УДК 004.8

ПОДХОД К ОЦЕНКЕ ПРИНАДЛЕЖНОСТИ ТЕКСТОВ ОДНОЙ ТЕМАТИКЕ

Каспранская А.И., Сметанина О.Н.

ФГБОУ ВО «Уфимский государственный авиационный технический университет», Уфа, e-mail: annakaspranskaya@gmail.com, smoljushka@mail.ru

Данная статья посвящена определению принадлежности двух текстов одной теме. В ходе работы был проведен анализ современного состояния проблемы, показавший актуальность работы и выявивший слабые стороны в предложенных решениях: работа только с большим объемом текстов, использование заранее подготовленных корпусов близости языка. Важной особенностью поставленной перед авторами задачи является произвольная и заранее неизвестная тематика текстов, которая не позволяет обучить собственную модель на некотором ограниченном наборе данных. Для преодоления этого ограничения была использована готовая модель для построения эмбеддингов, которая обучена на большом наборе текстов русской художественной литературы, включающем в себя более 300 тыс. текстов, с размером словаря 5×105 элементов. В основу предложенного решения положены перевод текстов в векторное представление и нахождение косинусовой близости между ними. Точность работы была определена в ходе вычислительного эксперимента и составила 97,9 % на наборе объёмом 1000 пар текстов, основанном на наборе парафраз для русского языка. Данное значение точности показало работоспособность предлагаемого автором решения проблемы определения принадлежности текстов общей теме.

Ключевые слова: обработка текстов на естественном языке, эмбеддинг, косинусовая близость, принадлежность текстов, семантическая близость текстов, модель машинного обучения

APPROACH TO THE ASSESSMENT OF TEXTS BELONGING TO THE SAME SUBJECT

Kaspranskaya A.I., Smetanina O.N.

Ufa State Aviation Technical University, Ufa, e-mail: annakaspranskaya@gmail.com, smoljushka@mail.ru

This article is devoted to determining whether two texts belong to the same topic. In the course of the work, an analysis of the current state of the problem was carried out, which showed the relevance of the work and revealed the weaknesses in the proposed solutions: work only with a large amount of texts, the use of pre-prepared corpora of language proximity. An important feature of the task set before the authors is the arbitrary and previously unknown subject matter of the texts, which does not allow training your own model on some limited data set. To overcome this limitation, a ready-made model for building embeddings was used, which was trained on a large set of texts of Russian fiction, including more than 300 thousand texts, with a dictionary size of 5 × 105 elements. The proposed solution is based on the translation of texts into a vector representation and finding the cosine proximity between them. The accuracy of the work was determined in the course of a computational experiment and amounted to 97.9 % on a set of 1000 pairs of texts based on a set of paraphrases for the Russian language. This value of accuracy showed the operability of the solution proposed by the author of the problem of determining whether texts belong to a common theme.

Keywords: natural language processing, embedding, cosine proximity, text ownership, semantic proximity of texts, machine learning model

Технологии искусственного интеллекта стремительно развиваются и позволяют решить многие задачи, что зачастую ранее сделать было невозможно. Вопросы лингвистического взаимодействия машины и человека связаны с областью обработки естественного языка. Одним из примеров решения задач в этой области является создание чат-ботов, таких что человек не всегда поймет, что с ним общается «искусственный интеллект». В целом область обработки естественного языка включает в себя множества задач, таких как распознавание речи, машинный перевод, выявление спама, голосовые помощники и пр.

В области анализа текста немаловажную роль играет задача оценки сходства текстов, в том числе и семантического. Среди исследователей в области обработки естественного языка, посвятивших свои работы оценке сходства текстов, можно отметить таких авторов, как В.Б. Барахнин, В.А. Нехаева, А.М. Федотов [1], А.Е. Письмак, А.Е. Харитонова, Е.А. Цопа, С.В. Клименков [2], А.В. Крюкова [3], Хиен Т. Нгуен, Фук Х. Дуонг [4], А. Розева, С. Зеркова [5], А.Х. Хакимова, М.М. Чарнин, А.А. Клоков, Е.Г. Соколов [6], Б. Мааке, С. Оджо, Т. Зува [7].

