Перенос авторского стиля текста на основе модифицированного tf-idf взвешивания слов

С. В. Мохнаткин, email: mokhnatkin.sv@gmail.com [1]

[1] Вятский государственный университет

Аннотация. *В данной работе рассматривается модель для переноса авторского стиля русскоязычных текстов. Для решения задачи переноса авторского стиля был собран текстовый корпус из произведений десяти авторов XIX века. Для определения важности слов авторского стиля использована модифицированная модель tf-idf взвешивания. Для поиска близких по смыслу слов применяются векторные представления word2vec. Результаты тестирования показали, что предложенная модель работоспособна и позволяет определять наиболее характерные для автора слова и использовать их для переноса авторского стиля текста.*

Ключевые слова: *Перенос стиля текста, авторский стиль, взвешивание слов, векторные представления, word2vec, tf-idf.*

Введение

Перенос стиля текста (англ. text style transfer) – это важная и сложная задача в области обработки естественного языка. В работе [1] приведены следующие примеры использования переноса стиля текста: персонализированные диалоговые системы, генерация новостных заголовков, машинный перевод в определенном стиле, анонимизация автора текста. В работе [11] представлен интеллектуальный текстовый редактор, позволяющий менять стиль текста.

Для английского языка существует множество работ по данному направлению [1, 2], однако для русского языка существует лишь одно исследование [3], в котором рассмотрены методы детоксификации (замены оскорбительных слов) русскоязычных текстов с использованием больших языковых моделей.

Перенос авторского стиля также был исследован в работах [4, 12] только для англоязычных текстов, например с использованием корпуса [5] переводов текстов У. Шекспира на современный английский.

Таким образом, настоящая работа является первым исследованием методов переноса авторского стиля текста для русского языка.

Вклад настоящего исследования заключается в следующем:

– предложен метод взвешивания важности слов для стиля автора, основанный на модификации модели tf-idf [8];
– разработана модель для применения к тексту стиля заданного автора, основанная на поиске семантически близких слов с помощью модели word2vec [6];
– приведены результаты работы модели с использованием корпуса русской литературы XIX века, включающего произведения десяти авторов.

1. Постановка задачи

Рассмотрим формальное описание задачи переноса стиля текста. В работе [3] стилем считается характеристика корпуса, отдельная от текстов, представленная множеством категориальных значений (меток), например: {*позитивный*, *негативный*}, {*формальный*, *неформальный*}, {*стиль автора 1, стиль автора 2, …}* и т. д. Эта характеристика может быть измерена функцией (например классификатором), получающей текст и возвращающей метку стиля. В настоящей работе рассматриваются авторские стили текстов.

Пусть даны два текстовых корпуса:

$$D^X = \{x_1, x_2, \ldots, x_n\}, \quad (1)$$

$$D^Y = \{y_1, y_2, \ldots, y_n\}, \quad (2)$$

имеющие отличные друг от друга стили s^X и s^Y, где x_i и y_i – предложения. Задача состоит в построении модели

$$f : X \rightarrow Y, \quad (3)$$

где X и Y – все возможные предложения, имеющие, соответственно, стили s^X и s^Y. Результатом применения модели для предложения x_i является предложение y_i', сгенерированное моделью и имеющее смысловое содержание предложения x_i и стиль s^Y.

При оценке качества переноса стиля учитываются следующие характеристики предложения y_i':

1. Соответствие целевому стилю s^Y.
2. Соответствие смысловому содержанию исходного предложения x_i.
3. Естественность языка.

2. Описание корпуса

В рамках исследования был подготовлен корпус русской литературы XIX века, включающий 260 757 предложений, взятых из произведений 10 различных авторов. Предложения, входящие в корпус, отбирались из исходных текстов произведений по диапазону длины от 40 до 170 символов (промежуток выбран исходя из распределения длин предложений в корпусе, а также просмотра структуры предложений различной длины), после предварительной обработки (удаление примечаний редакции, удаление выражений на иностранных языках и чисел, удаление тегов разметки, отделение знаков препинания пробелами, удаление прочих шумовых символов, разделение предложений по знаку препинания «.»). В табл. 1 указаны авторы, присутствующие в корпусе, а также количество предложений для каждого автора.

