ЯЗЫКОЗНАНИЕ

УДК 81’32:811.163.1

Гудков Вадим Вадимович

Санкт-Петербургский государственный университет,
Россия, 199034, Санкт-Петербург, Университетская наб., 7–9
st071220@student.spbu.ru

Митренина Ольга Владимировна

Санкт-Петербургский государственный университет,
Россия, 199034, Санкт-Петербург, Университетская наб., 7–9
o.mitrenina@gmail.com

Соколов Евгений Геннадьевич

Институт лингвистических исследований Российской академии наук,
Россия, 199053, Санкт-Петербург, Тучков пер., 9
pan_liwerij@mail.ru

Коваль Ангелина Александровна

Санкт-Петербургский государственный университет,
Россия, 199034, Санкт-Петербург, Университетская наб., 7–9
st069276@student.spbu.ru

Языковой перенос нейросетевого обучения для частеречной разметки Санкт-Петербургского корпуса агиографических текстов (СКАТ)

Для цитирования: Гудков В. В., Митренина О. В., Соколов Е. Г., Коваль А. А. Языковой перенос нейросетевого обучения для частеречной разметки Санкт-Петербургского корпуса агиографических текстов (СКАТ). *Вестник Санкт-Петербургского университета. Язык и литература.* 2023, 20 (2): 268–282. https://doi.org/10.21638/spbu09.2023.205

В статье рассказывается об эксперименте по обучению морфологического анализатора на основе искусственных нейронных сетей для разметки Санкт-Петербургского корпуса агиографических текстов (СКАТ), который разрабатывается на кафедре математической лингвистики СПбГУ. Корпус содержит тексты 23 рукописей XV–XVIII вв.

объемом около 190 тыс. словоупотреблений, четыре из которых размечены вручную. Для создания автоматического частеречного анализатора использовались модели bi-LSTM, дистиллированная RuBERT-tiny2 и модель RuBERT. Все они были обучены на корпусах текстов на современном русском языке и перенастроены для разметки древнерусских текстов с помощью языкового переноса. Для дообучения языковых моделей на основе архитектуры трансформера необходимо было сформировать свой токенизатор на основе техники byte pair encoding и соотнести токены из оригинального русскоязычного токенизатора и нового на основе индексов. Затем модель дообучалась на задачу классификации токенов. Для настройки модели использовался размеченный подкорпус из трех житий объемом 35 603 токена, 2885 предложений. В эксперименте учитывалась только разметка с указанием части речи, классификация проводилась по 17 тегам, 13 из которых соответствовали частям речи, а оставшиеся четыре отмечали знаки препинания. Для оценки качества модели использовались стандартные метрики F1 и Accuracy. Согласно автоматическим метрикам оценки наилучший результат показала модель RuBERT. С помощью нее была проведена частеречная разметка «Жития Александра Свирского», ошибки разметки были проанализированы вручную. Большинство ошибок были связаны с неверным обобщением закономерностей линейного положения или со сходством словоформ как в крайней левой, так и в крайней правой позиции.

Ключевые слова: агиография, корпус древнерусских текстов, нейросетевая разметка, языковой перенос нейросетевого обучения, частеречная разметка.

Проблема автоматической разметки корпусов древних текстов

По самой своей сути историческая лингвистика — дисциплина корпусная [Haug 2015: 185]. Не имея доступа к живым носителям, историк языка невольно работает с коллекцией текстов. Появление компьютеров позволило заменить собрания текстов на материальных носителях — глине, камне, папирусе, пергамене или бумаге — электронными корпусами, что сильно упростило работу исследователя. Особенностью корпуса является многоуровневая система помет (тегов), описывающих лексические, грамматические и прочие характеристики слов и других компонентов текста [Захаров 2017]. Аннотированный (или размеченный) корпус может быть полезен различным ученым: если лингвист найдет в нем удобный инструмент для извлечения языковых данных и статистической проверки гипотез, то филолога и историка может заинтересовать поиск текстовых заимствований или культурных реалий [Sokolov 2019: 338]. Так, специалистам по истории языка уже доступна построенная А. Е. Поляковым на основе данных корпуса церковнославянских текстов Национального корпуса русского языка (НКРЯ) [Поляков 2014: 245] эмпирическая модель словоизменения церковнославянского языка [Поляков]. Исследователь агиографической традиции может благодаря особой разметке корпусов сравнивать объем и расположение библейских и святоотеческих цитат в определенных группах житийных текстов [Алексеева и др. 2022], текстолог благодаря компьютерным средствам может эффективно находить разночтения в сотнях списков [Алексеева, Миронова 2017: 265–271], историк — сопоставлять описания событий на широком массиве источников.

Ручная разметка корпусов (расстановка тегов) — это трудоемкая операция, требующая долгого и однообразного труда квалифицированных лингвистов. Поэтому большой популярностью стали пользоваться автоматические способы раз-

метки корпусов. В частности, в Университете Фрайбурга был создан автоматический лемматизатор для средневековых церковнославянских текстов [Podtergera et al. 2016: 88–90], а для упомянутого выше церковнославянского корпуса НКРЯ применялась комбинация машинного и ручного грамматического анализа [Поляков 2014: 252–253].

Однако существующие инструменты грамматического разбора нередко оказываются малоэффективны при разметке древнерусских и церковнославянских памятников. Причина этого кроется в отсутствии для таких текстов единого графико-орфографического стандарта, многочисленных сокращенных и титлованных написаниях, а также значительной вариативности словоизменительных морфем [Podtergera et al. 2016: 68–82].