Активный интерес исследователей к вопросу автоматической обработки текста подтверждает актуальность тематики в области решения вопроса об оценке семантической близости текстов. Ряд работ посвящен математической оценке сход-

ства текстов, что не всегда может учитывать смысловую часть слова. Например, использование меры близости вместо меры сходства может породить ошибочные результаты (возможно ошибочное использование меры Левенштейна и др.). В другой части исследований можно обратить внимание, что решение связано с определенной заранее известной тематикой, построением ключевых слов либо векторного представления слов по заранее заданным текстам. Анализ современного состояния проблемы позволил сделать заключение, что такие решения не подходят для заранее неизвестных текстов произвольной тематики.

Исследование посвящено определению принадлежности двух текстов одной теме. В ходе работы необходимо провести анализ современного состояния проблемы, рассмотреть предложенные решения. Важной особенностью поставленной задачи является произвольная и заранее неизвестная тематика текстов, которая не позволит обучить собственную модель на некотором ограниченном наборе данных.

Современное состояние проблемы

В рамках решения задач NLP, например расширения запросов при поиске, оценке схожести текстов и др., используются различные меры сходства.

Для оценки меры близости текстов могут использоваться следующие метрики: с использованием n-грамм, расстояние Левенштейна, вычисление самых длинных подпоследовательностей, подстрок, косинусовая близость и пр.

Автор А.В. Крюкова в своей работе [3] при сравнении различных мер показала, что наибольшая важность для итогового решения имеет косинусовая близость.

Для оценки близости наборов текстов также используются алгоритмы кластеризации с выделением некоторых характеристик текстов [1], при этом методы кластеризации нельзя применить при оценке смысловой близости двух текстов.

Одним из способов оценки могут служить алгоритмы с использованием заранее подготовленных корпусов для языка [8], в которых для каждого слова определена его семантическая близость с другими словами. К сожалению, такие корпуса разработаны не для всех языков или не находятся в свободном доступе.

Результаты анализа современного состояния проблемы показали, что нужно разработать подход, позволяющий оценить принадлежность двух текстов одной тематике.

Постановка задачи

Формальная постановка задачи может быть представлена с использованием нотации IDEF0 (рис. 1).

Для задачи оценки сходства заранее неизвестных текстов на русском языке на произвольные темы используется функция определения принадлежности текстов одной теме.

Рис. 1. Формальная постановка задачи определения семантической близости текстов

Дано: два произвольных текста – $text_j$ где j=1, 2, состоящих из отдельных слов $text_i = (t_{ij}, .., t_{ni})$,$text_j$ – j-й текст, t_{ij} – i-е слово в j-м тексте, n – количество уникальных слов.

Определить: значение принадлежности –

$$sim(text1, text2) = \begin{cases} 1, (text1 \in A) \wedge (text2 \in A) \\ 0, (text1 \notin A) \vee (text2 \notin A) \end{cases},$$

где $sim(text1, text2)$ – функция определения принадлежности текстов одной теме, $text1$ и $text2$ – два текста, для которых определяется принадлежность, A – некоторая тема, которой могут принадлежать тексты.

На вход программы подается два текста, на выходе получаем числовое значение семантической близости из диапазона [0, 1]. При этом, чем значение ближе к 1, тем ближе семантическая близость текстов.

Предлагаемый подход к решению задачи

В постановке задачи имеется два текста на различные, заранее не определенные темы. Неопределенность темы является важной особенностью данной задачи, так как это ограничение не позволяет натренировать собственную модель машинного обучения для задачи классификации. Тексты представлены на русском языке. Предлагается следующий алгоритм решения (рис. 2).

Для обработки текста машиной необходимо привести его к более подходящему для обработки виду. Все шаги алгоритма насмотрены на примере текста о центральном процессоре: «Центра́льный проце́ссор – электронный блок либо интегральная схема, исполняющая машинные инструкции (код программ), главная часть аппаратного обеспечения компьютера или программируемого логического контроллера. Иногда называют микропроцессором или просто процессором».