Таблица 1

Авторы текстов, присутствующих в корпусе

Автор	**Количество предложений**	**Доля от общего количества, %**
А. П. Чехов	15 310	5,9
Ф. М. Достоевский	54 177	20,8
Н. В. Гоголь	13 262	5,1
И. А. Гончаров	37 878	14,5
Н. С. Лесков	37 575	14,4
А. Н. Островский	19 202	7,4
А. С. Пушкин	6 836	2,6
М. Е. Салтыков-Щедрин	24 382	9,4
Л. Н. Толстой	41 469	15,9
И. С. Тургенев	16 666	6,4
Всего	260 757	

3. Взвешивание слов

Работа предлагаемой модели делится на два этапа. На первом этапе определяются наиболее характерные для каждого автора слова. На втором этапе полученные слова используются для замены семантически близких слов из исходного предложения.

Для определения наиболее характерных для автора слов использована модифицированная модель tf-idf взвешивания [8].

В подготовленном корпусе тексты автора (далее авторский корпус, D^{Y}), стиль которого нужно применить к некоторому предложению,

отделяются от остальных текстов, которые при этом объединяются в один общий корпус (далее общий корпус, D^X). Затем составляются словари общего (далее V^X) и авторского (далее V^Y) корпусов – множества уникальных слов, содержащихся, соответственно, в общем и авторском корпусах. Для того, чтобы учесть все формы слов при вычислении весов, слова в корпусе были приведены к нижнему регистру, а также в нормальную форму с помощью парсера pymorphy2 [9].

В стандартной формулировке tf-idf частота слова tf рассчитывается в рамках одного предложения (документа) и только для слов этого предложения [8]. Для того, чтобы подсчитать веса для всего словаря корпуса, в данной работе tf рассчитывается суммарно для всех предложений в корпусе и для каждого слова из словаря. Кроме того, для приведения весов tf и idf к одному порядку от tf взят логарифм. Также особенность предложенной модификации заключается в том, что веса рассчитываются только для словаря авторского корпуса, при этом tf рассчитывается по авторскому корпусу, а idf – по общему корпусу.

idf для слов из словаря V^Y вычисляется по следующей формуле:

$$idf(t^Y, D^X) = \log \frac{|D^X| + 1}{|\{d_i^X \in D^X \mid t^Y \in d_i^X\}| + 1} + 1, \quad (4)$$

где t^Y – слово из словаря V^Y, d_i^X – документ общего корпуса, в котором встретился t^Y.

tf для слов из словаря V^Y вычисляется по следующей формуле:

$$tf(t^Y, D^Y) = \log(N(t^Y, D^Y)) + 1, \quad (5)$$

где $N(t^Y, D^Y)$ – количество вхождений слова t^Y в D^Y.

Таким образом, idf максимизирует вес слов, которые редко встречаются в D^X, а tf максимизирует вес слов, которые часто встречаются в D^Y. Произведение tf и idf позволяет определить какие слова наиболее свойственны для выбранного автора, так как такие слова будут иметь наибольший вес.

В табл. 2 приведены топ-20 слов по убыванию весов tf-idf для четырех авторов.

Таблица 2

Топ-20 слов по весам tf-idf

Н. В. Гоголь	А. С. Пушкин	Л. Н. Толстой	И. А. Гончаров
чичиков	пугачёв	левин	обломов
козак	дубровский	нехлюдов	марфинька
запорожец	оренбург	вронский	райский
акакий	германн	кить	тарантьев
ноздрево	мятежник	пьер	адуев
ковалёв	ибрагим	кутузов	фрегат
акакиевич	яицкий	денисов	японец
манилов	михельсон	долли	марк
андрий	швабрин	ростов	леонтий
собакевич	кузмич	анатоль	штольц
бульб	кирилович	маслов	якут
остап	оренбургский	долох	викентьев
селифан	бибиков	ростовый	посьета
костанжочь	кирил	болконский	адмирал
козацкий	савельич	элен	фаддеев
кошев	сильвио	аркадьич	едо
оксана	комендант	дутлов	Кичиб
парубок	кирджать	свияжский	колония
козаков	самозванец	катюша	савич
платонов	троекур	балашева	бен

4. Реализация модели

В предлагаемой модели переноса авторского стиля предусмотрено определение слов, близких по значению, так как модель должна сохранять смысл исходного предложения. Для этого на подготовленном текстовом корпусе была обучена модель word2vec [6] в реализации Gensim [7], позволяющая искать близкие по смыслу слова на основе их векторных представлений.