В последние годы бурное развитие получили технологии, основанные на нейросетевых моделях, в том числе в сфере морфологической разметки текстов на естественном языке [Dereza et al. 2016; NMT]. Свойственная им эффективность позволяет надеяться на успех в применении их и к древним рукописным памятникам.

В нашей статье на примере Санкт-Петербургского корпуса агиографических текстов (СКАТ) рассматриваются перспективы автоматизации процесса морфологической разметки электронных корпусов на основе обучения искусственных нейронных сетей.

Состав и особенности СКАТ

СКАТ — это развивающийся проект, реализуемый с конца 1970-х гг. усилиями сотрудников и студентов кафедры математической лингвистики Санкт-Петербургского государственного университета и ставший предметом в том числе и международного сотрудничества. Работа над корпусом в последние годы ведется в кооперации с лионской лабораторией IHRIM, имеющей богатый опыт работы с размеченным корпусом старофранцузского языка [Azarova et al. 2021], участники проекта выступают на крупных конференциях, таких как El'Manuscript, Interdisciplinary Perspectives on Data: 2nd International Conference of the European Association for Digital Humanities (EADH) и XLIX Международная научная филологическая конференция им. Л. А. Вербицкой.

Существование проекта СКАТ делает СПбГУ одним из немногих вузов мира, обладающих собственным историческим корпусом славянских языков [Mitrenina 2014], что подчеркивает высокий научный уровень Петербургского университета.

СКАТ включает тексты 23 рукописей XV–XVIII вв. объемом около 190 тыс. словоупотреблений. В основном это жития севернорусских святых, основателей монастырей Русского Севера на территории современных Вологодской, Архангельской и Ярославской областей. Четыре жития корпуса (50 тыс. словоупотреблений) вручную снабжены полной морфологической разметкой, которая была проведена силами студентов кафедры математической лингвистики под руководством Е. Л. Алексеевой. Эти тексты представлены в формате XML в соответствии с рекомендациями Text Encoding Initiative (TEI); в 2021 г. корпус был реализован на платформе TXM [Azarova et al. 2021]. Размеченные жития используются в качестве обучающей выборки для ряда экспериментов по автоматизации процесса морфологического аннотирования.

Структура морфологической разметки СКАТ

В морфологически размеченной части корпуса для токена могут быть указаны следующие шесть параметров: начальная форма слова («lemma»), грамматические характеристики («msd»), часть речи («pos»), словоформа, записанная с помощью символов кириллицы («reg»), словоформа, записанная с помощью специально разработанного шрифта («scr»), а также код, несущий информацию о тексте, которому принадлежит токен, и его порядковом номере в этом тексте («xml:id»). Ниже приведен пример разметки текста:

(1)
```
<w lemma="и" pos="союз" reg="и" src="И" xml:id="DmPrlc.188">и</w>
<w lemma="наказатель" msd="jo;дат;мн;м" pos="сущ" reg="наказателемъ"
src="НАКАЗАТЕЛЕМЪ" xml:id="DmPrlc.189">наказателемъ</w>
<lb n="10"/>
<w lemma="иже" msd="м;им;мн;м" pos="мест" reg="иже" src="ИЖЕ"
xml:id="DmPrlc.190">иже</w>
<w lemma="въ" pos="пред" reg="въ" src="ВЪ" xml:id="DmPrlc.191">въ</w>
<w lemma="правда" msd="а;мест;ед;ж" pos="сущ" reg="правд+" src="ПРАВД+"
xml:id="DmPrlc.192">правдѣ</w>
<w msd="м;наст;дат;мн;м" pos="прич" reg="св+дущимъ" src="СВ+ДЏЩИМЪ"
xml:id="DmPrlc.193">свѣдоущимъ</w>
<pc force="weak" xml:id="DmPrlc.194">.</pc>
<w lemma="паче" pos="нар" reg="паче" src="ПА&amp;ЧЕ" xml:id="DmPrlc.195">
па
<lb n="11"/>
че
</w>
<w lemma="же" pos="част" reg="же" src="ЖЕ" xml:id="DmPrlc.196">же</w>
<w lemma="тои" msd="тв;дат;мн;м" pos="мест" reg="т+мъ" src="Т+МЪ"
xml:id="DmPrlc.197">тѣмъ</w>
<pc force="weak" xml:id="DmPrlc.198">.</pc>
```

Набор грамматических характеристик в параметре «msd» зависит от части речи, к которой принадлежит слово. Для существительных, прилагательных и числительных указываются тип склонения, падеж, число и род. Например, для слова *рвенїю* тег будет выглядеть следующим образом: «jo;дат;ед;ср».

Для местоимений в зависимости от разряда могут указываться тип склонения, тип местоимения (1-е, 2-е лицо или возвратность), падеж, число и род; при этом сам разряд явным образом не называется. Различия в составе тегов можно увидеть при сравнении характеристик личного местоимения *мы* (личн;1;им;мн), возвратного местоимения *себе* (личн;возвр;вин/род) и неличного местоимения *сами* (тв;им;мн;м).

Грамматический тег для причастий состоит из пяти компонентов: типа склонения, времени, падежа, числа и рода (ср. *помышлѧющи* — «jo;наст;им;мн;м»). Для глагола указываются наклонение, лицо, число, а также время (для изъявительного наклонения), род (для прошедшего времени), класс (для настоящего/будущего времени и повелительного наклонения). Отметим, что такая характеристика, как возвратность, отображается не в теге, а в части речи («прич/в», «гл/в»).

