Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет
«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

Гавришина Екатерина Ильинична

МОДЕЛИРОВАНИЕ РАСПОЗНАВАНИЯ СЕМАНТИЧЕСКИХ РОЛЕЙ ДЛЯ РУССКОГО ЯЗЫКА

Выпускная квалификационная работа студента 4 курса бакалавриата группы 172

Академический руководитель образовательной программы	Научный руководитель
канд. филологических наук, доц.	канд. филологических наук, проф.
Ю.А. Ландер	О.Н. Ляшевская
_____	_____

« » _____ 2021 г.

Москва 2021

ОГЛАВЛЕНИЕ

Abstract

Semantic role labeling (SRL), or shallow semantic parsing, involves assigning semantic labels to the participants of predicate-argument structure. SRL and neural SRL tasks in particular have received much attention during the past decade. Results so far have been very promising, yet only several authors investigated this field on the Russian language data. Thus we carried out the SRL research based on semantically annotated resource FrameBank (Lyashevskaya, Kashkin 2015) by means of neural networks. Our approach includes preprocessing stage with cleaning and reorganizing the data, the comparison of different pretrained embeddings benefits, building stage with multiple layers stacked and architecture variants observed, and the overall manual and statistical analysis of achieved results at the final stage. As a result, we propose the multi-layered neural network with bidirectional long short-term memory (BiLSTM) applied for semantic parsing of Russian sentences.

Keywords: natural language processing, semantic role labeling, semantic parsing, the Russian language, neural networks.

Аннотация

Разметка семантических ролей (SRL), или неглубокий семантический парсинг, предполагает присвоение семантических меток участникам предикатно-аргументной структуры. В последнее десятилетие задаче SRL в целом и, в частности, нейронным подходам уделяется большое внимание. Результаты очень многообещающие, однако лишь несколько авторов исследовали эту область на материале русского языка. Таким образом, мы провели исследование SRL на основе семантически аннотированного ресурса FrameBank (Кашкин, Ляшевская 2013) с помощью нейронных сетей. Наш подход включает этап предобработки данных, сравнение предварительно обученных эмбеддингов, этап построения модели с несколькими слоями и выбора оптимальной архитектуры и заключительный этап критического анализа полученных результатов на внешнем тесте. В результате работы мы представляем многослойную нейронную сеть с двунаправленной долгой краткосрочной памятью (BiLSTM), применяемую для семантического парсинга русских предложений.

Ключевые слова: обработка естественного языка, распознавание семантических ролей, семантический парсинг, нейронные сети, русский язык.

1. Введение

Одной из самых сложных задач обработки естественного языка (NLP) является распознавание семантических ролей (SRL), или неглубокий семантический парсинг, который играет важную роль в широком спектре задач NLP, таких как извлечение информации, вопросно-ответные системы, машинный перевод и др. Парсинг включает в себя распознавание предикатно-аргументной структуры конструкции и присвоение меток именам и именным группам через определение их семантической роли в предложении (агенс, пациенс, цель и т.д.) с учетом предиката. Так, в предложении (1) ‘задержать’ — это предикат, его аргумент ‘полицейский’ играет роль агенса, ‘подозреваемый’ — тема конструкции, а ‘место преступления’ — место.

(1) *Полицейский задержал подозреваемого на месте преступления*.

Разметка семантических ролей включает в себя выявление и дизамбигуацию предикатов (для правильного выделения предикатных конструкций и их семантических особенностей), идентификацию аргументов (для определения того, является ли каждая зависимая группа релевантной в качестве аргумента) и классификацию аргументов, или собственно присвоение семантических ролей.

Концепция семантических участников ситуации была впервые упомянута в (Fillmore 1968), где они назывались “глубинными падежами”, но позже стали известны как “семантические роли” или “тематические роли”. Как только появились FrameNet (Baker et al. 1998) и PropBank (Palmer et al. 2005), семантически аннотированные корпуса для английского языка, исследований в области SRL стало значительно больше. Различные системы и архитектуры были протестированы с использованием данных FrameNet (Gildea, Jurafsky 2002; Das et al. 2014) и PropBank (Xue, Palmer 2004; Johansson, Nugues 2008; Punyakanok et al. 2008) в качестве обучающего материала. Исследования начинались со статистических методов, а затем развились в сторону нейронных подходов “с учителем” (англ. *supervised*) и “без учителя” (англ. *unsupervised*).

