П

УДК 004.912

Я. С. Пикалёв
Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк
83048, г. Донецк, ул. Артема, 118-б

РАЗРАБОТКА АВТОМАТИЧЕСКОЙ СИСТЕМЫ ТРАНСФОРМАЦИИ АНГЛИЙСКИХ ВСТАВОК В РУССКИХ ТЕКСТАХ С ПРИМЕНЕНИЕМ ГЛУБОКОГО ОБУЧЕНИЯ

Ya. S. Pikalyov
Public institution «Institute of Problems of Artificial intelligence», c. Donetsk
83048, Donetsk, Artema str., 118-b

THE DEVELOPMENT OF THE AUTOMATIC TRANSFORMATION OF ENGLISH ACCENTS IN RUSSIAN TEXTS WITH THE APPLICATION OF DEEP LEARNING

Данная работа посвящена задаче разработки автоматической системы трансформации английских вставок в русских текстах с применением глубокого обучения. Автором предложен гибридный метод получения транскрипции, который был разработан, основываясь на работах лингвистов, а также с применением глубокого обучения. В данной работе изложен декларативно-процедурный подход, использующий как словарь, так и правила англо-русской практической транскрипции, для трансформации английских вставок, встречающихся в русских текстах. Подготовлен словарь англо-русской практической транскрипции с использованием механизма конечных автоматов. Обучена нейронная сеть для классификации языка текста с применением многослойных свёрточных нейронных сетей. Обучена нейронная сеть для трансформации слов на латинице, не найденных в словаре с использованием нейросетевой архитектуры типа энкодер-декодер.
Ключевые слова: обработка естественного языка; практическая транскрипция; механизм конечных автоматов; transformer; свёрточные нейронные сети.

This work is devoted to the task of developing an automatic system for transforming English inserts in Russian texts using deep learning. The author proposed a hybrid method of transcription, which was developed based on the work of linguists, as well as using deep learning. This paper presents a declarative-procedural approach, using both the dictionary and the rules of English-Russian practical transcription, for the transformation of English inserts found in Russian texts. The dictionary of English-Russian practical transcription using the finite state machine mechanism is prepared. Trained neural network for the classification of the language of the text using a multilayer convolutional neural networks. A neural network for the transformation of words in Latin not found in the dictionary using neural network architecture such as encoder-decoder has been trained.
Key words:natural language processing; practical transcription; finite state machine mechanism; transformer; convolutional neural networks.

Введение

На сегодняшний день в любом тексте на русском языке (статья, книга, новостная лента и т.п.) можно встретить большое количество вставок на латинице, чаще всего – на английском языке, которые, в основном, представлены названиями компаний и организаций («Apple», «Manchester United»), масс-медиа («Forbes»), географическими названиями («New York»), произведениями («Yesterday»), компьютерными программами («Microsoft Office»), интернет-сервисами («Amazon»), именами собственными («James 6Bond») и т.п. Реже встречаются цитаты или фразы на английском языке («To be or not to be»).

Данные вставки осложняют сбор данных для задач, связанных с обработкой естественного языка (например, классификация текста), а также с задачами распознавания и синтеза речи (сбор данных для формирования языковой модели; текстовая разметка аудиоданных). То есть для того, чтобы система синтеза/распознавания речи могла «работать» со словами, написанными на латинице, их необходимо трансформировать в ту же фонетическую систему, что и для русских слов. Для решения этой проблемы систему синтеза/распознавания речи, как правило, дополняют отдельным модулем для английского языка. Например, для системы распознавания речи необходим набор обучающих данных для дополнительного языка со своей транскрипцией и языковой моделью, а также модуль для классификации языка.

Таким образом, данная работа ориентирована на создание автоматической системы формирования практической транскрипции из слов, написанных на латинице, с использованием лингвистических знаний совместно с глубоким обучением, без необходимости в дополнительном языковом модуле. Универсальность данного метода заключается в том, что на выходе блока нормализации формируются «нормализованные» слова на кириллице, которые далее обрабатываются по тем же правилам, что и обычные слова русского языка. Благодаря этому одни и те же правила практической транскрипции можно включить в любую русскоязычную систему синтеза/ распознавания речи, независимо от фонетической транскрипции, которая в ней используется.

Характеристика существующих методов трансформации английских слов в кириллицу

1. Наиболее распространённым методом для трансформации английских слов в кириллицу является транслитерация, когда символу или набору символов из одного алфавита ставится в соответствие символ или набор символов из другого алфавита, причём соответствие осуществляется только по их графическому сходству. Недостаток данного подхода состоит в том, что восстанавливается исходное написание слова, но без учёта его произношения.

2. В качестве одного из методов автоматической транскрипции используют перевод, при котором некоторому часто встречающемуся имени ставится в соответствие его эквивалент, устоявшийся в языке, на который осуществляется перевод. Недостаток данного подхода состоит в том, что восстанавливается лишь семантическая информация, без учёта его произношения.

3. Другим способом является словарный метод, когда словам из языка N приводится в соответствие слова языка M при помощи некого словаря. Недостаток данного подхода состоит в том, что практически невозможно учесть все имеющиеся слова, а в случае формирования такого словаря данный алгоритм будет иметь высокую вычислительную сложность.