Для слов других частей речи параметр «msd» не указывается.

Знаки препинания также размечаются, однако для них указываются только две характеристики: «force» (длительность паузы) и «xml:id» (код). Кроме того, в корпусе размечены номера страниц и строк рукописи (параметры «n»).

Тег «src» используется для сохранения написания графических особенностей текста в формате plain text. Так, например, для кодировки буквы *ѣ* (ять) используется знак «+». Обучение модели проводилось на основании тегов «src», однако для удобства чтения примеры в статье будут указаны в оригинальном написании.

У каждого памятника имеется также экстралингвистическая разметка, включающая в себя сведения о его печатном издании, информацию о названии, датировке и библиотечном шифре взятой за основу издания рукописи, а также о лицах, ответственных за редактирование текста и его конвертацию в формат XML-TEI.

В нашем эксперименте учитывалась только разметка с указанием части речи (тег «pos»).

Вопрос о применимости к анализу церковнославянских текстов понятия «предложение» и конкретном его определении и наполнении имеет долгую и сложную историю [Николенкова 2000: 38 и далее]. Поскольку наша статья посвящена не синтаксической, а морфологической разметке житийного корпуса, мы предпочтем не углубляться в проблему выделения минимальных и максимальных синтаксических единиц членения церковнославянских памятников, принимая за предложение отрезок текста любой степени синтаксической сложности и связности, ограниченный с обеих сторон точками.

Искусственные нейронные сети в лингвистических исследованиях

В современной науке о языке для исследовательских и практических целей активно применяются нейросетевые модели. В основе искусственной нейросети лежит идея модели нейрона [Jurafsky, Martin 2020: 137], предложенная в 1943 г. У. Мак-Каллоком и У. Питтсом [McCulloch, Pitts 1943]. Искусственный нейрон, или узел нейросети, принимая на вход некоторый набор вещественных чисел, по определенной закономерности их обрабатывает и возвращает результат обработки [Jurafsky, Martin 2020: 138]. С помощью так называемой функции активации этот результат преобразуется в выходной сигнал, принадлежащий обычно отрезку от –1 (или 0) до 1 [Jurafsky, Martin 2020: 138; Букия, Протопопова 2016: 131–132]. Узлы нейросети располагаются слоями. Слои должны быть трех типов: по одному входному и выходному и один или несколько промежуточных, скрытых [Васильев 2021: 33]. В каждом последующем слое всякий нейрон получает на вход выходной сигнал всех нейронов предшествующего слоя [Jurafsky, Martin 2020: 147], причем основные преобразования происходят в скрытых слоях. В выходном слое, как правило, бывает столько узлов, сколько предусмотрено решений для стоящей перед нейросетью задачи [Васильев 2021: 34]. К примеру, в нейросетевом морфологическом анализаторе каждый узел выходного слоя будет соответствовать отдельной части речи или варианту части речи с дополнительными морфологическими показателями (как это будет показано далее в статье).

Применяемые в настоящее время нейронные сети обычно имеют большое количество скрытых слоев, из-за чего называются глубокими. Использование та-

кой нейросети, соответственно, называется глубоким обучением [Jurafsky, Martin 2020: 137].

В нейронных сетях применяется так называемый эмбеддинг (от англ. *embedding* ‘вложение’) — установление соответствия между собственно лексемами и векторами вещественных чисел. Близкие по употреблению единицы имеют схожее векторное представление, что позволяет нейросети предсказывать их поведение в высказывании и свойства, даже если одна из этих единиц еще не встречалась в данной конкретной речевой цепочке [Jurafsky, Martin 2020: 143].

Искусственная нейронная сеть не может работать с текстами напрямую, она может обрабатывать только числа. Искусственный нейрон принимает на вход набор вещественных чисел, поэтому данные перед обработкой нейросетью должны быть переведены в числовой вид.

Нейросетевая модель BERT и ее дообучение

Для проведения морфологической разметки корпуса СКАТ мы использовали нейросетевую языковую модель BERT (англ. Bidirectional Encoder Representations from Transformers), обученную на русскоязычном корпусе и перенастроенную для разметки древнерусских текстов[1]. Технология перенастройки модели описана в работе [Kuratov, Arkhipov 2019].

Тип моделей BERT появился в 2018 г. Это двунаправленная языковая модель, которую можно дообучить для большинства известных задач обработки естественного языка, таких как машинный перевод, морфологическая разметка, извлечение именованных сущностей и др.

Модель BERT обучается на задаче предсказания слова по контексту. Первые модели, учитывающие контекст, были однонаправленными, то есть могли работать, ориентируясь только на левую часть контекста. Однако чтобы повысить точность предсказания, необходимо было учитывать также и слова, стоящие справа. Для решения этой проблемы были предложены двунаправленные нейросети — по сути, ансамбль из двух сетей, работающих в противоположных направлениях (одна — с учетом левого контекста, вторая — правого) и приводящих свои решения к общему результату. Разработчики BERT придумали более эффективное и изящное решение: применили механизм внимания, позволяющий смотреть сразу по обе стороны искомого токена. Сам токен при этом заменяется маской, из-за чего языковая модель, на которой обучается алгоритм, получила название «маскированной». Еще одним свойством механизма внимания является умение увеличивать или уменьшать вес слов в контексте. Вес отвечает за значимость слова: чем сильнее оно влияет на решение модели, тем больше он должен быть.