Тем не менее, для языков, отличных от английского, как синтаксически, так и семантически аннотированные корпуса малочисленны, но необходимы. Поэтому в российском научном сообществе авторы были вынуждены использовать для

обучения и валидации автоматически размеченные данные (Shelmanov, Smirnov 2014). Однако с появлением ресурса FrameBank (Кашкин, Ляшевская 2013), созданного по аналогии с FrameNet, он широко используется в различных методах решения задач NLP на материале русского языка.

Наше исследование (Гавришина 2020) стало пилотным методом создания эталонного корпуса русского языка. Данные корпуса FrameBank и градиентный бустинг были использованы для автоматической разметки семантических ролей на материале входного корпуса СинТагРус с эталонной синтаксической аннотацией (Дяченко и др. 2015) и получения примера полностью аннотированного корпуса с синтаксической и семантической информацией.

Цель данной работы — реализовать последовательную автоматическую разметку семантических ролей на материале СинТагРус, начиная с соответствующей предварительной обработки, с применением нейронной модели, обученной на данных корпуса FrameBank. С этой целью было поставлено несколько конкретных задач:

1) предобработка обучающего корпуса и приведение его к формату внешнего тестового корпуса СинТагРус;
2) разработка специальных метрик оценки качества модели, адаптированных под задачу разметки семантических ролей;
3) построение нейросетевых моделей с различными архитектурами;
4) сравнение качества работы моделей;
5) распознавание семантических ролей на тестовом корпусе;
6) анализ результатов.

Далее в работе после обзора существующих, релевантных для нашего исследования ресурсов и методов семантического парсинга, будут описаны корпуса, использованные для обучения и тестирования модели. Затем будет подробно изложена методология решения нашей задачи, включая предобработку данных, различные архитектуры модели, разработанные метрики оценки и предобученные эмбеддинги. После сравнения качества описанных моделей, полученные на тестовых данных результаты будут подвергнуты критическому анализу с целью повышения качества работы модели в будущем.

2. Обзор литературы

2.1. Ресурсы

FrameNet (Baker et al. 1998) является первым большим лексическим ресурсом, снабженным семантической информацией. Он был создан в рамках теории фреймов Ч. Филлмора (Fillmore 1968) — ситуаций с известным набором участников и ролей, а не глагольно-ориентированных подходов. Например, фрейм “commerce_buy” (рус. *купить*) состоит из двух основных семантических ролей — Покупатель и Товар, и ряда периферийных ролей, таких как Продавец, Деньги и т.д. В этот фрейм включены следующие лексические единицы: *купить, покупатель, клиент, покупка, приобрести, приобретение.*

В проекте PropBank (Palmer et al. 2005), однако, предложения из Penn Treebank (Kingsbury et al. 2002) представлены с предикатно-аргументной разметкой, добавленной поверх синтаксических деревьев. PropBank определяет набор семантических ролей для маркирования аргументов. Эти роли делятся на две категории: основные и неосновные роли. Основные роли (A0-A5 и AA) указывают на различную семантику в предикатно-аргументной структуре, в то время как неосновные роли являются модификаторами. В этой системе, ориентированной только на глагольные предикаты, глагол “buy” (рус. *купить*) описан при помощи следующих основных ролей и фреймов: A-0 “buyer” (рус. *покупатель*), A-1 “thing bought” (рус. *то, что куплено*), A-2 “seller” (рус. *продавец*), A-3 “price paid” (рус. *цена*), etc. Этот ресурс получил более широкое распространение, чем FrameNet, благодаря более общей репрезентации реальной языковой ситуации и более наглядному сопоставлению с синтаксической структурой.

VerbNet (Kipper et al. 2006) — еще один ресурс для английского языка, в котором глаголы сгруппированы в определенные семантические классы. Для примера, в системе VerbNet глагол “buy” (рус. *купить*) входит в класс под названием “get” (рус. *получить*), который состоит из почти двадцати пяти других глаголов (*найти, нанять, поймать* и т.п.) и ряда отличительных семантических ролей: Агенс, Тема, Источник, Бенефициар и Актив/Имущество.

