

Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.89

Евтушенко
Елизавета Юрьевна

Перенос стиля текста с помощью нейронной сети

АВТОРЕФЕРАТ

на соискание степени

магистра информатики и вычислительной техники

по специальности 1-40 81 01 – **Информатика и технологии разработки
программного обеспечения**

Научный руководитель
Тихоненко Татьяна Владимировна
к.ф.-м.н., доцент

Минск 2020

Работа выполнена на кафедре информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

Научный руководитель: **ТИХОНЕНКО Татьяна Владимировна**,
кандидат физико-математических наук, доцент,
главный специалист отдела по подбору персонала
№1 ИООО «ЭПАМ Системз»

Рецензент: **ПЛЮЩ Олег Борисович**,
кандидат физико-математических наук, доцент
кафедры управления информационными
ресурсами учреждения образования «Академия
управления при Президенте РБ»

Защита диссертации состоится «25» июня 2020 г. в 15⁰⁰ часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Гикало, 9, корп. 4, ауд. 111, тел. 293-85-91, e-mail: inform@bsuir.by.

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

Нейронные сети применяются для работы с естественным языком с момента своего возникновения и продолжают достижение прогресса в задачах, которые начинали решаться в более ранних областях машинного обучения. Простые, механические задачи, вроде разметки текста, сегодня могут быть решены с высокой точностью. Но есть и другой класс задач, требующий от систем искусственного интеллекта работы со смыслом, заключённым в тексте, в определённой степени его понимания. Такие задачи требуют более сложных подходов, но при этом представляют особый интерес для исследователей.

Одним из видов таких задач можно назвать перенос стиля текста. Эта задача чаще всего формулируется так – перенос смысла предложения в определённый стиль.

В существующих работах рассматривались тексты на английском языке. Для русского языка широко известных результатов пока нет, и готовых наборов данных совсем немного. Кроме того, русский язык сложнее морфологически, что требует особого внимания к согласованию слов в генерируемых текстах. Это дополнительно усложняет задачу переноса стиля текста на русском языке.

Впрочем, работа с текстами на родном языке даёт исследователям преимущество – возможность наиболее точной оценки полученных результатов. Таким образом, два обстоятельства – новизна данной задачи применительно к русскому языку и возможность компетентной оценки – обусловили выбор текстов на русском языке в качестве данных для исследуемой модели.

Кроме научного, данная задача представляет и практический интерес. Применение полученных результатов может быть широким: персонализация автоматических систем ответов на вопросы, автогенерация различного контента, автоматическая модерация текста и решение других прикладных задач.

С одной стороны, эта задача похожа на задачу машинного перевода, с другой стороны, она явно была вдохновлена успехами переноса стиля в области изображений. Сходство с этими хорошо изученными задачами обуславливает выбор методов для достижения поставленной цели. Это широко используемая в задачах генерации данных генеративно-состязательная архитектура, с тем отличием, что генерирующий модуль представляет собой модифицированный автокодировщик. Такой подход уже был успешно использован для переноса содержимого отзывов на английском языке между доменами тональности.

Остаются нерассмотренными вопросы регуляризации полученной нейронной сети и работы с предложениями произвольной длины.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Существующие работы в области переноса стиля текста в основном посвящены изменению тональности текстов англоязычных отзывов на услуги.

Актуальной является разработка модели, рассматривающей иные категории текстов и стиля, а именно, дневников и их авторского стиля, к тому же написанных на русском языке.

Степень разработанности проблемы

Исследование переноса стиля текста с помощью нейронных сетей осуществлялось на основе построения теоретических моделей и с использованием работ зарубежных учёных: О. Ашуала (O. Ashual), Т. Юргенсона (T. Jurgenson), Д. Гринберга (D. Grinberg), Г. Лампля (G. Lample), С. Субраманиана (S. Subramanian), Э. Смита (E. Smith), Л. Денуайе (L. Denoyer), М.'А. Ранзато (M.'A. Ranzato), Ю.-Л. Бюро (Y-L. Boureau) и др.

Упущением является рассмотрение текстов только на английском языке. Предложенное исследование направлено на устранение этого пробела.

Цели и задачи исследования

Целью диссертационной работы является разработка архитектуры, нахождение оптимальных параметров и обучение нейронной сети для решения задачи переноса стиля между корпусами текстов на русском языке.

Для достижения поставленной цели необходимо решить следующие *основные задачи*:

1. Определить поставленную задачу в рамках понятий машинного обучения.
2. Изучить существующие подходы в данной области и найти наиболее подходящий для решения данной задачи.
3. Разработать архитектуру нейронной сети на основе выбранного подхода.
4. Обучить построенную нейронную сеть.
5. Проанализировать полученные результаты и обозначить пути дальнейшего улучшения модели.

