СТАТЬИ

УДК 004.89

ПРИМЕНЕНИЕ СТРУКТУРНОГО АНАЛИЗА ДЛЯ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТА НА РУССКОМ ЯЗЫКЕ БЕЗ ИСПОЛЬЗОВАНИЯ ЛЕКСИЧЕСКИХ ПРИЗНАКОВ

Батурин М.М., Белов Ю.С.

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», Калужский филиал, Калуга, e-mail: k4dys@yandex.ru

Определение стиля письма представляет собой комбинацию последовательных решений на разных уровнях обработки текстов на естественном языке, включая лексический, синтаксический и структурный, связанные с конкретным автором. Лексические, синтаксические и структурные признаки составляют три основных семейства стилистических признаков. Лексические признаки отражают характер автора и предпочтения в использовании слов, а синтаксические признаки фиксируют синтаксические модели предложений в документе. Структурные особенности раскрывают информацию о том, как автор организует структуру текста. Одной из основных проблем, редко затрагиваемой в литературе, является взаимодействие стиля и содержания. В то время как содержательные слова могут быть признаками авторского стиля письма из-за того, что они несут информацию о лексическом выборе авторов, исключение содержательных слов в качестве характеристик является фундаментальным шагом для предотвращения определения темы, а не определения стиля. Однако синтаксические и структурные особенности не зависят от содержания, что делает их устойчивыми к расхождению тем. Предлагаемое решение определяет стиль автора текста исходя из структуры написанного им текста, что делает модель устойчивой к изменению темы.

Ключевые слова: определение авторства текста, структурный анализ текста, теги POS

APPLAYING STRUCTURAL ANALYSIS TO DETERMINE THE AUTHORSHIP OF A TEXT IN RUSSIAN WITHOUT USING OF LEXICAL FEATURES

Baturin M.M., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: k4dys@yandex.ru

The definition of writing style is a combination of sequential solutions at different levels of natural language text processing, including lexical, syntactic and structural, related to a specific author. Lexical, syntactic and structural features make up three main families of stylistic features. Lexical features reflect the author's character and preferences in the use of words, and syntactic features fix syntactic models of sentences in the document. Structural features reveal information about how the author organizes the structure of the text. One of the main problems rarely touched upon in literature is the interaction of style and content. While meaningful words can be signs of the author's writing style due to the fact that they carry information about the lexical choice of the authors, the exclusion of meaningful words as characteristics is a fundamental step to prevent the definition of the topic, not the definition of style. However, syntactic and structural features do not depend on the content, which makes them resistant to divergence of themes. The proposed solution determines the style of the author of the text based on the structure of the text written by him, which makes the model resistant to changing the topic.

Keywords: determining the authorship of the text, structural analysis of the text, POS tags

При решении проблемы определения авторства текста часто используются модели, так или иначе опирающиеся на лексические данные текста [1–3], однако подобный подход теряет эффективность в случае, когда изменяется тема текста, так как с ней значительные изменения претерпевает и набор используемых слов. Структурные же особенности, как правило, не зависят от содержания, что означает, что они в основном согласуются между различными документами, написанными конкретным автором. Предложенная модель реализует идею определения авторства текста на основе структурных признаков.

Цель исследования – изучить способы определения авторства текста при помощи структурного анализа.

Общая структура модели

Для кодирования синтаксических шаблонов документа в иерархической структуре используется синтаксическая рекуррентная нейронная сеть. Во-первых, каждое предложение представляется как последовательность тегов POS (part of speech) [4], и каждый тег POS встраивается в малоразмерный вектор [5], а кодировщик тегов POS, представляющий собой двунаправленную LSTM (long short term memory), создает синтаксическое представление предложений [6]. Впоследствии полученные представления предложений объединяются в представление документа [7] посредством двунаправленной LSTM.