4. Наиболее оптимальным подходом является метод транскрипции, в котором звучание слова в языке N записывается средствами языка M. Однако язык M может быть знаком лишь узкому кругу специалистов, поэтому используют метод практической транскрипции [1], который генерирует запись транскрипции иноязычных слов с помощью орфоэпических норм языка N, используя только обычные знаки (буквы) этого языка без введения дополнительных знаков.

Проблемы методов трансформации английских слов в кириллицу

При трансформации английских слов в кириллицу возникают следующие проблемы:

– неполное соответствие фонемного состава двух языков;

– использование запрещённых в целевом языке морфем (ряд затруднений с обозначением звуков, отсутствующих в данном языке). Поэтому зачастую при транскрипции слов приходится ставить в примерное соответствие звукам одного языка звуки другого;

– частичная потеря информации как в виде строк, так и в виде фонетической информации (длительность, палатализованность, высота тона и др.);

– возможное отсутствие семантики для транслитерированного слова;

– отсутствие единого стандарта транскрипции и транслитерации. Правила для транскрипции на кириллицу либо ещё совсем не разработаны, либо разработаны, но вызывают много вопросов (т.е. даны лишь основные соответствия, а правильная передача многих буквосочетаний остается неясной);

– различные варианты произношения одного и того же слова, связанные с носителями языка или традициями. В подобных случаях при транскрипции возникают несколько потенциально возможных вариантов транскрипции, выбрать один из которых не представляется возможным;

– отсутствие взаимнооднозначного соответствия при транскрипции слова с исходного языка на язык перевода и обратно. То есть, если слово одного языка транскрибировать в другой язык, а затем транскрибировать его обратно, то полученное слово в значительном количестве случаев будет отличаться от исходного;

– транскрипции с одного языка (например, английского) имён собственных, исконно принадлежащих другому языку (например, испанскому);

– проблеме трансформации английских слов в кириллицу посвящены работы, в которых упор делается на использование фонетических правил [2-4] или на классическое машинное обучение [5], [6]. Стоит выделить тот факт, что данные работы, в основном, направлены на трансформацию имён собственных, написанных на английском языке.

Особенности произношения английских вставок русскоязычными дикторами

На основании работ [2-4] и [7-18] были выделены следующие ключевые особенности произношения английских вставок носителями русского языка.

1. Фонетическая модификация английских слов, заимствованных в русскую речь. Важным фактором является распространенность англоязычных слов в повседневной жизни носителей русского языка. Чем выше частотность слова, тем выше вероятность того, что оно будет озвучено диктором «по-русски» (для английских

звуков, отсутствующих в русском языке, подыскиваются ближайшие по звучанию звуки-замены). Например, для слов «bluetooth», «word», «amazon», «twitter», «microsoft», «facebook» и т.п. Однако даже менее распространенные слова («Whats App», «Slack» и др.), как правило, произносятся «по-русски». Так, например, в окончании «-ing», как правило, произносится русское [инк], т.е. происходит редукция («драйвинг» – «дра+йвинк»). То же можно сказать о фразе «I and you» («I & U» – «а+йэ+нт ю+»).

2. Существенным фактором является длина англоязычной вставки. Целые фразы на английском языке и просто длинные словосочетания редко употребляются в русскоязычных текстах. Тем не менее, длинные англоязычные фрагменты иногда озвучиваются русифицировано, особенно если данные словосочетания достаточно известны («Work&Travel», «Apple Watch», «Amazon Kindle Paperwhite»).

3. Целый ряд английских слов был заимствован в русский язык со смещением ударения на последний (или предпоследний) слог. Например, «email» (['i:meɪl]; «и+мэйл») – «имэ+йл»; «facebook» 'feɪsbʊk]; «фэ+йсбук») – «фэйсбу+к».

4. Английская фонема [w], которую лингвисты предлагают передавать через «в» перед буквой «у» и через «у» во всех остальных случаях («woods» – «ву+дс», «windows» – «уи+ндоус»). Однако, на практике, данная фонема произносится как «в». Например: «twitter» («туи+ттэр», «тви+ттэр»), «windows» («уи+ндовс», «ви+ндовс»)

5. Суффиксы «-er» и «-ed». По фонетическим правилам фонема [ə] передаётся транслитерацией «э». А в таких словах, как partner['pɑ:tnə]/па+ртнер/ гласный [е] смягчал предшествующий согласный [н] и по правилам ассимиляции смягчался и предшествующий зубной [т] ([па́ рт'н'эр]). На практике, при произнесении англоязычных вставок дикторы смягчают согласные по большей части так же, как и в русском языке: Christies– [кр'и́с't'ьс]; Acoustic– [ъку́с't'ьк]. Но для таких суффиксов, как -er («эр»), -ed («эд»), -ment («мэнт») смягчение обычно не происходит.

6. На практике, фонема [ð] передаётся звуком «з», а фонема [θ] через звуки «з» и «с». Пример Bluetooth – «блютуз».

7. Английские аббревиатуры, являющиеся словом или состоящие из более, чем одного слога, произносятся «по-русски». В противном случае – каждой букве аббревиатуры ставится в соответствие её транскрипция как отдельного звука («IBM» - «а+й би+ э+м»).

Стоит отметить, что большое значение имеет уровень владения английским языком говорящего. Но, на практике, говорящий лишь в редких случаях полностью переключается на английский язык вне зависимости от его уровня владения языком.