Проводится процесс лемматизации, то есть слова заменяются нормальной формой. Такой набор слов называется набором токенов. Далее токены собираются в мешок слов. Мешок слов – представление текста в виде набора нормализованных уникальных слов текста. Мешок слов переводится в мешок векторов, то есть каждому элементу сопоставляется вектор, полученный из модели машинного обучения.

Текст подготовлен, и для определения близости текстов было расширено следующее правило – чтобы определить близость двух слов, нужно перевести их в векторы и посчитать косинус угла между ними.

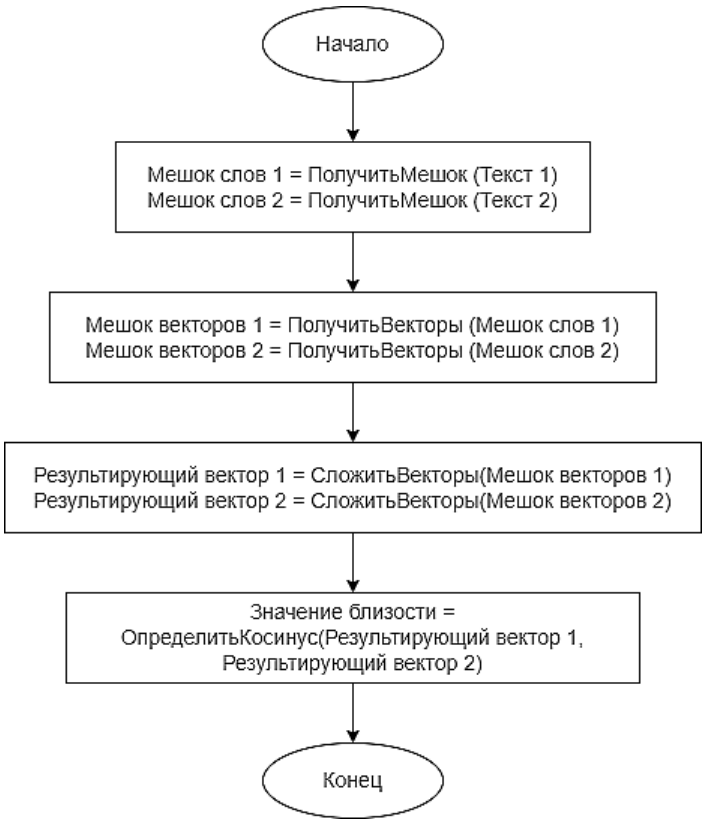

Рис. 2. Алгоритм решения задачи оценки близости текстов

Алгоритм перевода текстов можно описать следующим образом:

1. Для каждого токена из текста подбираем эмбеддинг (вектор). Можно представить текст как

$$text_j = (t_{ij}, .., t_{ni}) => (e_{ij}, .., e_{nj}) = emb_j,$$

где $text_j$ – набор токенов входящих в j-й текст, t_{ij} – i-й токен в j-м тексте, e_{ij} – i-й эмбеддинг (вектор) в j-м тексте, n – количество уникальных токенов в тексте и соответствующих векторов, emb_j – представление текста в виде набора векторов.

2. Складываем все вектора и получаем результирующий для всего текста

$$ResV_j = \sum emb_j,$$

где $ResV_j$ – результирующий вектор для текста j, emb_j – набор соответствующих токенам текста векторов.

Так, тексты определяются в n-мерном векторном пространстве (рис. 3).