Объектом исследования являются нейронные сети.

Предметом исследования являются методы переноса стиля текста с помощью нейронных сетей.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования нейронной сети для решения задачи переноса стиля между корпусами текстов на русском языке. Современные методы глубокого обучения позволяют решать подобные задачи, не имеющие чётких метрик и заранее известных правильных ответов. Исследование данной задачи применительно к текстам на русском языке позволит ускорить прогресс в этом направлении.

Теоретическая и методологическая основа исследования

В основу диссертации легли работы зарубежных учёных в области переноса стиля текста, а также анализ российских и зарубежных учебников по тематике глубокого обучения.

Информационная база исследования сформирована на основе литературы, открытой информации, сведений из электронных ресурсов и материалов научных конференций.

Научная новизна

Научная новизна и значимость полученных результатов работы заключается в разработке модели нейронной сети, решающей задачу переноса стиля текста на русском языке.

Теоретическая значимость работы заключается в анализе применимости существующих моделей NLP, решающих задачу переноса стиля, к текстам на русском языке.

Практическая значимость диссертации состоит в разработанной модели нейронной сети, решающей поставленную задачу.

Основные положения, выносимые на защиту

1. Распределённые представления слов успешно решают задачу представления текстов на русском языке в машинном обучении.
2. Модель, построенная на основе GAN, позволяет осуществлять перенос стиля текста с сохранением его смыслового содержания.

Личный вклад соискателя

Результаты, приведённые в диссертации, получены соискателем лично. Вклад научного руководителя Т. В. Тихоненко заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Тезисы некоторых результатов исследований, вошедших в диссертацию, докладывались на Международной научной конференции «Информационные технологии и системы 2019» (Минск, Беларусь, 2019).

Публикации

По теме диссертации опубликована 1 печатная работа в сборнике материалов международной конференции.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трёх глав, заключения, списка использованных источников, списка публикаций автора и приложений.

В первой главе представлен анализ предметной области, выявлены задачи, смежные с поставленной, проведён обзор существующих подходов к её решению.

Во второй главе проведено исследование методов глубокого обучения, позволяющих решать задачу переноса стиля текста, и последовательно построена архитектура соответствующей модели.

В третьей главе на основе полученной модели реализована нейронная сеть, представлены использованные технические средства, описан процесс сбора и подготовки данных, проанализированы полученные результаты.

В приложениях представлены листинги кода построенной нейронной сети и вспомогательных модулей обработки данных.

Общий объем работы составляет 62 страницы, из которых основного текста – 41 страница, 5 рисунков на 2 страницах, 2 таблицы на 2 страницах, 17 формул, список использованных источников из 52 наименований на 5 страницах и 5 приложений на 9 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении определена область и указаны направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В первой главе проведён обзор смежных задач и существующих работ в области переноса стиля текста с помощью нейронных сетей.

Обзор задач обработки естественного языка показал, что на сегодняшний день наиболее интересными и имеющими не до конца исследованные пути

решения являются задачи генерации текста на основе некоторой информации, полученной из различных источников, в том числе, и другого текста.

Для задачи переноса стиля текста нет единого определения, так как под стилем текста может пониматься множество атрибутов. В данной работе за стиль принимается авторский стиль Льва Толстого, извлечённый из корпуса текстов его дневников.

Задача переноса стиля текста возникла как отражение успехов в области переноса стиля изображений, однако специфика тестовых данных не позволяет применить те же подходы и требует иных решений.

Задача переноса стиля текста похожа на машинный перевод, и именно решения в этой области служат основой подходов к решению данной задачи, что и было обнаружено в ходе анализа существующих работ.

Существующие работы, как показал обзор [1], в основном посвящены изменению тональности текстов англоязычных отзывов на услуги, что показывает научную новизну данной работы, в которой рассматривается иная категория текстов, к тому же написанных на русском языке.

Результаты предыдущих исследований в данной области отражены в работах О. Ашуала (O. Ashual), Т. Юргенсона (T. Jurgenson), Д. Гринберга (D. Grinberg), Г. Лампля (G. Lample), С. Субраманиана (S. Subramanian), Э. Смита (E. Smith), Л. Денуайе (L. Denoyer), М.'А. Ранзато (M.'A. Ranzato), Ю.-Л. Бюро (Y-L. Boureau) и др.

Вторая глава посвящена анализу подходящих моделей и алгоритмов в области глубокого обучения и последовательному построению модели нейронной сети, решающей поставленную задачу.