Таким образом, при озвучивании отдельных английских слов и не очень длинных словосочетаний актуально использовать практическую транскрипцию английского языка, из тех соображений, что носители русского языка привыкли озвучивать и воспринимать на слух английские слова в их русифицированной форме. Следовательно, транскрипция английских слов по правилам англо-русской практической транскрипции, может повысить точность распознавания речи. Дополнительно актуальность подтверждается тем, что в настоящее время в открытом доступе отсутствует большое количество данных для построения таких систем, а также готовые решения.

Постановка задачи

Английские слова, написанные в виде латиницы, должны быть преобразованы в формат кириллицы, используя практическую транслитерацию. В отличие от обычной

транслитерации, практическая основывается на произношении слов, что позволяет распознавать английские звуки, используя единую лингвистическую и акустическую базу. После того как английские слова будут преобразованы в последовательность символов кириллицы, на всех этапах их дальнейшей обработки они ничем не будут отличаться от русских. Данная методика может быть использована для любой системы синтеза или распознавания речи.

Одним из преимуществ предлагаемого способа обработки слов на латинице является его оптимальность: вместо написания дополнительного языкового модуля используется система англо-русской практической транскрипции для перевода английских слов на кириллицу. Кроме того, предлагаемый метод универсален, то есть может быть использован в любой русскоязычной системе синтеза/распознавания речи: после того, как вставка на латинице переведена в кириллическую графическую систему, она может обрабатываться на всех дальнейших этапах синтеза речи по тем же правилам, что и обычные (нормализованные) русские слова на кириллице.

Разрабатываемый подход является декларативно-процедурным, он использует как словарь, так и правила англо-русской практической транскрипции, изложенные в [7-18]. Для уменьшения комплексности используется нейросетевой (НС) подход и механизм конечных автоматов [19]. В качестве основных нейросетевых архитектур используется архитектура типа энкодер-декодер – Transformer [20], а также многослойной свёрточной нейронной сети (Convolional Neural Netwok, CNN) [21].

Предложенный метод получения практической транскрипции использует две нейросетевые модели для:

- определения языка текста, используя CNN (чтобы отсеять тексты на иноязычном тексте);
- трансформации графем в формат кириллицы, используя архитектуру Transformer.

Стоит отметить, что при сборе данных для обучения составляется словарь, использующий направленные ациклические графы слов (directed acyclic word graphs, DAWG) [22], основанные на конечных состояниях автоматов. В табл. 1 приведены примеры записей этого словаря.

Таблица 1 – Пример строк из словаря

Слово на латинице	«Нормализованное» слово
airways	эйрвэйс
bloomberg	блумберг
british	бритиш
microsoft	майкрософт

Описание общего алгоритма

Алгоритм построения практической англо-русской транскрипции, схема которого изображена на рис.1, состоит в следующем.

Рисунок 1 – Общая схема метода трансформации английских вставок

1. Считывается документ (T) после проведения нормализации, т.е. в документе отсутствуют цифро-буквенные комплексы (20th), сокращения, неконтекстные токены и т.п.

2. T разделяется на предложения ($S = \{s_i\}$, $i = 0,\ldots, N$, где N – общее количество предложений).

3. s_i разделяется на токены ($W = \{w_j\}$, $j = 0,\ldots, M$, где M – общее кол-во токенов в s_i), используя набор фильтров.

4. Используя модель для определения языка, определяем язык текста для s_i, если модель детектирует не русский язык (большее количество слов не из русского алфавита) считываем s_{i+1} и повторяем шаг 4 для этого предложения. То есть переходим к следующему шагу, если текущее предложение на русском языке.

5. Проверяем, содержит ли w_j символ не из русского алфавита, и добавляем токен в переменную w_cur. В случае, когда токен w_j представляет собой русско-английскую структуру (best-вещь, вещь-best), то w_j разделяется на токены, состоящие из символов одного алфавита без использования фильтров, применяемых для разделения предложения на токены. После разделения w_cur получаем массив E ($E = \{e_k\}$, $k = 2,\ldots,L$, где L – общее количество токенов, получившихся при разделении). Затем аналогично проверяем первый символ для e_k, если первый символ e_k не является символом для русского алфавита, то переходим к следующему шагу.

6. Если e_k не является аббревиатурой, то применяем модель трансформации токена e_k в формат кириллицы. В итоге получаем слово в формате кириллицы и заносим его в изменённый массив E. Соответственно, повторяем шаги 6 и 7 для всего массива E.

7. Объединяем массив E в строку через разделитель «-» получаем токен w_j в формате кириллицы.

8. Получаем массив транспонированных в кириллицу токенов $W_{ed} = \{w_j\}$, j = 0,…, M, которые объединяем и получаем транспонированное в кириллицу предложение s_i.

9. Получаем массив трансформированных в кириллицу предложений $S = \{s_i\}$, i = 0,…, N, которые объединяем и получаем транформированный в кириллицу документ T_{ed}.

Обучение модели определения языка текста

В качестве обучающего материала были использованы текстовые расшифровки для видеовыступлений из проекта Ted-Talks [23] для языков, содержащих символы кириллицы (белорусский (bel), болгарский (bol), киргизский, казахский, русский, сербский, таджикский, македонский, украинский, азербайджанский), а также для наиболее распространённых языков, содержащих символы латиницы (английский, немецкий, испанский, французский).