BERT использует bite pair encoding (BPE) — передовой метод токенизации, при котором текст разбивается не на слова, а на части слов (подтокены), которые и хранятся в словарях, при этом неначальные элементы помечаются специальным символом ( ). Сегментация производится исключительно статистическими методами и не соотносится с морфемным составом слова. Еще одним новшеством в устройстве BERT является способ предобучения. Все эти особенности

[1] Авторы благодарят Даниила Гаврилова за предложение использовать этот подход.

выделяют модель BERT на фоне других систем и делают ее весьма полезной для нашей задачи.

Стандартная предобученная версия BERT довольно громоздкая: весит больше 600 Мбайт, обрабатывает предложение около 120 мс на CPU. Для тестирования нашего подхода, особенно если учесть небольшой размер имеющегося корпуса текстов, таких мощностей не требуется. В связи с этим было принято решение использовать дистиллированную версию модели: RuBERT-tiny2. В ее основе лежит классическая основа BERT (bert-multilingual), параметры которой были уменьшены: словарь сокращен до 83 тыс. токенов, размер эмбеддинга — в два раза, число слоев — с 12 до 3. Веса инициализированы случайным образом. Данные для обучающей выборки (2,5 млн коротких текстов) были взяты из параллельных русско-английских корпусов — от «Яндекс.Переводчика», OPUS-100 и Tatoeba. По оценкам разработчика, модель демонстрирует скорость предсказания одного токена 6 мс и весит всего 45 Мбайт.

Существующие модели BERT можно дообучать, чтобы использовать их для решения других задач. Дообучение, также известное как файнтьюнинг (от англ. *fine tuning* ‘тонкая настройка’), основано на предпосылке, что знания, полученные при решении общих заданий, помогают модели справиться с узконаправленными задачами. Суть этого метода состоит в том, что в уже обученную модель добавляется новый слой нейронов. Его веса задаются произвольно, и модель корректирует их, опираясь на параметры остальных слоев. За корректировку весов в модели отвечает функция потерь: это функция, которая сравнивает полученный результат предсказания с ожидаемым и вычисляет ошибку. В процессе дообучения модель старается подбирать веса так, чтобы ошибка была как можно меньше. После дообучения модель можно проверять в действии и оценить ее работу с помощью существующих метрик оценки качества.

Модификация модели и результаты

Чтобы модель BERT смогла работать с текстами на древнерусском языке, ее необходимо было модифицировать. В ее первом слое содержится «словарь» — индексы и соответствующие им токены. В задачах, связанных с автоматической обработкой языка, кодирование является важнейшим этапом, так как компьютер работает с числовыми векторами, а не со словами естественного языка.

Поскольку индексов для древнерусских слов в оригинальном BERT не предусмотрено, необходимо было произвести кодировку самостоятельно и заменить исходный словарь новым. Важным условием было соблюдение размерности: количество токенов в новом слое должно было быть не больше, чем в старом (в противном случае сети не хватило бы нейронов для их обработки).

Предобучение токенизатора проводилось на неразмеченной части корпуса СКАТ. Далее была проведена настройка на задачу частеречной разметки с помощью размеченного подкорпуса из трех житий: Димитрия Прилуцкого, Дионисия Глушицкого и Кирилла Новоезерского. Общий объем размеченного корпуса — 35 603 токена, 2885 предложений. Было проведено предварительное разбиение корпуса на обучающую и тестовую выборки в соотношении 0,85:0,15. Модель дообучалась на задачу классификации токенов по стандартной процедуре, описанной в работе [Akbik et al. 2018].

Модели были обучены классифицировать токены в соответствии со следующим набором тегов: «COLON», «COMMA», «DOT», «SEMICOLON», «гл», «инф», «межд», «мест», «нар», «посл», «пред», «прил», «прич», «союз», «сущ», «част», «числ».

В качестве исходной модели была выбрана модель bi-LSTM со следующими параметрами:

Window size 2
LSTM state size 200
Optimiser Adagrad
Initial learning rate 0.05
Decay rate 0.05
Dropout rate 0.5
Кроме того, были обучены модели RuBERT-tiny2 и RuBERT.
Были выбраны следующие гиперпараметры дообучения:
learning_rate=3e-5,
num_train_epochs=100,
weight_decay=0.01,

Размер батча был установлен на отметке 64 текстов для всех экспериментов.

Для оценки качества работы модели использовались стандартные метрики F1 (среднее гармоническое точности и полноты) и Accuracy (количество верно классифицированных объектов относительно общего количества всех объектов). На первых двух эпохах обучения показатели были низкими, поскольку вначале модель использовала существующие у нее веса для текстов на русском языке; однако затем она перестроилась, качество предсказания стало расти. Результаты экспериментов приведены в таблице.

Название модели	**F1**	**Accuracy**
bi-LSTM	0,72	0,81
RuBERT-tiny2	0,80	0,87
RuBERT	0,81	0,88

Ручная проверка результатов

С помощью полученной модели была произведена разметка нового файла — «Жития Александра Свирского». Результат разметки был проанализирован вручную. Для этого из текста случайным образом были выбраны 100 предложений. В целом модель показала достаточно хорошие результаты. Приведем пример предложения, не имеющего ошибок в морфологической разметке:

(2) сты҃ѧ мл҃твою во|ѡрѫжаше҆ .
прил част сущ гл DOT

Однако существует определенное количество контекстов, в которых распознавание морфологической категории части словоформ оказалось для модели затруднительным. Ниже рассмотрим некоторые распространенные случаи.