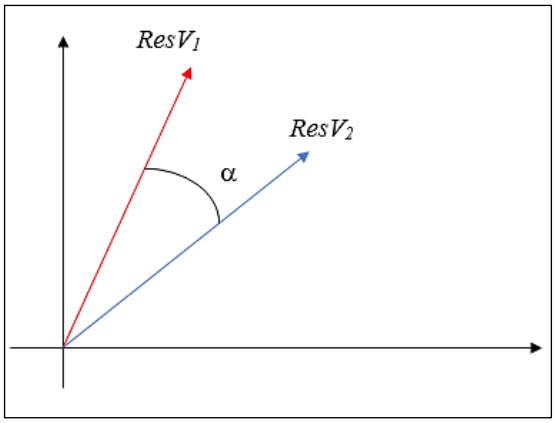

Рис. 3. Представление текстов в виде векторов

Представление в векторном пространстве такое, что чем ближе находятся векторы друг к другу, тем более близкие значения они представляют с точки зрения семантики. Из чего следует, что угол а1) на техническую тему – «Микропроце́ссор – процессор (устройство, отвечающее за выполнение арифметических, логических операций и операций управления, записанных в машинном коде), реализованный в виде одной микросхемы или комплекта из нескольких специализированных микросхем (в отличие от реализации процессора в виде электрической схемы на элементной базе общего назначения или в виде программной модели)»;

2) текст из ботаники – «Рома́шка – род многолетних цветковых растений семейства Астровые, или Сложноцветные, объединяет около двадцати видов невысоких пахучих трав, цветущих с первого года жизни».

Для всех текстов были посчитаны результирующие векторы, а после косинусовые меры близости для пар текстов:

1. Центральный процессор – микропроцессор (близость – 0,8698).

2. Центральный процессор – ромашка (близость – 0,1428).

Полученные результаты показывают, что тексты на одну тематику, техническую, имеют более близкое значение семантической близости, чем тексты на разные тематики.

При решении задачи использованы библиотеки проекта Natasha, которые предоставляют инструменты для решения базовых задач обработки естественного русского языка, а именно – сегментацию на токены и предложения, синтаксический и морфологический анализ, лемматизацию, извлечение именованных сущностей. Для сопоставления токенам эмбеддингов использовалась модель hudlit_12B_500K_300d_100q из библиотеки Navec. Эта модель имеет размер словаря в 500 000 записей и была обучена на текстах художественной литературы объемом 145 GB, благодаря этому она покрывает 98 % слов в художественных текстах [9].

Вычислительный эксперимент

Для проверки работоспособности предлагаемого решения задачи и оценки его корректности был составлен набор из 1000 парафраз на основании корпуса парафраз В. Гудкова и О. Митрофановой [10]. Оригинальный корпус парафраз был размечен по смыслу предложений. Фразы: 1. «У меня есть пять яблок» и 2. «У меня нет пяти яблок» ранжировались как различные. Для рассматриваемого исследования такое ранжирование не подходит, так как цель – определить принадлежность текстов одной теме. А в данном случае обе фразы о яблоках. Поэтому способ ранжирования был изменен.

Для определения границы принадлежности текста теме была проведена серия из 8 запусков алгоритма с разными показателями в диапазоне [0,4; 0,57]. Границы диапазона определялись в случае существенного ухудшения точности алгоритма. Результаты прогонов представлены на рис. 4, и итоговая граница была определена – 0,51. Это значит, что если значение косинусовой близости текстов больше или равно этому значению, то тексты определяются как принадлежащие одной теме, в противном случае тексты принадлежат различным темам.