Для тестирования использовалось по 100 случайно выбранных предложений для каждого языка. Для обучения использовалось от 2 000 до 10 000 предложений для каждого языка.

Для обучения нейросети, предназначенной для распознавания языка текста предложения, необходимо получить векторное представление символа. Алгоритм получения векторного представления символа состоял в следующем:

- создать общий словарь встречаемых символов;
- обучить char2vec – нейросетевую модель для представления символа в виде вектора, размерностью 100, используя обучающий набор;
- используя обученную char2vec модель, векторизовать набор обучающих данных;
- используя набор векторизованных данных с соответствующим классом (языком), обучить НС для определения языка текста.

Архитектура модели классификации состоит из трёх слоевременной свёртки, размером 128×5,которые свёрнуты с входным слоем по одному пространственному (или временному) измерению, а также из трёх субдискретизирующих слоёв, размером 5 и 35 (последний слой). После этого выходные данные трансформируются в одномерный вектор и проходят два полносвязных слоя. Перед последним полносвязным слоём выполняется dropout-регуляризация. На всех слоях, кроме выходного полносвязного слоя, используется функция активации rectified linear unit, последний же слой использует softmax.

При обучении модели классификации были использованы следующие гипер-параметры:

- размер батча: 128 (данный датасет объёмный, поэтому данные делим на пакеты или батчи, что позволяет оптимизировать процесс обучения модели);
- размер векторных представлений для символов: 100;
- коэффициент скорости обучения: 0.01;
- кол-во эпох обучения: 50;
- коэффициент dropout-регуляризации: 0.5;
- функция регуляризации: L2-регуляризация;

– оптимизатор для градиентного спуска: Адам;
– функция потерь: перекрестная энтропия.

В качестве метрик использовался F1-score (рис. 2), а также стандартные метрики потерь и точности классификации (рис. 3).