Анализ возможных моделей представлений слов в задачах машинного обучения показал, что именно распределённые представления слов отвечают необходимым требованиям, а именно: сохранению семантических связей между словами в их представлениях (рисунок 1), небольшой размерности пространства полученных векторов и дифференцируемости этого пространства. В качестве формата представления входных и выходных данных модели выбраны распределённые представления слов в виде предобученных по модели *word2vec* 300-мерных векторов 4 млн. слов русского языка, полученных как результат проекта *FastText*.

Последовательная природа текста находит своё отражение в архитектуре моделей, предназначенных для его обработки. И на сегодняшний день общепринятой архитектурой нейронных сетей в области NLP является рекуррентная. Анализ существующих моделей рекуррентных слоёв показал, что наилучшим решением в случае обработки текстовых последовательностей будет двунаправленный рекуррентный слой с LSTM-ячейками.

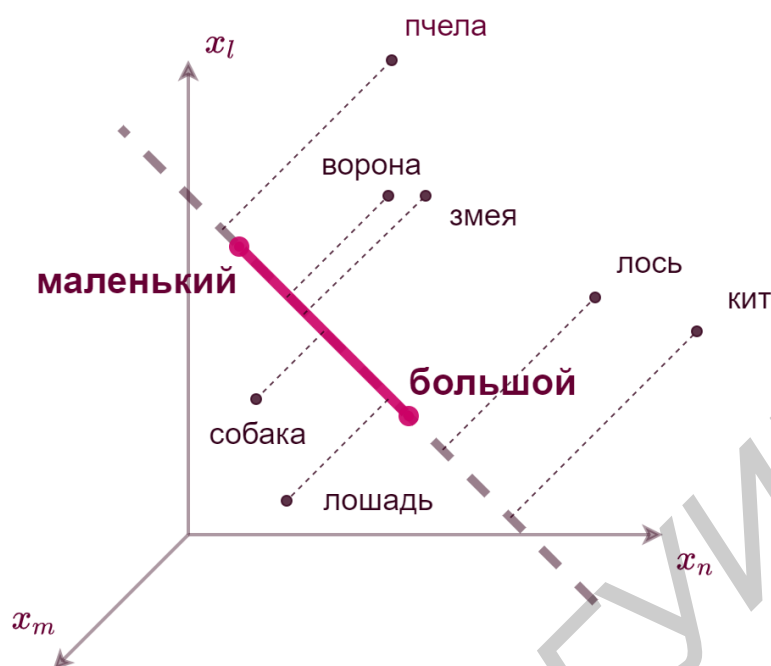


Рисунок 1 – Семантические связи в векторном пространстве

В задаче выделения из входных данных некоторых полезных признаков, которая в данной работе может быть определена как задача извлечения смыслового содержания текста, широко применяются высокоуровневые архитектуры нейронных сетей, основанные на идее автокодировщика. Для второй части задачи – переноса стиля предложений — применены идеи модификации автокодировщика, применяемые в машинном переводе. Полученный автокодировщик имеет сложную функцию потерь

$$L_{G \text{ full}} = L_{\text{recon}} + L_G = \begin{cases} -\frac{1}{n} \sum_{i=1}^n D(\hat{y}^{(i)}), & y \in S, \\ \sum_{i=1}^n \sum_{j=1}^m \frac{(\hat{y}_j^{(i)} - y_j^{(i)})^2}{nm}, & y \in T. \end{cases} \quad (1)$$

где S – домен-источник, T – целевой домен, n – количество образцов, m – длина предложения.

Для оценки качества переноса стиля в данной работе применяется решение, имеющее сегодня абсолютное превосходство в области генерации искусственных данных, – генеративно-сопоставительная архитектура (рисунок 2). Показано, что, несмотря на пока что слабую распространённость GAN в области генерации текста, решения, применённые в данной работе, а именно, представления слов в дифференцируемом пространстве, устраняют препятствия к их использованию.

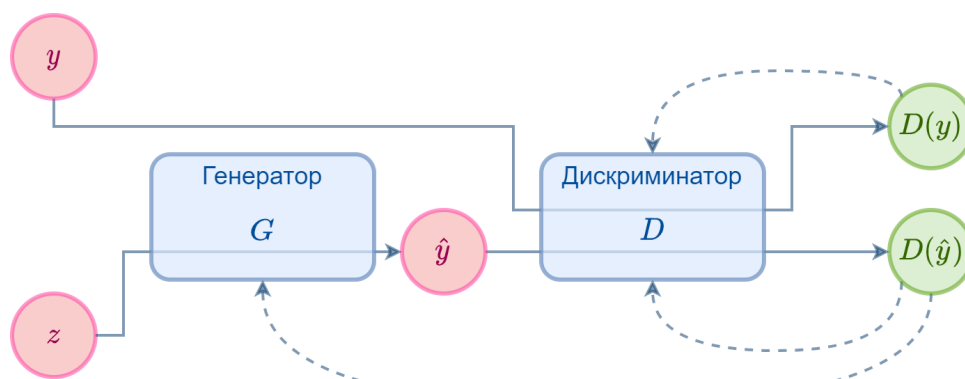


Рисунок 2 – Генеративно-сопоставительная сеть

Проблемы, связанные с возможными трудностями в процессе обучения GAN и сложностью оценки прогресса этого процесса, могут быть успешно решены модификациями, связанными с применением функции потерь Вассерштейна

$$-\frac{1}{n} \sum_{i=1}^n (y_i p_i). \quad (2)$$

где n – количество образцов, y_i – истинная метка образца, p_i – рассчитанная вероятность истинности образца.