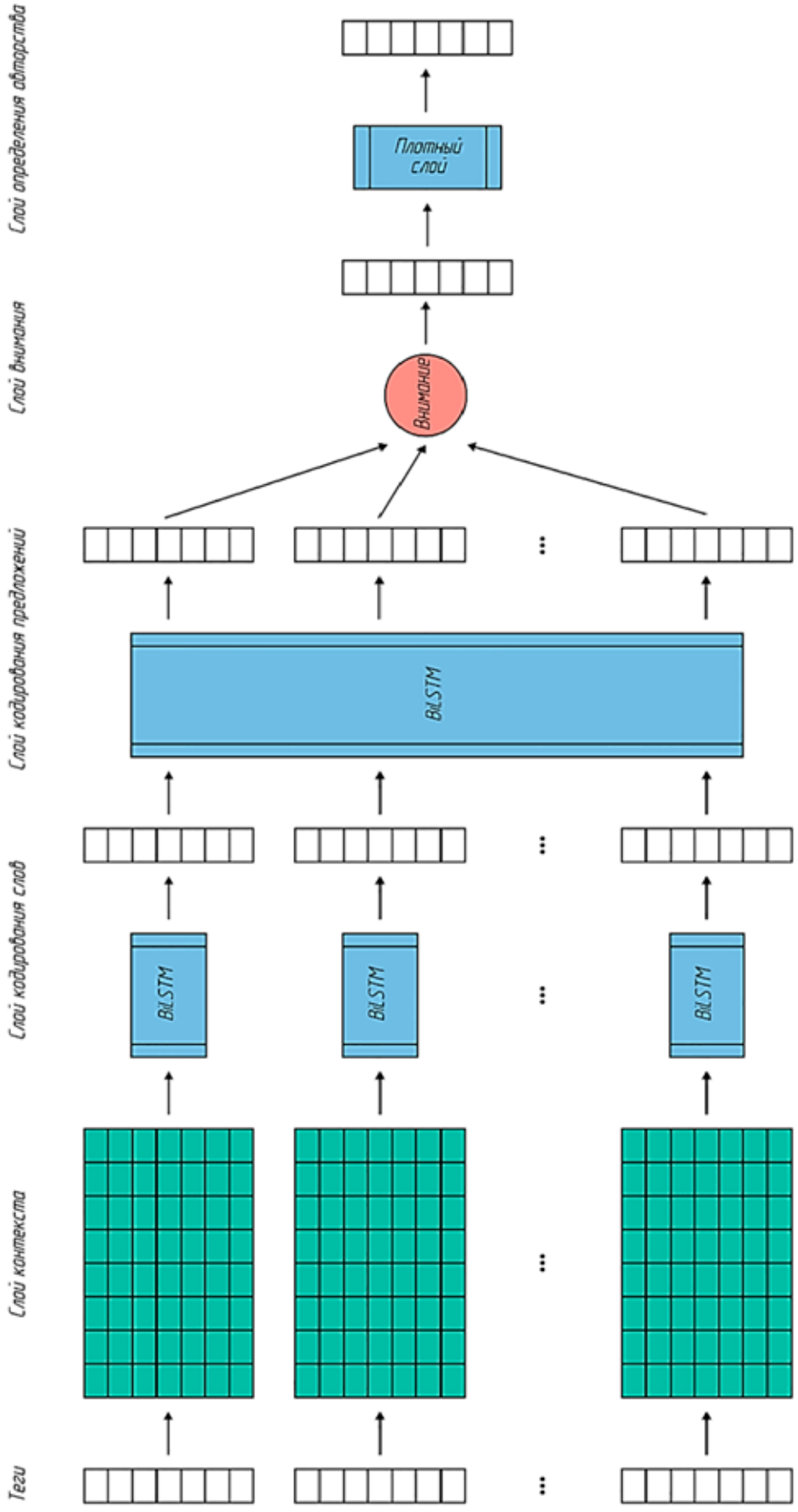

Рис. 1. Структура модели

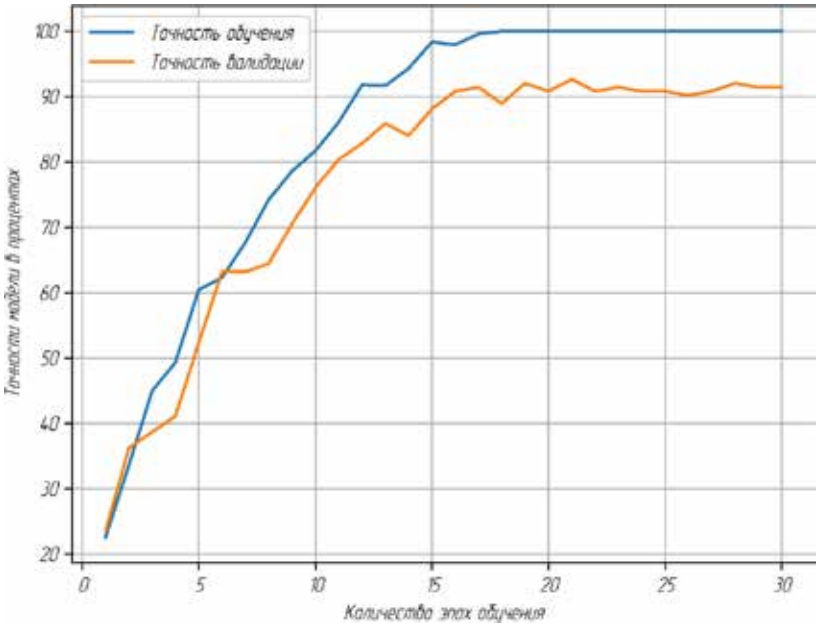

Рис. 2. Ошибка на уровне сегмента обучение/валидация

Кроме того, используется механизм внимания, чтобы вознаграждать предложения, которые больше способствуют предсказанию меток. После этого для вычисления распределения вероятностей по меткам классов используется плотный слой (рис. 1).

Точность модели на уровне сегмента в зависимости от количества эпох обучения

Для определения точности модели при обучении используются две следующие метрики: значение функции потерь и точность. Поскольку в задаче определения авторства присутствуют несколько выходных меток, в качестве функции потерь используется категориальная кросс-энтропия [8].

График зависимости значения функции потерь от количества эпох обучения представлен на рис. 2. Из представленного графика видно, что потери стремительно уменьшаются первые 16 эпох, после чего практически не изменяются. После первой эпохи потери при обучении равняются приблизительно 2,4. После 16 эпох значение функции при обучении стремится к нулю. При валидации после первой эпохи потери приблизительно равны 2,2, после 16 эпох обучения значение функции потерь приближается к 0,24 и далее продолжает колебаться около этого значения.

Точность модели на уровне сегмента высчитывается исходя из вероятности приписывания моделью выбранного сегмента истинному автору. Как показано на рис. 3, она, так же как и функция потерь, существенно изменяется первые 16 эпох обучения, достигая после них значений, близких к максимальным. Одна эпоха обучения дает модели точность около 23% при обучении и валидации. После 16 эпох точность при обучении приближается к 100%, точность же при валидации достигает значения в 91% и далее колеблется около него.

Важно отметить, что представленные выше показатели измеряются на уровне сегмента, итоговое же предсказание наиболее вероятного автора осуществляется на основе слияния результатов всех сегментов данного на вход модели текста при помощи *Fuse*-функции. В силу этой особенности также здесь не приводится зависимость точности классификации тестовых текстов от количества эпох обучения модели.