Рассмотрим алгоритм работы модели:

1. На вход модели поступает исходное предложение.
2. В предложении каждое слово рассматривается отдельно.
3. Для текущего слова находим N близких слов с помощью word2vec (в экспериментах N было установлено равным 5).

4. Найденные близкие слова проверяем на соответствие следующим условиям: слово есть в авторском корпусе; части речи совпадают; нормальные формы не совпадают; вес tf-idf не меньше, чем у исходного.
5. Если на предыдущем этапе нашлось несколько подходящих слов, выбираем по наибольшему весу tf-idf. Если не нашлось ни одного слова, оставляем исходное.

5. Результаты

В табл. 3–6 приведены примеры работы модели. В первом столбце указаны номера примеров, во втором столбце – авторы, в третьем столбце – сначала исходное предложение, затем предложение, сгенерированное моделью. Жирным выделены слова, измененные моделью.

Таблица 3

Перенос стиля Достоевский (Д) → Гоголь (Г)

1	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная** светская старушонка
	Г	ведь это вы тогда вошли, а? **старуха** ее просто **странная** светская старушонка
2	**Д**	но **девок** всего пришло только три, да и **марьи** еще не было
	Г	но **баб** всего пришло только три, да и **анны** еще не было
3	**Д**	еще от дворника узнал он, что **петрушка** и не **думал** являться
	Г	еще от дворника узнал он, что **селифан** и не **подумал** являться

Таблица 4

Перенос стиля Достоевский (Д) → Пушкин (П)

4	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная** светская старушонка
	П	ведь это вы тогда вошли, а? **дочь** ее просто **славная** светская старушонка
5	**Д**	но девок всего **пришло** только три, да и марьи еще не было
	П	но девок всего **приходило** только три, да и марьи еще не было
6	**Д**	еще от **дворника** узнал он, что **петрушка** и не **думал** являться
	П	еще от **кабака** узнал он, что **ямщик** и не **чувствовал** являться

Таблица 5

Перенос стиля Достоевский (Д) → Толстой (Т)

7	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная** светская **старушонка**
	Т	ведь это вы тогда вошли, а? **жена** ее просто **гадкая** светская **арестантка**
8	**Д**	но **девок всего пришло** только три, да и **марьи** еще не было
	Т	но **баб служебная приходило** только три, да и **княжны** еще не было
9	**Д**	еще от дворника **узнал** он, что **петрушка** и не **думал являться**
	Т	еще от дворника **увидал** он, что **ямщик** и не **чувствовал подниматься**

Таблица 6

Перенос стиля Достоевский (Д) → Гончаров (Г)

10	**Д**	ведь это вы тогда вошли, а? **мать** ее просто **смешная светская** старушонка
	Г	ведь это вы тогда вошли, а? **старуха** ее просто **испанская пожилая** старушонка
11	**Д**	но **девок** всего пришло только три, да и **марьи** еще не было
	Г	но **баб** всего пришло только три, да и **прасковьи** еще не было
12	**Д**	еще от дворника узнал он, что **петрушка** и не **думал являться**
	Г	еще от дворника узнал он, что **ямщик** и не **подумал ходить**

В примерах 1, 2, 3, 4, 10, 11, 12 модель удачно выполнила замену слов и эти примеры удачно передают стиль автора.

В примере 5 заменено всего одно слово, чего недостаточно для изменения стиля.

В примерах 6 (*не чувствовал являться*), 7 (*жена ее*), 9 (*не чувствовал подниматься*) нарушена естественность языка.

В примерах 6 (*от кабака узнал он*), 8 (*баб служебная приходило*) нарушена естественность языка, а также утерян смысл предложения.