Ошибки, объяснимые линейным положением

В некоторых случаях словоформа, находящаяся между двумя элементами одинаковой категориальной принадлежности, ошибочно получает ту же морфологическую помету, что и эти два элемента. Так, в примерах (3) и (4) частица *же*, находящаяся между двумя местоимениями, определена как местоимение.

(3) сеⷤ| все бы҃ да з҄бꙋде||тсѧ̇ бгоⷨ҇ реченое̓ .
мест ***мест*** *мест гл част гл сущ прич DOT*

(4) … и тыⷤ ми реклъ е̓си …
союз мест ***мест*** *мест гл гл*

Однако возникает подозрение, что модель вообще оказывается склонной к постановке как минимум двух одинаковых категориальных помет подряд, даже если одна из них не соответствует действительности. Скажем, в примере (5) существительное *имя* перед двумя прилагательными *святыя* и *живоначальныя* опознано также как прилагательное, в примере (6) возвратное местоимение *си* после глагола *сокрушаеши* помечено как глагол.

(5) въ и̓мѧ̓ ст҃ыа̀| живоначаⷧ҇ны̑ трⷪ҇ца .
пред ***прил*** *прил прил сущ DOT*

(6) По|ч҄то та сокрꙋ|шаешиⷭꙵ҇ тѣло …
нар нар гл ***гл*** *сущ …*

В примере (7) аорист с имперфектным значением *бѣ* перед существительным *недугомъ* ошибочно определен как существительное, а частица *же* в примере (8), следующая за наречием, указана моделью как наречие.

(7) ѿдер҄жима бѣ недꙋгомъ ве|лїи̓мъ …
прич ***сущ*** *сущ прил …*

(8) … и̓ послⷤ҇ѣ всⷯѣ| и̓схожаше .
союз нар ***нар*** *мест гл DOT*

Наконец, в примере (9) частица *же*, следующая за двумя наречиями, также названа наречием.

(9) и̓ тако па|ки послⷤ҇ѣ в҄сⷯѣ из цр҃квї и̓схо|жаше .
союз нар нар нар ***нар*** *мест мест сущ гл DOT*

Ошибки, объяснимые сходством

Смешение финитного глагола и существительного. Неоднократно в выбранных предложениях повторяется смешение финитной глагольной формы с именем существительным.

(10) ѿдер҄жима бѣ недꙋгомъ ве|лїи̓мъ на̇ м҄ноⷢ҇о время .
прич ***сущ*** *сущ прил пред прил сущ DOT*

(11) ꙗ̓ко| е̓диномꙋ хотѧщꙋ| бⷥѣмⷧ҇ѡствовати …
союз числ прич ***сущ***

При этом возможно отождествление не только глагола с существительным, но и именной части речи (местоимения или существительного) с глаголом.

(12) ꙗ҆ м҆нѣ|тѝ е҆и … съкрꙋ|шитѝ .
союз ***сущ гл*** *… инф DOT*

(13) правѣнѝ бо рⷮ҆е ꙗ҆ко финѝ процвѣ|тⷮ҆е .
гл *союз гл союз сущ гл DOT*

(14) но и҆ ближ҆ни|мъ сꙋсѣдо҆ и҆| глти҆ .
союз союз прил ***гл*** *мест инф DOT*

Возможно, впрочем, такое отождествление обязано своим появлением определенному сходству конечных элементов некоторых существительных с отдельными показателями глагольного словоизменения, как в примере (15), где слово ризу можно, не принимая во внимание контекст и его значение, счесть формой первого лица единственного числа настоящего времени (как *везу*).

(15) нйч҆то҆ в҆зⷣе| раз҆вѣ потре|в҆нꙋю ризꙋ .
мест гл ***сущ*** *прил гл DOT*

Смешение прилагательного с другими частями речи

Словоизменительные элементы кратких форм имени прилагательного зачастую совпадают со словоизменительными или словообразовательными средствами других частей речи. Поэтому прилагательное иногда опознается моделью как наречие или существительное:

(16) пожи҆ же| мало в҆рема в до|брѣ испове҆данїи҆ .
прич част ***нар*** *сущ пред* ***сущ*** *сущ DOT*

Случаи смешения прилагательного с существительным весьма многообразны:

(17) пришⷭ҆ши| нѣкаа҆ невиди|маа҆ бїа сила .
гл мест прич ***сущ*** *сущ DOT*

(18) конечне| и҆сцелєнїе полꙋчи҆ .
… ***сущ*** *сущ прич DOT*

(19) ж҃ и҆ м҆ногы҃ а҆испо|вѣда ѡ҆ преслав҆нⷭ҆ѣ мⷭ҆жи сⷭ҆ѣ .
союз союз ***сущ*** *гл пред прил сущ мест DOT*

При этом иногда прилагательное может стоять и в полной форме:

(20) бл҆вит҃ѣ҃ гл҃а ѡ҆ц҃ы ч҆тнїи҆ .
гл ***гл*** *прич сущ* ***сущ*** *DOT*

Видимо, по той же причине — из-за фонетической близости конечных элементов словоформ — наречия, как и прилагательные, могут время от времени опознаваться как существительные:

(21) и҆ еще҆ ж҃ и҆ въ пекар҆ню| часто хожⷭ҆аше .
союз нар част союз пред сущ ***сущ*** *гл DOT*

(22) и҆ глада҃ сѣмо и҆ ѡ҆вамо …
союз гл нар союз ***сущ*** *…*

Особенности разметки словоформ с элементом ѿ в начальной позиции. Отдельную подгруппу составляют ошибки, связанные со словоформами, содержащими букву ѿ (в формате plain text передается буквой *w* с последующим выносным элементом, заключенным в скобки).