Настройка гиперпараметров модели

Настройка модели является важным этапом, ведь от выбранных параметров в значительной степени будет зависеть конечная производительность модели, ее точность и временные показатели. Необходимо подобрать оптимальные формат предложения и длину сегмента. Данные, исходя из которых осуществляется подборка, получаем в результате проведения экспериментов на корпусе текстов на русском языке.

Обобщим полученные в ходе тестирования точности модели на уровне сегмента данные:

1. Малая длина предложения в 10 или 20 слов дает существенную разницу, а также слишком большой разброс в точности в сравнении с длиной в 30 слов.

2. Длина предложения в 40 слов является избыточной, поскольку прибавка к точности при большой длине сегмента (100–200 предложений) крайне мала.

3. Размер сегмента в 20 и 50 предложений дает большой разброс по точности при любой длине предложения, так же результаты на уровне сегмента при таком размере значимо меньше, чем при длине в 100 или 200 предложений.

4. Сегмент в 200 предложений на первый взгляд кажется лучшим решением, так как при длине предложения в 30 слов дает прирост минимальной точности в 2% и медианной в 3%, однако нужно помнить, что проводится сравнение показателей точности модели на уровне сегмента, а не текста, и увеличение длины одного фрагмента данных в 2 раза даст также и двукратное уменьшение данных для следующего этапа – слияния.

Исходя из полученных данных, оптимальными параметрами модели были выбраны длина предложения в 30 слов, что позволит получать близкую к максимальной точность при минимальной возможной вычислительной сложности и длина сегмента в 100 либо 200 предложений, окончательный выбор длины сегмента будет произведен на следующем этапе. Результаты тестирования на уровне сегмента послужат основой для последующих экспериментов. Далее, настроив гиперпараметры модели, можно переходить с уровня сегмента непосредственно к определению авторства на уровне всего текста.