Именно такое решение применено в модели, разработанной в рамках данной работы.

Определены в окончательном виде модули полученной модели, входные и выходные данные и функции потерь, которые будут применены в процессе обучения построенной нейронной сети.

В третьей главе на основе полученной модели построена нейронная сеть, которая затем обучена на подготовленных данных. Проведён анализ результатов, выявлены недостатки реализации и указаны пути их преодоления.

В качестве основы для реализации модели, решающей поставленную задачу, выбран наиболее развитый на сегодняшний день фреймворк для глубокого обучения *TensorFlow* версии 2, имеющей встроенную поддержку другого популярного фреймворка *Keras*, предоставляющего более высокоуровневые компоненты для построения нейронной сети. Этот выбор обусловил использование языка *Python* версии 3 как для работы с упомянутым фреймворком, так и для реализации вспомогательных модулей. В качестве формата организации кода и среды выполнения выбраны *Jupyter Notebook* и *Google Colab*, соответственно.

Подготовка данных произведена самостоятельно с помощью конвейера модулей, последовательно очищающих и преобразовывающих текст в массивы векторов вещественных чисел, представляющих данные в форме, пригодной

для использования в качестве входных данных нейронной сети.

Также создан модуль, преобразовывающий выходные данные сети обратно в обыкновенный текст. Сложность задачи поиска точек в многомерном пространстве потребовала применения специализированных структур и дополнительных оптимизаций.

В соответствии с разработанной в Главе 2 моделью построена нейронная сеть, реализованы функции потерь и логика потоков данных между модулями этой сети.

Полученная нейронная сеть обучена на подготовленных ранее данных, и получены хотя и недостаточно впечатляющие, но всё же интересные результаты.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. В данной работе была исследована применимость методов глубокого обучения к задаче переноса авторского стиля текста на русском языке, определено её место среди других задач обработки естественного языка. Было показано взаимосвязь данной задачи с такими направлениями искусственного интеллекта, как перенос стиля изображений и машинный перевод. Изучены существующие решения данной задачи с помощью нейронных сетей [1-А].

2. Исследованы архитектуры нейронных сетей и представлений данных, подходящих для решения поставленной задачи. Последовательно построена генеративно-сопоставительная модель нейронной сети, использующая модифицированный рекуррентный автокодировщик в качестве генератора, критик Вассерштейна в качестве дискриминатора и распределённые представления слов в качестве входных и выходных данных. Определены выражения для функций потерь каждого модуля нейронной сети.

3. Собраны и обработаны данные для домена-источника переноса стиля – тексты блога А. Лебедева – и данные для целевого домена – тексты дневников Л. Толстого. Построена и обучена реализация полученной модели нейронной сети с использованием фреймворка *TensorFlow 2*. Рассчитаны результаты работы сети на тестовых данных. Результаты проанализированы и сделан вывод, что полученная модель, с одной стороны, достигла в некоторой мере поставленной цели и обнаружила интересные побочные эффекты, а с другой – имеет недостатки, требующие её доработки. Следующими шагами может быть применение более сложных архитектурных решений, таких как механизм внимания.

Рекомендации по практическому использованию результатов

1. Полученные результаты, хотя и требуют доработки, составляют теоретическую и практическую основу для дальнейшего развития программных средств, предназначенных для решения задач переноса стиля текстов на русском языке. Также найденные решения могут быть применены для расширения функционала существующих решений в области обработки текста.

2. Разработанная модель нейронной сети для переноса стиля текста может быть обучена на более глобальных наборах данных и затем использоваться в системах автоматической модерации текстов, например, для замены сленговых выражений нейтральными аналогами, или же в системах автокоррекции текста для исправления речевых ошибок и неправильных согласований слов.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

Тезисы конференций

1-А. Евтушенко, Е.Ю. Обзор подходов к решению задачи переноса стиля текста с помощью нейронных сетей / Е.Ю. Евтушенко // Информационные технологии и системы 2019: материалы международной научной конференции. – Минск: БГУИР, 2019. – с. 304–305.