально-ментальных единиц в русской и английской лингвоконцептосферах в ракурсе закономерностей исторических изменений языка / Т.А. Чубур // Моделирование систем и процессов. – 2017. – Т. 10, № 1. – С. 76-80.

8. Лавлинский, В.В. Применение математического описания действий для целенаправленных систем на основе методов нейронных сетей / В.В. Лавлин-

ский, С.Н. Яньшин // Моделирование систем и процессов. – 2017. – Т. 10, № 2. – С. 17-23.

9. Лавлинский, В.В. Теоретические основы формирования моделей и методов взаимодействия информационных процессов / В.В. Лавлинский, И.И. Струков // Моделирование систем и процессов. – 2018. – Т. 11, № 2. – С.31-37.