Так, в примерах (23) и (24) существительное *отець в* графическом представлении о҃ смешивается с предлогом ѿ — и наоборот. Поскольку обе словоформы в формате plain text различаются только содержимым скобок (титлом), их спорадическое смешение объяснимо.

(23) и҆ о҃| мъногы҃ бꙋ҆деши| дх҃вны҃ .
союз ***пред*** *прил гл прил DOT*

(24) и҆зыди| ѿ зе҃мла твоеѧ҆| и҆ ѿ рожениꙗ҆ сво|е҆го .
гл пред сущ мест союз ***сущ*** *сущ мест DOT*

Однако следует отметить, что и достаточно длинные словоформы, если они начинаются сочетанием «*ω + скобка*» (титло), распознаются моделью как предлоги. Именно предлогами модель сочла аорист ѿдо҃ и причастие ѿгнанꙋ҆ в примерах (25) и (26).

(25) … и҆ ѿдо҃ в҆ до҃ сво҃ с мирꙋ҃ .
… *союз* ***пред*** *пред сущ мест пред сущ DOT*

(26) … ѿгнанꙋ҆ бы|в҆шꙋ҆ нечи҃томꙋ҆| дх҃ꙋ҆
… ***пред*** *прич прил сущ*

Ошибки при разметке глагольных форм

Смешение причастий с другими глагольными формами. С финитным глаголом может смешиваться как пассивное (примеры (27)–(28)), так и активное (примеры (29)–(30)) причастие.

(27) и ѿ всѣ҆ поносими| бꙋ҆дѣ҆
союз пред мест ***гл*** *гл*

(28) списано ї҆роди|ѡ҆но҃ и҆гꙋ҆мено҃
гл *сущ сущ*

(29) и блг҃внъ бы҃ ѿ ст҃го
союз прил ***гл*** *пред прил*

(30) …по|стническыми сѣдинна҃ оу҆краси|вса .
…*прил сущ* ***гл*** *DOT*

Определение одной и той же грамматической формы иногда как причастия, иногда как глагола может встречаться в пределах одного предложения, как в примере (31):

(31) паче҆ | стоя҆ на мл҃тваχъ| и҆ моля҃ бо҃ .
нар част прич пред сущ союз ***гл*** *сущ DOT*

Путаница причастий и прилагательных. Видимо, по причине морфологического сходства, а также из-за возможности перехода отдельных причастий в разряд отглагольных прилагательных модель время от времени смешивает эти категории.

(32) … чл҃къ [нѣ]|кыꙵ … живыи блⷥи ѿби|тели ст҃го .
… сущ мест … ***прил*** *сущ сущ прил DOT*

(33) и гл҃ше вола гн҃а чⷣаⷪ да бꙋдеⷮ .
союз гл ***прич прич*** *сущ част гл DOT*

Смешение существительных с инфинитивами. Видимо, из-за фонетического сходства падежных форм (и даже просто основ) некоторых склонений существительных с глагольной формой инфинитива на *-ти* зафиксированы случаи неверной разметки такого рода:

(34) сїє … предахⷪⷨ къ ре|вⸯности послꙋша|телєⷨ .
Мест … гл пред ***инф*** *сущ DOT*

(35) вѣдꙋщи бо ст҃ость| жиⷮаꙵ єⷢ .
прич союз сущ ***инф*** *мест DOT*

Общий итог ручной проверки

Как можно видеть, в основе неудачных морфологических разборов с большой долей вероятности лежат ошибки модели, связанные либо с неверным обобщением закономерностей линейного положения («за элементом A с большей вероятностью следует элемент B, чем элемент C»), либо со сходством словоформ, причем как в крайней левой (приставка ѿ-), так и в крайней правой (различные фонетически подобные суффиксы и флексии) позиции. Ошибки, возникающие при сходстве словоформ, объясняются тем, что BERT использует сжатие данных BPE, при котором в словарях хранятся не целые слова, а части слов (подтокены).

Следует упомянуть, что тексты СКАТ могут содержать символы — разделители строк, но из размеченного корпуса, на основе которого проводилась настройка модели, разделители строк были удалены. Однако это не повлияло на конечный результат: при обработке вариантов текста с разделителями и без них модель производит разметку примерно одинаково.

Выводы

Опыт использования нейросетевой модели с языковым переносом для частеречной разметки древнерусских текстов можно считать в целом успешным, учитывая, что в эксперименте применялись небольшие дистиллированные модели, которые быстро обучаются и не требуют больших ресурсов. Предложенный подход позволяет использовать предобученные языковые модели для дообучения на материале малоресурсных языков, т. е. языков, для которых трудно собрать большие корпусы. Описанная модель может использоваться для частеречной разметки других текстов в рамках корпуса СКАТ и — в перспективе — за его пределами.

Словари

Поляков — Поляков А. Е. *Грамматический словарь церковнославянского языка (по материалам корпуса)*. http://dic.feb-web.ru/slavonic/dicgram/ (дата обращения: 04.07.2021).

Литература

Алексеева и др. 2022 — Алексеева Е. Л., Азарова И. В., Рогозина Е. А., Сипунин К. В. Корпусное выделение библейских цитат в севернорусских житийных текстах XVI–XVII вв. В сб.: *Источниковедение литературы и языка (археография, текстология, поэтика): Памяти Елены Ивановны Дергачевой-Скоп*. Новосибирск: ГПНТБ СО РАН, 2022. С. 237–242.