Существует также проект NomBank (Meyers et al. 2004), который предоставляет структуру аргументов для распространенных существительных из корпуса Penn Treebank в дополнение к информации, указанной в PropBank. Например, в предложении “Her gift of a book to John” (рус. *Ее подарок Джону — книга*) аргументная структура имени “gift” (рус. *подарок*) выглядит следующим образом: A-0 *ее*, A-1 *книга*, A-2 *Джону*. В работе (Jiang, Ng 2006) авторы применили модель SRL на основе PropBank к структуре NomBank с частичным совпадением признаков, используемых для обучения классификации.

Некоторые работы по SRL опираются сразу на несколько перечисленных выше ресурсов, например (Guiglea, Moschitti 2006). Авторы объединяют данные FrameNet, PropBank и VerbNet: сначала фреймы сопоставляются с классами глаголов, а затем предикаты PropBank используются для расширения лексикона. Однако исследователи столкнулись с трудностями, связанными с недостаточным количеством обучающих примеров для идентификации фреймов. В исследовании (Маркова 2013) на данные корпусов FrameNet и VerbNet были отображены синонимические ряды (синсеты) из лексического ресурса WordNet (Fellbaum 1998), состоящего из семантических сетей для глаголов, существительных, прилагательных и наречий.

Другие не менее полезные ресурсы для английского языка включают OntoNotes (Hovy et al. 2006), объединяющий разнообразные по жанру источники (новости, блоги, телефонные разговоры), систему Bridge (Bobrow et al. 2007), корпус Discourse Representation Structures (Bos et al. 2017), основанный на теории представления дискурса, Parallel Meaning Bank (Abzianidze et al. 2017; Bos et al. 2017) и др.

Ряд семантически аннотированных ресурсов для русского языка не так разнообразен. Наиболее распространенным материалом для решения задач NLP на русском языке является FrameBank (Кашкин, Ляшевская 2013), который был построен по аналогии с FrameNet, но отличается тем, что основан на конструкциях конкретных лексем, а не на типовых ситуациях (фреймах). Корпус состоит из около 800 глаголов и словаря русских предикатных конструкций. Представлена аннотация фреймов, семантических ролей, лексико-семантических ограничений и морфосинтаксических структур.

Также в 2020 году был выпущен новый проект для русского языка — Russian PropBank (Moeller et al. 2020). Исследовательская группа реализовала идею автоматического проецирования английских семантических ролей на русские тексты. Они постарались сохранить параллели и выравнивание между двумя языками, несмотря на возникающие проблемы с дизамбигуацией смыслов и языковой спецификой. Результаты исследования находятся в свободном доступе в сети, но количество представленных предикатов не так велико. Глагол “купить” имеет ту же аргументную структуру, что и в PropBank (A-0 *покупатель*, A-1 *купленная вещь* и т.д.), с примерами контекста на русском языке.

С использованием семантических ролей строится ресурс ABBYY Compreno (Anisimovich et al. 2012), в основе которого лежит семантико-синтаксический парсер, предоставляющий лес деревьев разбора, однако технология находится в закрытом доступе.

2.2. Методы

Существует несколько подходов к решению задачи SRL, таких как алгоритмы, основанные на правилах, вероятностные методы, методы с использованием алгоритмов машинного обучения, а также разнообразные методы, основанные на применении нейронных сетей. Ранние системы SRL использовали правила, опирающиеся на грамматику, поскольку не было надежных аннотированных данных. Например, в (Sokirko 2001) семантический анализатор русского языка, основанный на правилах, преобразует предложение в семантический граф — такое представление существенно отличается от современных подходов. В (Shelmanov, Smirnov 2014) правиловый семантический анализатор применяется для автоматического аннотирования корпуса, который в дальнейшем используется при обучении модели.

Статистические методы распространились в сообществе исследователей английского языка вскоре после появления FrameNet. Используя его, (Gildea, Jurafsky 2002) впервые применили статистический подход к машинному обучению — их исследование было основано на статистическом синтаксическом парсере, который опирался на различные морфо-синтаксические особенности. Кроме того, авторы описали важность синтаксической информации для общей

производительности модели.