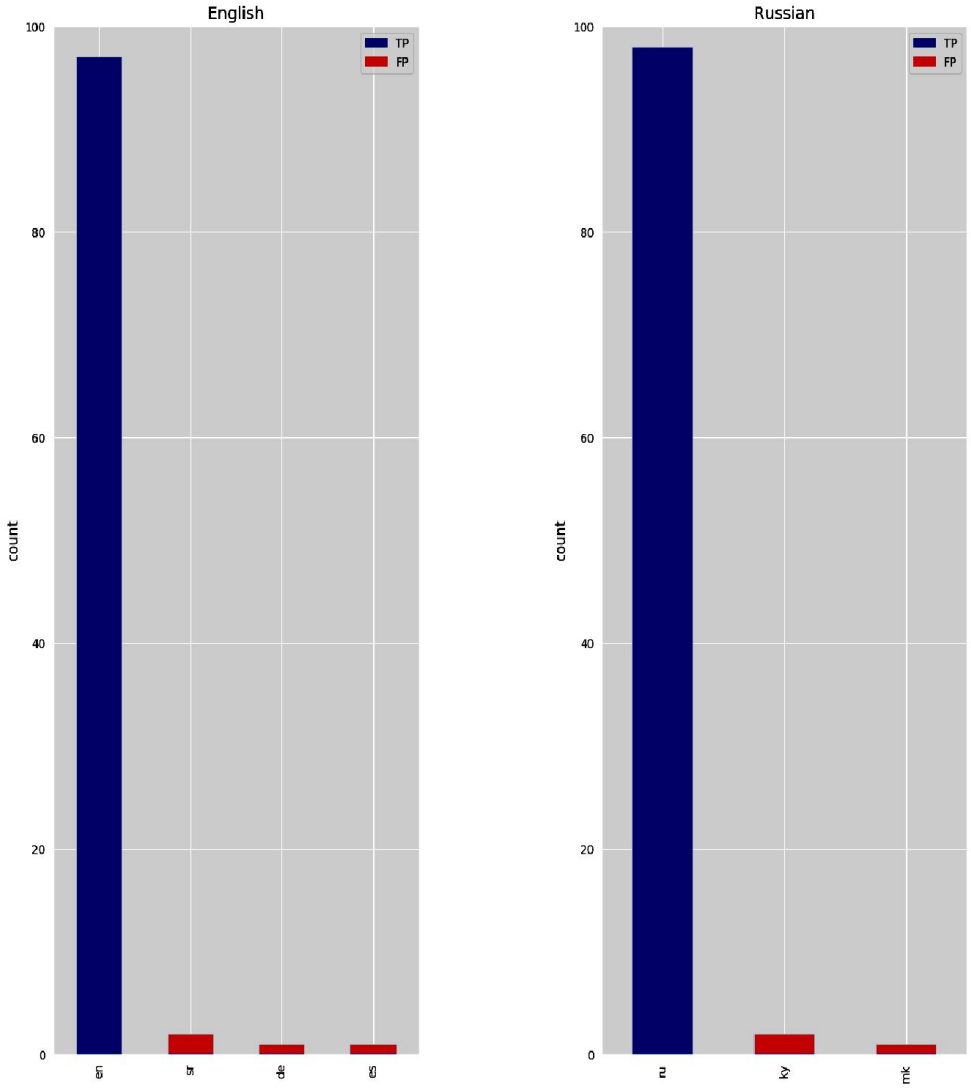

Рисунок 2 – Диаграмма F1-score для классификации английского и русского языков

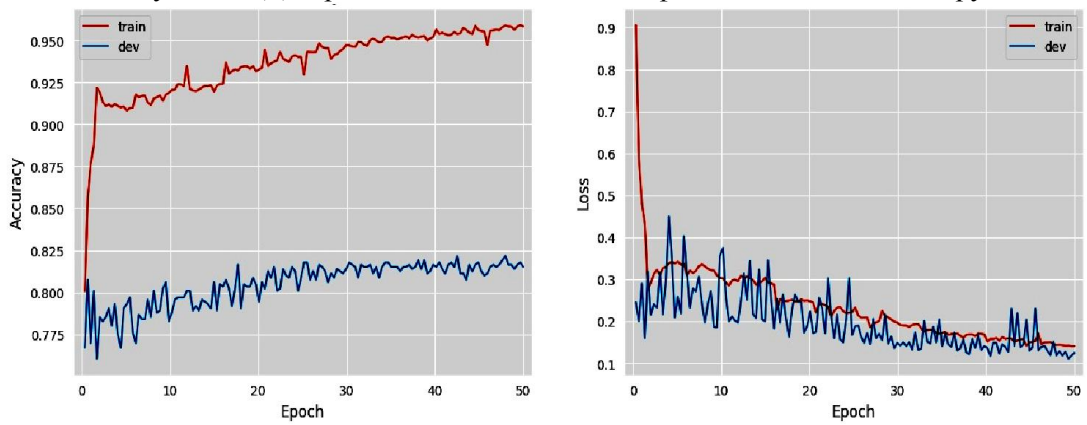

Рисунок 3– Графики точности (accuracy) и потерь (loss) для обучающей (train) и для тестовой (dev) выборок при обучении НС-модели классификации языка по тексту

Обучение модели трансформации в формат кириллицы

Подготовлен набор обучающих данных, собранный из информации, извлеченной из работ [11-18]. В общей сложности количество элементов обучающей выборки составило около 300 тыс. пар.

В качестве основной архитектуры для обучения использовалась архитектура Transformer. При обучении НС использовались следующие гиперпараметры:

- количество скрытых слоёв: 512;
- размер батча: 256;
- кол-во блоков в энкодере (энкодер переводит входной сигнал в более компактное представление, при этом сохраняя семантическую информацию): 5,
- количество блоков в декодере (восстанавливает исходный сигнал из компактного представления): 3;
- функция активации для скрытых слоёв: rectified linear unit;
- функция активации выходного слоя: softmax;
- функция потерь: разреженная кросс-энтропия;
- коэффициент dropout-регуляризации: 0.2;
- функция регуляризации: L2-регуляризация;
- оптимизатор для градиентного спуска: Адам;
- коэффициент скорости обучения: 0.0001;
- кол-во эпох: 100 тыс.

Для оценки результатов использовалась метрики: WER – отношение количества неверно трансформированных слов к общему количеству слов (Word Error Rate) и PER – отношение количества неверно трансформированных символов к общему количеству символов (Phoneme Error Rate). Графики результатов обучения приведены на рис. 4.

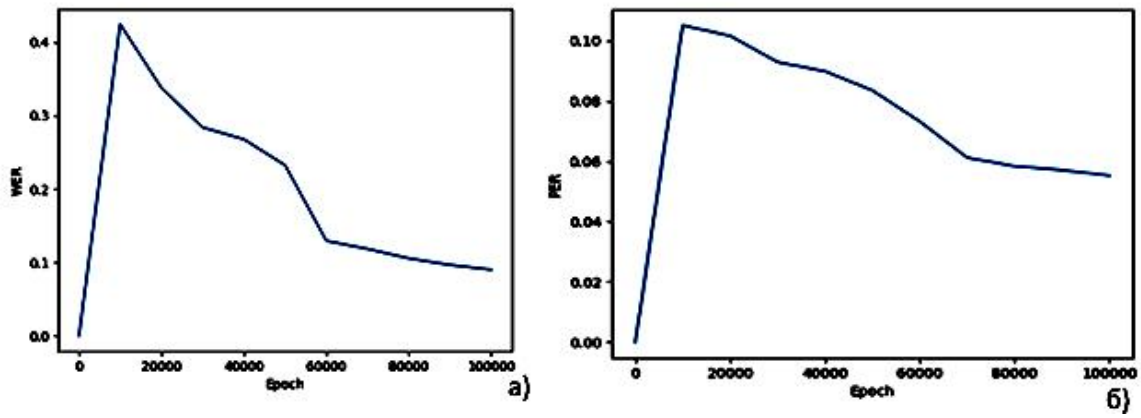

Рисунок 4– Графики WER (а) и PER (б) для НС-модели трансформации английских вставок в слова на кириллице

Выводы

Предложенный подход для создания автоматической системы формирования практической транскрипции слов английского языка позволяет получать англо-русскую практическую транскрипцию с точностью более 90% для слов и более 95% для символов. Такое высокое качество обеспечивает учет орфоэпических норм русского языка не только на основе фонетических правил, но и с использованием глубокого обучения.

Предложенная технология получения практической транскрипции слов английского языка может быть использована в системах синтеза/распознавания русской речи с целью адаптации англоязычных слов на этапе формирования «нормализованных» слов на кириллице для их дальнейшей обработки системой по правилам, применяемым для слов русского языка. Универсальность данного метода заключается в том, что на выходе блока нормализации формируются «нормализованные» слова на кириллице, которые далее обрабатываются по тем же правилам, что и обычные слова русского языка. Благодаря этому одни и те же правила практической транскрипции можно включить в любую русскоязычную систему синтеза/распознавания речи, независимо от фонетической транскрипции, которая в ней используется.