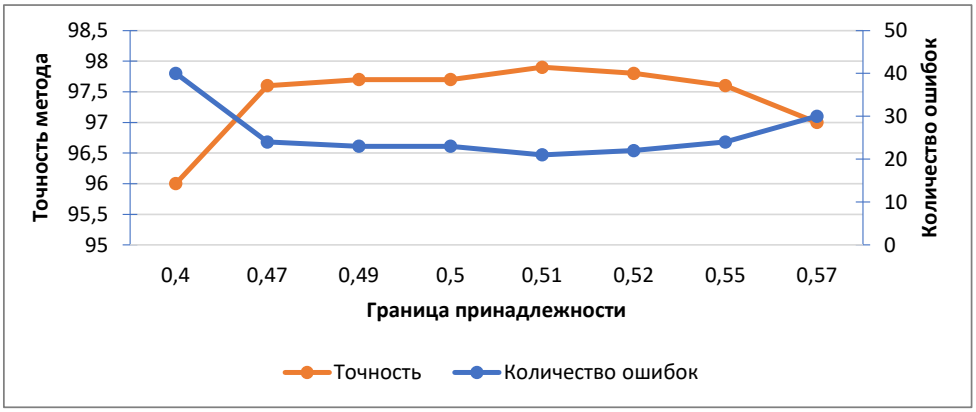

Рис. 4. Подбор значения для определения границы принадлежности текста

В результате прогона набора из тысячи пар текстов была получена точность решения – 97,9 %. Это является хорошим показателем для модели, обученной на художественной литературе русского языка без дополнительного дообучения узким предметным областям, которым могли принадлежать тексты.

Заключение

Результаты анализа текущего состояния проблемы в области оценки сходства текстов показал ее актуальность и необходимость более эффективного решения задачи.

В основу предложенного решения положены перевод текстов в векторное представление и нахождение косинусовой близости между ними. Особенностью решения является использование модели, обученной на больших наборах разнообразных текстов на русском языке. Это позволяет процесс распознавания близости между текстами представить наиболее похожим на то, как это делает человек. Однако модель представляет и самое уязвимое место в решении: если тексты специфические, то модель может не знать используемых слов; если сравниваются тексты не очень близкие друг другу по смыслу, то для корректной работы требуется дополнительная обработка текстов – нахождение уникальных слов внутри тем с отбрасыванием лишних.

При решении была использована готовая модель для построения эмбеддингов (библиотеки проекта Natasha), которая обучена на большом наборе текстов русской художественной литературы, включающем в себя более 300 тыс. текстов, с размером словаря 5×105 элементов. Точность работы была определена в ходе вычислительного эксперимента и составила 97,9 % на наборе объёмом 1000 пар текстов, основанном на наборе парафраз для русского языка. Идея показала свою работоспособность и может иметь практическое применение в решении задач определения близости текстов для различных тематик.

Результаты исследований, приведенные в статье, частично поддержаны грантом РНФ 22-19-00471.

Список литературы

1. Барахнин В.Б., Нехаева В.А., Федотов А.М. О задании меры сходства для кластеризации текстовых документов // Вестник НГУ. Серия: Информационные технологии. 2008. № 1. С. 87–97.

2. Письмак А.Е., Харитонова А.Е., Цопа Е.А., Клименков С.В. Оценка семантической близости предложений на естественном языке методами математической статистики // Научно-технический вестник информационных технологий, механики и оптики. 2016. № 2. С. 324–330.

3. Крюкова А.В. Определение семантической близости текстов с использованием инструмента DK Pro Similarity // Компьютерная лингвистика и вычислительные онтологии. Выпуск 1. 2017. С. 87–97.

4. Hien T. Nguyen, Phuc H. Duong, Cambria E. Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowledge-Based Systems. 2019. Vol. 182. 104842.

5. Rozeva A., Zerkova S. Assessing semantic similarity of texts. Methods and algorithms. AIP Conference Proceedings. 2017. P. 060012-1–060012-8.

6. Khakimova A.Kh., Charnine M.M., Klokov A.A., Sokolov E.G. Approaches to assessing the semantic similarity of texts in a multilingual space. Physics and technology proceedings (CPT2020). Nizhny Novgorod, 2020. P. 3–9.

7. Maake, Benard Zuva, Tranos. A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems. Conference on Information Communications Technology and Society (ICTAS). 2018. P. 1–5.

8. Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare Voss, Jiawei Han. Scalable Topical Phrase Mining from Text Corpora. Proceedings of the VLDB Endowment, 2014. Vol. 8. P. 305–316.

9. Проект Natasha – набор Python-библиотек для обработки текстов на естественном русском языке: официальный сайт. URL: https://natasha.github.io/navec/ (дата обращения: 01.04.2022).

10. Гудков В., Митрофанова О. Автоматически ранжируемый русский корпус парафраз для генерации текста // Материалы четвертого семинара по нейронной генерации и трансляции. 2020. С. 54–59.