Сравним точность определения автора текста на выборке из 19 текстов, по одному на каждого представленного в корпусе автора. Эти тексты исключаются из выборки, на оставшихся текстах обучается модель с установленными гиперпараметрами. Из рис. 4 видно, что модель с установленной длиной сегмента в 100 предложений правильно классифицировала все тексты, при длине сегмента в 200 предложений модель не смогла правильно распознать авторов трех текстов.

В результате тестирования модели на уровне текста была выбрана длина сегмента, равная 100 предложениям, дающая меньшую точность классификации при обработке каждого отдельного сегмента, но показывающая значительное улучшение после слияния результатов ее работы *Fuse*-функцией на уровне всего текста.

Тестирование модели на романе Л.Н. Толстого «Анна Каренина»

В качестве первого примера работы предложенной модели рассмотрим случай определения авторства романа «Анна Каренина», написанного Л.Н. Толстым.

Как видно из рис. 5, модель правильно определила автора 80% сегментов текста, приписав их настоящему автору. Другие 20% сегментов были поделены между другими присутствующими в наборе авторами.

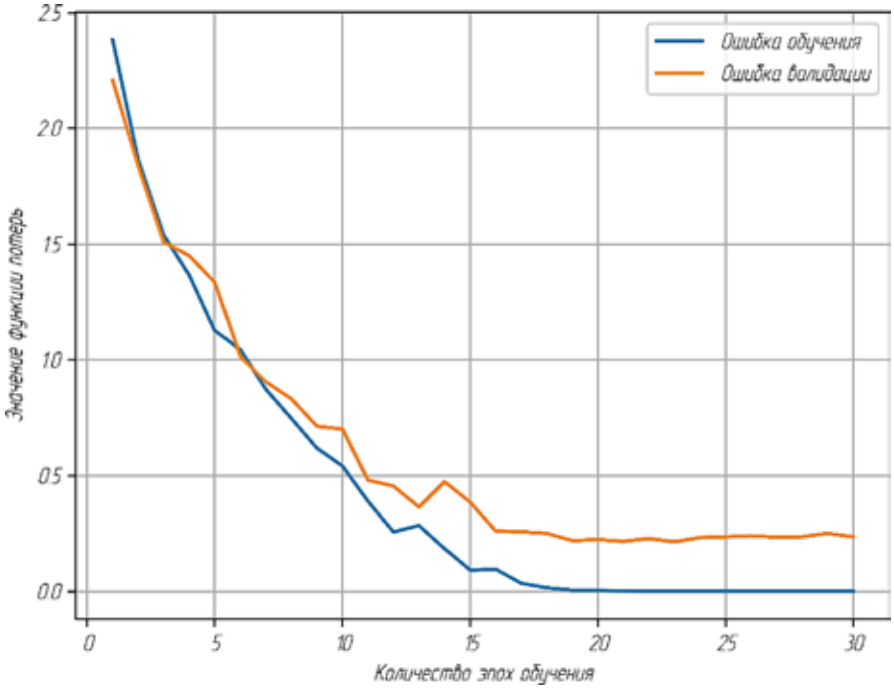

Рис. 3. Точность на уровне сегмента обучение/валидация

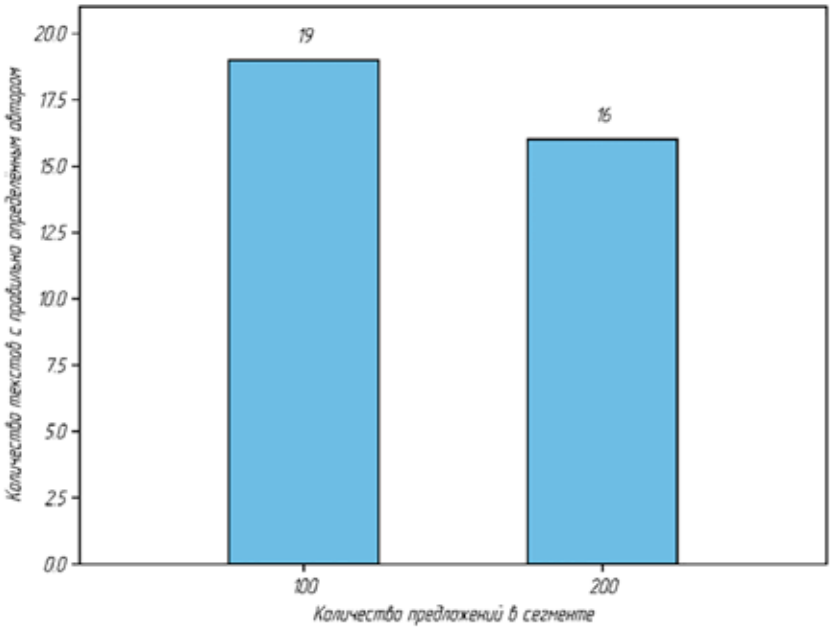

Рис. 4. Сравнение точности модели на уровне текста при разной длине сегмента

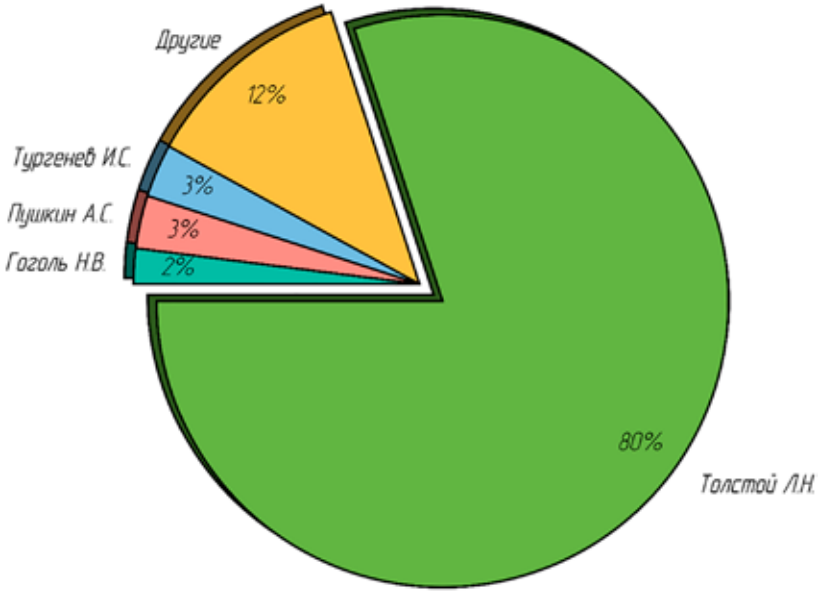

Рис. 5. Предсказанные авторы сегментов из произведения Толстого Л.Н.

Наибольшее сходство было выявлено с Тургеневым, Пушкиным и Гоголем, оставшиеся 12% сегментов были распределены между другими авторами, однако для понятного отображения данных на графике все авторы с маленьким процентом сходства (то есть с малым количеством приписанных им сегментов текста) помещены в общую графу «Другие».

Вероятностные показатели корректного определения авторства текста для различных авторов

Рассмотрим показатели точности классификации модели на уровне сегмента для нескольких авторов, как видно из столбчатой диаграммы на рис. 6, показатели точности классификации модели на уровне сегмента варьируются от 68 до 80%.