Список литературы

1. Гируцкий А. А. Введение в языкознание [Текст] /А. А. Гируцкий; [рецензенты: к.фил.н., доц. Е. С. Садовская, к.фил.н., доц. Ж. С. Спливеня]. – Минск : Вышэйшая школа, 2016. – 238 с.
2. Формальный метод транскрипции иностранных имен собственных на русский язык [Электронный ресурс] / А. В. Бондаренко, Ю. В. Визильтер, В. И. Горемычкин, Э. С. Клышинский // Программные продукты и системы. – 2010. – № 1. – URL: https://cyberleninka.ru/article/n/formalnyy-metod-transkriptsii-inostrannyh-imen-sobstvennyh-na-russkiy-yazyk (дата обращения: 19.02.2019).
3. Черепанова О. Д. Озвучивание англоязычных употреблений в системе русскоязычного синтеза речи «ТЕКСТ-РЕЧЬ» с помощью практической транскрипции [Текст] / О. Д. Черепанова // Проблемы компьютерной лингвистики и типологии: сборник. – 2016. – С. 226.
4. Черепанова О. Д. Лингвистическое обеспечение речевых технологий: использование англо-русской практической транскрипции в системе русскоязычного синтеза «ТЕКСТ-РЕЧЬ» [Текст] / О. Д. Черепанова // Вестник Московского университета. – Серия 9: Филология. – 2017. – № 3. – С. 156–167.
5. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Rcogniiton [Текст] / L.R. Rabiner // Proceedings of the IEEE.–1989. – № 77. – P. 257–286.
6. Hermjakob U. Name translation in statistical machine translation-learning when to transliterate [Текст] / Hermjakob U., Knight K., Daumé III H // Proceedings of ACL-08: HLT. – 2008. – P. 389–397.
7. Речевые технологии в задаче обучения студентов-носителей русского языка произношению на иностранном языке [Текст] / Мещеряков Р. В., Тиунов С. Д., Лирмак Ю. М., Шевкунова А. Е. // «Анализ разговорной русской речи» (АРЗ-2011): Труды Пятого междисциплинарного семинара. – СПб. : ГУАП. – 2011. – С. 78–82.
8. Успенский В. А. Труды по нематематике [Текст] / В. А.Успенский.– М. : ОГИ,2002.– С. 390–412.
9. Зиндер Л.Р. Общая фонетика [Текст] / Л.Р. Зиндер. – М. : Высш. шк., 1979. – 309 с.
10. Буркова С. С. Англицизмы в современном русском интернет-языке [Электронный ресурс] / С. С. Буркова, А. И. Дергабузов // Актуальные проблемы филологии: материалы III Междунар. науч. конф. (г. Казань, май 2018 г.). – Казань: Молодой ученый, 2018. – С. 11-13. – URL : https://moluch.ru/conf/phil/archive/301/14116/ (дата обращения: 20.02.2019).
11. Суперанская А. В. Теоретические основы практической транскрипции [Текст] / А. В. Суперанская. – М.: Наука, 1978. – 283 с.
12. Гитляревский Р. С. Иностранные имена и названия в русском тексте: справочник [Текст] / Р. С. Гитляревский, Б. А. Старостин. – 3-е изд. – М. : Высш. шк., 1985. – 304 с.
13. Ермолович Д. И. Имена собственные на стыке языков и культур [Текст] / Д. И. Ермолович. – М.: Р. Валент. – 2001. – Т. 200. – С. 3.
14. Ермолович Д. И. Методика межъязыковой передачи имен собственных [Текст] / Д. И. Ермолович. – М., 2009.
15. Рыбакин А.И. Словарь английских фамилий [Текст] / А. И. Рыбакин. – М.: Астрель: АСТ, 2000. – 576 с.
16. Лидин Р. А. Иностранные фамилии и личные имена: Практика транскрипции на русский язык: Справочник [Текст] / Р. А. Лидин. – М. : ООО «Издательство Толмач», 2006. – 480 с.
17. Казакова Т. А. Практические основы перевода. English<=>Russian. Серия: Изучаем иностранные языки [Текст] / Т. А. Казакова. – СПб.: «Издательство Союз», 2001. – 320 с.
18. Таранов А. М. Кириллическая транслитерация [Текст] / А. М.Таранов. – М.: T&P Books Publishing, 2013. – 256 с.
19. Mohri M. Systems and methods for generating weighted finite-state automata representing grammars : пат. 7181386 США. [Текст] / M. Mohri, M. J. Nederhof. – 2007.

20. Vaswani A. et al. Tensor2tensor for neural machine translation [Текст] / A. Vaswani et al. // arXiv preprint arXiv:1803.07416. – 2018. https://arxiv.org/pdf/1803.07416.pdf
21. Kim Y. Convolutional neural networks for sentence classification [Текст] / Y. Kim // arXiv preprint arXiv:1408.5882. – 2014.
22. Crochemore M. Direct construction of compact directed acyclic word graphs [Текст] / M. Crochemore, R. Vérin // Annual Symposium on Combinatorial Pattern Matching. – Springer, Berlin, Heidelberg, 1997. – P. 116–129.
23. TED Talks [Электронный ресурс]. – Режим доступа : https://www.ted.com/talks/ – (Дата обращения: 25.05.2019).
24. Харламов А. А. Анализ текстов: лингвистика, семантика, прагматика в рамках когнитивного подхода[Текст] / А. А. Харламов, Т. В. Ермоленко // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2015. – № 0(1). – С. 107–115.