Алексеева, Миронова 2017 — Алексеева Е. Л., Миронова Д. М. Компьютерная текстология. В кн.: *Прикладная и компьютерная лингвистика*. Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). М.: URSS, 2017. С. 259–272.

Букия, Протопопова 2016 — Букия Г. Т., Протопопова Е. В. Машинное обучение в лингвистике. В кн.: *Прикладная и компьютерная лингвистика*. Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). М.: URSS, 2017. С. 121–137.

Васильев 2021 — Васильев Ю. *Обработка естественного языка Python и SpaCy на практике*. СПб.: Питер, 2021.

Захаров 2017 — В. П. Захаров. Корпусная лингвистика В кн.: *Прикладная и компьютерная лингвистика*. Николаев И. С., Митренина О. В., Ландо Т. М. (ред.). М.: URSS, 2017. С. 138–155.

Николенкова 2000 — Николенкова Н. В. Некоторые принципы синтаксической организации церковнославянского текста: на примере житийных текстов XI–XIII веков: дис. … канд. филол. наук. М., 2000.

Поляков 2014 — Поляков А. Е. Корпус церковнославянских текстов: проблемы орфографии и грамматики. *Przegląd Wschodnioeuropejski*. 2014 (1): 245–254.

Akbik et al. 2018 — Akbik A., Blythe D., Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: *Proceedings of COLING 2018. The 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, 2018. P. 1638–1649.

Azarova et al. 2021 — Azarova I., Alekseeva E., Lavrentiev A., Rogozina E., Sipunin K. Content Structuring in the St Petersburg Corpus of Hagiographic Texts (SCAT) *Scripta & e-Scripta. The Journal of Interdisciplinary Mediaeval Studies*. 2021, (21): 69–78.

Dereza et al. 2016 — Dereza O. V., Kayutenko D. A., Fenogenova A. S. Automatic Morphological Analysis for Russian: a Comparative Study. Computational Linguistics and Intellectual Technologies. In: *Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies*. Student session (online publication). 2016. http://www.dialog-21.ru/media/3473/dereza.pdf (дата обращения: 04.07.2021).

Haug 2015 — Haug D. T. T. Treebanks in historical linguistic research. In: Viti, Carlotta (eds), *Perspectives on Historical Syntax*. Amsterdam: John Benjamins Publishing Company. P. 185–202.

Jurafsky, Martin 2020 — Jurafsky D., Martin J. H. Chapter 7. Neural Networks and Neural Language Models. In: *Speech and Language Processing*. Draft of December 30, 2020. P. 137–147. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (дата обращения: 04.07.2021).

Kuratov, Arkhipov 2019 — Kuratov Yu., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language 2019. https://arxiv.org/abs/1905.07213 (дата обращения: 04.07.2021).

McCulloch, Pitts 1943 — McCulloch W. S., Pitts W. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*. 1943 (5): 115–113.

Mitrenina 2014 — Mitrenina O. The Corpora of Old and Middle Russian Texts as an Advanced Tools for Exploring an Extinguished Language. *Scrinium*. 2014, (X): 455–461.

NMT — Neural Morphological Tagging. http://docs.deeppavlov.ai/en/master/features/models/morphotagger.html (дата обращения: 04.07.2021).

Podtergera et al. 2016 — Podtergera I., Mocken S., Besters-Dilger J. *SlaVaComp — COMPutergestützte Untersuchung von VAriabilität im kirchenSLAvischen*. Forschungsergebnisse. Freiburg: Universitätsbibliothek Freiburg, 2016.

Sokolov 2019 — Sokolov E. G. The project of a deeply tagged parallel corpus of Middle Russian translations from Latin. *Journal of Applied Linguistics and Lexicography*. 2019, (2): 337–364.

Статья поступила в редакцию 22 июня 2022 г.
Статья рекомендована к печати 3 февраля 2023 г.

Vadim V. Gudkov

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russia
st071220@student.spbu.ru

Olga V. Mitrenina

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russia
o.mitrenina@gmail.com

Evgenii G. Sokolov

Institute for Linguistic Studies, Russian Academy of Sciences,
9, Tuchkov per., St. Petersburg, 199053, Russia
pan_liwerij@mail.ru

Angelina A. Koval

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russia
st069276@student.spbu.ru

Language-based transfer learning approaches for part-of-speech tagging on Saint Petersburg Corpus of Hagiographic texts (SKAT)

For citation: Gudkov V. V., Mitrenina O. V., Sokolov E. G., Koval A. A. Language-based transfer learning approaches for part-of-speech tagging on Saint Petersburg Corpus of Hagiographic texts (SKAT). *Vestnik of Saint Petersburg University. Language and Literature.* 2023, 20 (2): 268–282. https://doi.org/10.21638/spbu09.2023.205 (In Russian)

The article describes an experiment about training a part-of-speech tagger using artificial neural networks on the St. Petersburg Corpus of Hagiographic Texts (SKAT), which is being developed at the Department of Mathematical Linguistics of St. Petersburg State University. The corpus includes the texts of 23 manuscripts dating from the 15th–18th centuries with about 190,000 words usages, four of which were labelled manually. The bi-LSTM, distilled RuBERT-tiny2 and RuBERT models were used to train a POS tagger. All of them were trained on modern Russian corpora and further fine-tuned to label Old Russian texts using a technique called language transfer. To fine-tune transformer-based language models it was necessary to tokenize the texts using byte pair encoding and map tokens from the original Russian-language tokenizer to the new one based on indices. Then the model was fine-tuned for the token classification task. To fine-tune the model, a tagged subcorpus of three hagiographical texts was used, which included 35,603 tokens and 2,885 sentences. The experiment took into account only the tags of the parts of speech, the classification included seventeen tags, thirteen of which corresponded to parts of speech, and the remaining four marked punctuation marks. To evaluate the quality of the model, the standard metrics F1 and Accuracy were used. According to automatic evaluation metrics, the RuBERT model showed the best result. Most of the errors were related to incorrect generalization of linear position patterns or to the similarity of word forms in both the extreme left and extreme right positions.