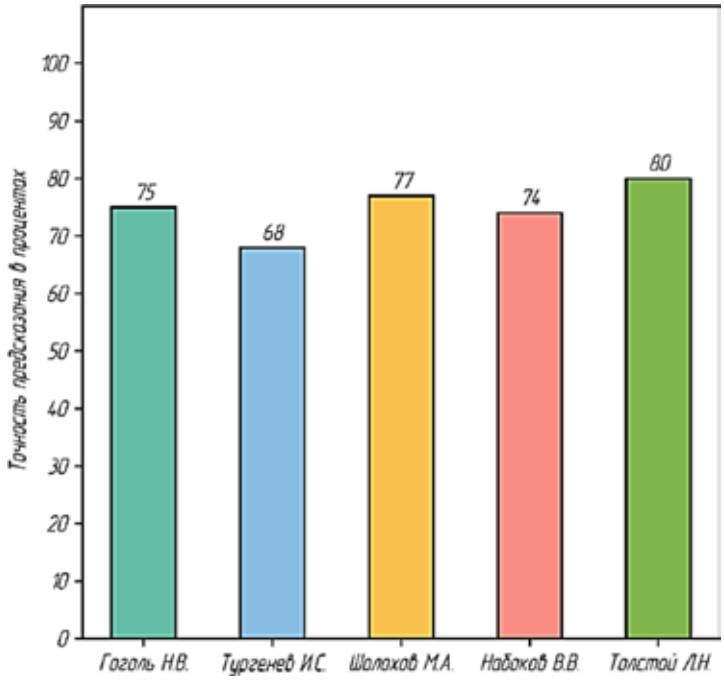

Рис. 6. Процент сегментов с правильно определенным автором

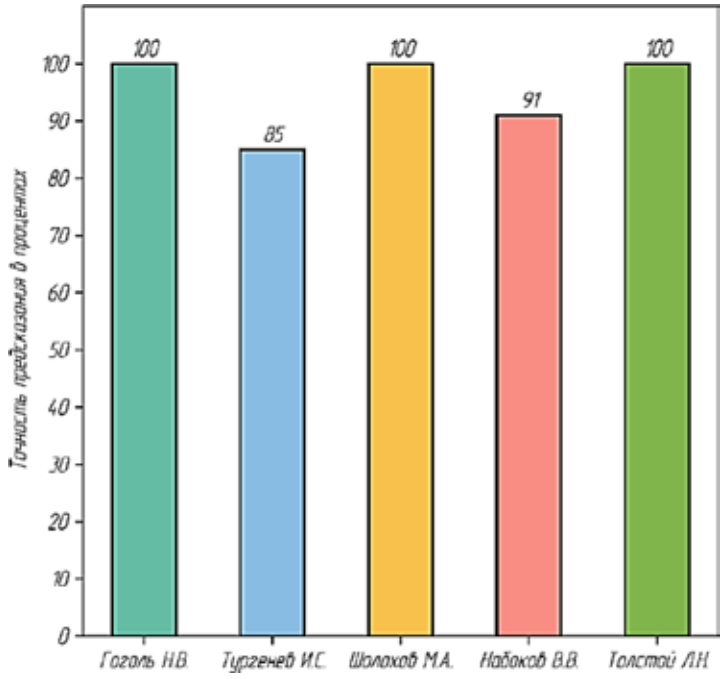

Рис. 7. Процент правильно определенных текстов

Также, исходя из полученных данных, можно вывести закономерность, что чем лучше отдельно взятый автор представлен в обучающей выборке, тем выше вероятность правильной классификации сегмента текста, относящегося к этому автору. Так Л.Н. Толстой, представленный в обучающем корпусе самым большим набором текстов (как по количеству самих произведений, так и по длине каждого из них), имеет самый большой шанс правильной классификации сегмента написанного им текста.

Далее от определения авторства сегмента перейдем к тестированию определения авторства полных текстов. Как видно из рис. 7, для нескольких авторов модель успешно определила авторство всех имеющихся в корпусе текстов. Так как результаты, полученные на отдельных сегментах текста, объединяются, теоретически достаточно иметь больше половины правильно классифицированных сегментов, чтобы определить истинного автора заданного текста. На практике же иногда случается, что один сегмент структурно можно с примерно равной вероятностью отнести сразу к двум авторам, вследствие чего результаты голосования по сегментам не всегда бывают на 100 % точными.

Заключение

В данной статье было произведено тестирование работы модели на представленном корпусе произведений на русском языке, с определением показателей точности на уровне сегментов и целых текстов для различных авторов. Представлены графики точности модели при обучении и валидации в зависимости от количества эпох обучения. Визуализированы результаты предсказания моделью авторства текста Л.Н. Толстого, с вероятностями принадлежности текста различным авторам.

Список литературы

1. Gómez-Adorno H., Posadas-Durán J.P., Sidorov G. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. Computing. 2018. P. 1–16.

2. Shrestha P., Sierra S., González F. Convolutional Neural Networks for Authorship Attribution of Short Texts. 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. Vol. 2. P. 669–674.

3. Stamatatos E. Authorship Attribution Using Text Distortion. 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. Vol. 1. P. 1138–1149.

4. Stein R., Jaques P., Valiati J. An Analysis of Hierarchical Text Classification Using Word Embeddings. Information Sciences. 2018. Vol. 471. P. 216.

5. Батурин М.М., Белов Ю.С. Применение многозадачного обучения для определения авторства текста на основе механизма конкурентного внимания // Научное обозрение. Технические науки. 2022. № 3. С. 5–9.

6. Zhang R., Hu Z., Guo H. Syntax encoding with application in authorship attribution. Conference on Empirical Methods in Natural Language Processing. 2018. P. 2742–2753.

7. Sundararajan K., Woodard D. What represents ‘style’ authorship attribution? 27th International Conference on Computational Linguistics. 2018. P. 2814–2822.

8. Soler J., Wanner L. On the relevance of syntactic and discourse features for author profiling and identification. 15th Conference of the European Chapter of the Association for Computational Linguistics. Vol. 2. 2017. P. 681–687.