References

1. Girutskyi, A. A. *Vvedeinie v yazykoznanie* [Introduction to linguistics] / [retsenzenty: k..phil.n., dots. E. S. Sadovskaya, k.phil.n., dots. Zh. S. Splivenya], Minsk, Vysheihsaya shkola, 2016, 238 c., ISBN 978-985-06-2720-9.
2. Bondarenko A. V., Vizilter Yu. V., Goremychkin V. I., Klyshinskiy E. S. Phormalnyi metod transkriptsii inostrannyh imyon sobstvennyh na russkiy yazyk [Formal method of transcription of foreign proper names into Russian].*Programmnye produkty i sistemy* [Software products and systems], 2010, No1, URL: https://cyberleninka.ru/article/n/formalnyy-metod-transkriptsii-inostrannyh-imen-sobstvennyh-na-russkiy-yazyk (appeal: 19.02.2019).
3. CherepanovaO. D. Ozvuchivanie angloyazychnyh upotrebleniy v sisteme russkoyazychnogo sinteza rechi «TEXT-RECHJ» s pomoschju prakticheskoy transkriptsii [Scoring of English-language uses in the system of Russian-language speech synthesis "TEXT-SPEECH" with the help of practical transcription]. *Problemy kompjuterno ylingvistikii tipologii* [Problems of computational linguistics and typology: collection], sbornik, 2016, pp. 226.
4. CherepanovaO. D. Lingvisticheskoe obespechenie rechevyh tehnologiy: ispolzovanie anglo-russkoy prakticheskoy transkriptsii v sisteme russkoyazychnogo sinteza "TEXT-RECHJ″ [Linguistic support of speech technologies: the use of English-Russian practical transcription in the system of Russian-language synthesis "TEXT-SPEECH"]. *Vestnik Moskovskogo universiteta* [Bulletin of Moscow University]. Seriya 9: Philologiya, 2017, No 3, pp. 156-167.
5. Rabiner L.R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Rcogniiton. *Proceedings of the IEEE*.1989. No 77, pp. 257-286.
6. Hermjakob U., Knight K., Daumé III H. Name translation in statistical machine translation-learning when to transliterate. *Proceedings of ACL-08: HLT*,2008,pp. 389-397.
7. Mescheryakov R. V., TiunovS. D., LirmakYu. M., Shevkunova A. Ye. (2011) Rechevye Tehnologii v zadache obucheniya studentov-nositeley russkogo yazyka proiznosheniyu na inostrannom yazyke [Speech technologies in the problem of teaching students-native speakers of Russian pronunciation in a foreign language]. *«Analiz razgovornoy russkoy rechi»* [Analysis of spoken Russian speech] (ARZ-2011): Trudy pyatogo mezhdistsiplinarnogo seminara, SPb.: GUAP, pp. 78–82.
8. UspenskiyV. A. (2002).*Trudy po nematematike* [Works on remotemachine] // M., OGI, S. 390–412.
9. Zinder L.R. *Obshaya phonetika* [General phonetics]. M., Vysh. shk., 1979, 309 s.
10. Burkova S. S., Dergabuzov A. I. Anglitsizmy v sovremennom russkom internet-yazyke [Anglicisms in the modern Russian Internet language]. *Aktualnye problem philologii* [Actual problems of Philology]: materialy III Mezhdunar. nauch. konf. (g. Kazanj, maj 2018 g.), Kazanj, Molodoy uchjonyi, 2018, pp. 11-13. URLhttps://moluch.ru/conf/phil/archive/301/14116/ (appeal: 20.02.2019).
11. Superanskaya. A.V. *Teoreticheskie osnov y prakticheskoy transkriprsii* [Theoretical basis of practical transcription], M., Nauka, 1978, 283 s.
12. Gitlyarevskyi R. S., Starostin B.A. *Inostrannye imena i nazvaniya v russkom texte* [Foreign names and titles in the Russian text], spravochnik, 3-eizd., M., Vyssh. shk., 1985, 304 p.
13. ErmolovichD. I. *Imena sobstvennye na styke yazykovi kultur* [Proper names at the crossroads of languages and cultures], M., R. Valent, 2001, T. 200, p. 3.
14. ErmolovichD. I. *Metodika mejyazykovoy peredachi imyon sobstvennyh* [The technique of cross-language transfer of proper names],M., 2009.
15. RybakinA.I. *Slovarj angliyskih familiy* [Dictionary of English surnames], M., Astrelj, AST, 2000, 576 s.
16. Lidin R. A. *Inostrannye familii i lichnye imena: Praktika transkriptsii na russkiy yazyk* [Foreign names and personal names: the Practice of transcription for Russian language]: Spravochnik,M., OOO «Izdatelstvo Tolmach», 2006, 480 s.