Автоматическое тестирование проводилось при помощи классификатора SVC в реализации Scikit-learn [13]. В качестве обучающих данных выбраны все имеющиеся в корпусе предложения двух авторов – Ф. М. Достоевский и Л. Н. Толстой. В качестве тестовых данных из обучающих данных отделено по одной тысяче предложений

каждого автора. Затем, стиль всех тестовых предложений одного автора изменен на стиль противоположного автора с помощью предлагаемой модели. Для векторизации предложений использована модель tf-idf в реализации Scikit-learn [13]. Доля правильно угаданных меток (accuracy) в результате бинарной классификации составила 58%.

На данный момент согласование форм слова осуществляется посредством обучения word2vec на исходных текстах (без приведения в нормальную форму) и проверки соответствия частей речи при замене. Это приводит к тому, что не всегда соблюдается естественность языка, а также word2vec менее эффективно определяет сходство слов ввиду большого количества словоформ. Данную проблему можно решить обучением word2vec на текстах в нормальной форме и согласованием форм слова с помощью вариаций моделей GPT [10] и BERT [14].

Другая проблема заключается в том, что не всегда достаточно заменить отдельные слова для изменения стиля. В качестве альтернативы можно использовать замену n-грамм.

Также требуется исследовать влияние на качество работы модели других вариантов формул tf и idf и их комбинирования с коэффициентом сходства слов word2vec.

Заключение

В данной работе была рассмотрена модель переноса авторского стиля текста. Для русскоязычных текстов такое исследование проводится впервые. В рамках исследования был подготовлен корпус русской литературы XIX века, предложен метод оценивания значимости авторских слов для формирования стиля, разработана модель переноса стиля и приведены результаты работы модели.

В дальнейших исследованиях должны быть изучены перспективные способы улучшения качества работы модели, а также возможности комбинации предложенных методов с языковыми моделями глубокого обучения.

Список литературы

1. Deep Learning for Text Style Transfer: A Survey / D. Jin [et al.] // Computational Linguistics. – 2021. – P. 1-51.

2. Text Style Transfer: A Review and Experimental Evaluation / Z. Hu [et al.] // arXiv preprint. – 2021. – 45 p. – https://arxiv.org/pdf/2010.12742.pdf

3. Methods for Detoxification of Texts for the Russian Language / D. Dementieva [et al.] // Multimodal Technol. Interact. – 2021. – № 5(9): 54. – 26 p.

4. Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models / H. Jhamtani [et al.] // Proceedings of the Workshop on Stylistic Variation (Copenhagen, Denmark, September 7-11, 2017) – Copenhagen, 2017. – P. 10-19.

5. Paraphrasing for Style / W. Xu [et al.] // Proceedings of COLING 2012: Technical Papers (Mumbai, India, December 2012) – Mumbai, 2012. – P. 2899-2914.

6. Efficient Estimation of Word Representations in Vector Space / T. Mikolov [et al.] // arXiv preprint. – 2013. – 12 p. – https://arxiv.org/pdf/1301.3781.pdf

7. Gensim [Электронный ресурс] : библиотека с открытым исходным кодом для неконтролируемого тематического моделирования и обработки естественного языка. – Режим доступа : https://radimrehurek.com/gensim/

8. Д. Маннинг, К. Введение в информационный поиск. Пер. с англ. / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – М. : ООО «И. Д. Вильямс», 2011. – 528 с.

9. Pymorphy2 [Электронный ресурс] : морфологический анализатор. – Режим доступа : https://pymorphy2.readthedocs.io/en/stable/

10. Improving Language Understanding by Generative Pre-Training / A. Radford [et al.] // Preprint. – 2018. – 12 p. – https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

11. A Recipe For Arbitrary Text Style Transfer with Large Language Models / E. Reif [et al.] // arXiv preprint. – 2021. – 12 p. – https://arxiv.org/pdf/2109.03910.pdf

12. A Probabilistic Formulation Of Unsupervised Text Style Transfer / J. He [et al.] // arXiv preprint. – 2020. – 14 p. – https://arxiv.org/pdf/2002.03912.pdf

13. Scikit-learn [Электронный ресурс] : библиотека машинного обучения для языка программирования Python. – Режим доступа : https://scikit-learn.org/stable/

14. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.] // Proceedings of NAACL-HLT 2019 (Minneapolis, Minnesota, June 2-7, 2019) – Minneapolis, 2019. – P. 4171-4186.