Keywords: hagiography, corpus of Old Russian texts, neural network tagging, language-based transfer learning, part-of-speech tagging.

References

Алексеева и др. 2022 — Alekseyeva Ye. L., Azarova I. V., Rogozina E. A., Sipunin K. V. Corpus selection of biblical quotations in northern Russian hagiographic texts of the 16th–17th centuries. In.: *Istochniko-*

vedenie literatury i iazyka (arkheografiia, tekstologiia, poetika): Pamiati Eleny Ivanovny Dergachevoi-Skop. Novosibirsk: GPNTB SO RAN Publ., 2022. P. 237–242. (In Russian)

Алексеева, Миронова 2017 — Alekseyeva Ye. L., Mironova D. M. Digital text studies. In: *Prikladnaia i kompiuternaia lingvistika*. Nikolayev I. S., Mitrenina O. V., Lando T. M. (red.). Moscow: URSS Publ., 2017. P. 259–272. (In Russian)

Букия, Протопопова 2016 — Bukiya G. T., Protopopova Ye. V. Deep learning applications in linguistics. In: *Prikladnaia i kompiuternaia lingvistika*. Nikolayev I. S., Mitrenina O. V., Lando T. M. (red.). M. : URSS Publ., 2017. P. 121–137. (In Russian)

Васильев 2021 — Vasilyev Yu. *Natural language processing in Python and SpaCy.* A practical introduction. St Petersburg: Piter Publ., 2021. (In Russian)

Захаров 2017 — Zakharov V. P. Corpus linguistics. In: *Prikladnaia i kompiuternaia lingvistika*. Nikolayev I. S., Mitrenina O. V., Lando T. M. (eds). Moscow: URSS Publ., 2017. P. 138–155. (In Russian)

Николенкова 2000 — Nukolenkova N. V. *Some principles of the syntactic organization of the Church Slavonic text: On the example of hagiographic texts of the 11th–13th centuries*. Thesis for PhD in Philological Sciences. Moscow, 2000. (In Russian)

Поляков 2014 — Polyakov A. Ye. Church Slavonic corpus: Issues in orthography and grammar. *Przegląd Wschodnioeuropejski*. 2014 (1). P. 245–254. (In Russian)

Akbik et al. 2018 — Akbik A., Blythe D., Vollgraf R. Contextual String Embeddings for Sequence Labeling. In: *Proceedings of COLING 2018. The 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, 2018. P. 1638–1649.

Azarova et al. 2021 — Azarova I., Alekseeva E., Lavrentiev A., Rogozina E., Sipunin K. Content Structuring in the St. Petersburg Corpus of Hagiographic Texts (SCAT) *Scripta & e-Scripta. The Journal of Interdisciplinary Mediaeval Studies.* 2021, (21): 69–78.

Dereza et al. 2016 — Dereza O. V., Kayutenko D. A., Fenogenova A. S. Automatic Morphological Analysis for Russian: a Comparative Study. Computational Linguistics and Intellectual Technologies. In: *Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies*. Student session (online publication). 2016. http://www.dialog-21.ru/media/3473/dereza.pdf (accessed: 04.07.2021).

Haug 2015 — Haug D. T. T. Treebanks in historical linguistic research. In: Viti, Carlotta (eds), *Perspectives on Historical Syntax*. Amsterdam: John Benjamins Publishing Company. P. 185–202.

Jurafsky, Martin 2020 — Jurafsky D., Martin J. H. Chapter 7. Neural Networks and Neural Language Models. In: *Speech and Language Processing*. Draft of December 30, 2020. P. 137–147. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed: 04.07.2021).

Kuratov, Arkhipov 2019 — Kuratov Yu., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language 2019. https://arxiv.org/abs/1905.07213 (accessed: 04.07.2021).

McCulloch, Pitts 1943 — McCulloch W. S., Pitts W. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*. 1943 (5): 115–113.

Mitrenina 2014 — Mitrenina O. The Corpora of Old and Middle Russian Texts as an Advanced Tools for Exploring an Extinguished Language. *Scrinium.* 2014, (X): 455–461.

NMT — Neural Morphological Tagging. http://docs.deeppavlov.ai/en/master/features/models/morphotagger.html (accessed: 04.07.2021).

Podtergera et al. 2016 — Podtergera I., Mocken S., Besters-Dilger J. *SlaVaComp — COMPutergestützte Untersuchung von VAriabilität im kirchenSLAvischen.* Forschungsergebnisse. Freiburg: Universitätsbibliothek Freiburg, 2016.

Sokolov 2019 — Sokolov E. G. The project of a deeply tagged parallel corpus of Middle Russian translations from Latin. *Journal of Applied Linguistics and Lexicography*. 2019, (2): 337–364.

Received: June 22, 2022
Accepted: February 3, 2023