17. Kazakova T. A. *Prakticheskie osnovy perevoda. English <=> Russian. Seriya: Izuchaem inostrannye yazyki* [Practical basics of translation. English <=> Russian. Series: Learning foreign languages]. SpB.: «Izdatelstvo Soyuz», 2001, 320 s.
18. Taranov A.M. *Kirillicheskaya transliteratsiya* [Cyrillic transliteration], M., T&P Books Publishing, 2013, 256 s.
19. Mohri M., Nederhof M. J. *Systems and methods for generating weighted finite-state automata representing grammars* : pat. 7181386 США, 2007.
20. Vaswani A. et al. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*, 2018, https://arxiv.org/pdf/1803.07416.pdf
21. Kim Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
22. Crochemore M., Vérin R. Direct construction of compact directed acyclic word graphs. *Annual Symposium on Combinatorial Pattern Matching*, Springer, Berlin, Heidelberg, 1997, pp. 116-129.
23. TED Talks [Electronicresource], Access mode: https://www.ted.com/talks/ (appeal: 25.05.2019).
24. Kharlamov A. A., Yermolenko T. V. Analiz tekstov: lingvistika, semantika, pragmatika v ramkakh kognitivnogo podkhoda [Text analysis: linguistics, semantics, pragmatics within the framework of the cognitive approach]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], Donetsk, GU «IPII», 2015, No 0(1), pp. 107–115.

RESUME

Ya. S. Pikalyov

The Development of the Automatic Transformation of English Accents in Russian Texts with The Application of Deep Learning

The problem considered in this paper relates to one of the main tasks of natural language processing text normalization, namely the processing of words and phrases in Latin, found in Russian texts. Development of the optimal (with an accuracy of more than 90%) of the system of normalization of Russian-language text is one of the key problems areas of natural language processing (natural language processing, NLP). The relevance of this problem is confirmed by the widespread introduction of such systems in most types of software products (search engines, recommendation systems, etc.), as well as the use of researchers and developers for other tasks of NLP and artificial intelligence. Currently, these systems are implemented using phonetic rules or machine learning. In addition, these works are mainly aimed at the transformation of proper names written in English. In this regard, the actual task is to develop an automatic system of transformation of English inserts in Russian texts using deep learning.

This article uses the following methods: classification methods, methods of tokenization of the text, the method of evaluation in terms of the error of recognition of matches; was used the author's dictionary containing more than 5 million words; was also used Python programming language for software implementation.

The algorithm of practical English-Russian transcription construction is formed; training data sets for the problems of language classification in text arrays are collected, as well as for the problem of transformation of text inserts in Latin into Cyrillic format; a model of text representations of symbols is obtained; a model of determining the language of the text is obtained; a model of transformation of inserts in Latin into Cyrillic format is obtained.

The proposed approach to create an automatic system of formation of practical transcription of words in English allows to obtain English-Russian practical transcription with an accuracy of more than 90% for words and more than 95% for symbols. Such high quality ensures the consideration of orthoepic norms of the Russian language not only on the basis of phonetic rules, but also with the use of deep learning.

The proposed technology of obtaining the practical transcription of words in the English language can be use in the synthesis of Russian speech recognition and adaptation of English words in the stage of formation of "standard" words in Cyrillic for further processing by the system of rules used for words in Russian.

РЕЗЮМЕ

Я. С. Пикалёв

Разработка автоматической системы трансформации английских вставок в русских текстах с применением глубокого обучения

Проблема, рассмотренная в данной работе, относится к одной из основных задач естественной обработки языка нормализации текста, а именно обработка слов и словосочетаний на латинице, встречающихся в русских текстах. Разработка оптимальной (с точностью более 90%) системы нормализации русскоязычного текста является одной из ключевых проблем направления обработки естественного языка (natural language processing, NLP). Актуальность данной проблемы подтверждается широким внедрением подобных систем в большинство видов программных продуктов (поисковые системы, системы рекомендаций и т.п.), а также использованием исследователями и разработчиками для других задач NLP и искусственного интеллекта. На текущий момент данные системы реализованы при помощи фонетических правил или с применением машинного обучения. К тому же данные работы, в основном, направлены на трансформацию имён собственных, написанных на английском языке. В связи с этим актуальной задачей является разработка автоматической системы трансформации английских вставок в русских текстах с применением глубокого обучения.

В данной статье использованы следующие методы: методы классификации, методы токенизации текста, метод оценки по показателю ошибки распознавания соответствий; был использован авторский словарь, содержащий более 5 млн слов; для программной реализации был использован язык программирования Python.

Сформирован алгоритм построения практической англо-русской транскрипции; собраны обучающие наборы данных для задач классификации языка в текстовых массивах, а также для задачи трансформации текстовых вставок на латинице в формат кириллицы; получена модель текстовых представлений символов; получена модель определения языка текста; получена модель трансформации вставок на латинице в формат кириллицы.

Предложенный подход для создания автоматической системы формирования практической транскрипции слов английского языка позволяет получать англо-русскую практическую транскрипцию с точностью более 90% для слов и более 95% для символов. Такое высокое качество обеспечивает учет орфоэпических норм русского языка не только на основе фонетических правил, но и с использованием глубокого обучения.

Предложенная технология получения практической транскрипции слов английского языка может быть использована в системах синтеза/распознавания русской речи с целью адаптации англоязычных слов на этапе формирования «нормализованных» слов на кириллице для их дальнейшей обработки системой по правилам, применяемым для слов русского языка.

Статья поступила в редакцию 01.03.2019.