За последние десять лет возросла частота использования нейронных сетей для решения задачи SRL. Авторы выбирают различные модели и архитектуры в соответствии со своими идеями и имеющимися ресурсами. Подходы на основе сверточных нейронных сетей (CNN) представлены в (Collobert et al. 2011; Fonseca, Rosa 2013; Foland, Martin 2015; Marcheggiani, Titov 2017). Коллоберт и исследовательская группа (2011) поставили перед собой цель обучить CNN с использованием предобученных эмбеддингов на практически неразмеченых данных, чтобы получить внутренние представления для различных задач NLP (разметка частей речи, группирование (чанкинг), распознавание именованных сущностей и маркировка семантических ролей). В (Fonseca, Rosa 2013) авторы используют многослойную CNN архитектуру и векторные представления входных слов для разметки семантических ролей в португальских текстах. (Foland, Martin 2015) применяют архитектуру из (Collobert et al. 2011) в сочетании с выходными признаками синтаксического парсера зависимостей для присвоения семантических ролей глагольным и именным предикатам. В (Marcheggiani, Titov 2017) предложен метод с использованием графовых сверточных сетей (GCN) над деревьями синтаксических зависимостей для получения признаковых представлений слов. Авторы объединили слои GCN и LSTM для улучшения производительности.

Наиболее актуальной работой, посвященной задаче SRL с использованием сверточных нейронных сетей, является (Li et al. 2020). В этом исследовании авторы в том числе проверяют вклад синтаксической информации в производительность нейронных сетей SRL, рассматривая две различных системы обучения — с использованием данных о предикате и без них. Во втором случае модель для разметки последовательности не только распознает аргументы, но и идентифицирует предикаты. Авторы рассматривают три различных подхода к решению задачи — sequence-based (основанный на последовательности), tree-based (основанный на синтаксическом дереве зависимостей конструкции) и graph-based (основанный на графах). Слово в модели представлено как конкатенация следующих признаков: является ли слово предикатом, посимвольные эмбеддинги, полученные при помощи CNN с двунаправленным

слоем LSTM (BiLSTM), случайно инициализированные и предобученные эмбеддинги для слов, эмбеддинги лемм и частей речи, а также представления, полученные из предобученных языковых моделей. В sequence-based модели авторы используют BiLSTM в качестве энкодера входных предложений. В результате они утверждают, что положительное влияние синтаксиса на качество модели постепенно достигает своего верхнего предела.

В работе (Shelmanov, Devyatkin 2017) авторы использовали два подхода к обработке признаков в модели: в “простой” модели на вход давался один вектор, объединяющий все известные признаки аргумента, а в “сложной” модели первый слой разделен на три входа — для категориальных признаков, для эмбеддинга аргумента и для эмбеддинга предиката при наличии. Архитектура “простой” модели состоит из трех полносвязных слоев (Dense), а в “сложной” модели перед вторым слоем происходит конкатенация категориальных и лексических эмбеддингов. В качестве обучающей выборки в исследовании был использован корпус FrameBank с автоматической синтаксической разметкой при помощи парсера Google SyntaxNet (Andor et al. 2016). Также для эксперимента были выбраны лексические эмбеддинги, предобученные на корпусе НКРЯ и Википедии, однако большинство предикатов не были представлены в модели. Авторы отдельно разработали модель для распознавания ролей аргументов “неизвестных” предикатов (out-of-domain), которая показала меньшую производительность, чем на “известных” предикатах, но доказала, что присутствие в обучающей выборке глаголов, близких к “неизвестным” в векторном пространстве, значительно повышает качество.

Другой широко используемой архитектурой для распределения семантических ролей с помощью нейронных сетей является механизм внимания (Vaswani et al. 2017). В работе (Strubell et al. 2018) авторы усовершенствовали механизм внимания, добавив лингвистическую информацию (linguistically-informed self-attention, LISA): модель со слоями “множественного внимания” (multi-head self-attention) и слоями прямого распространения (feed-forward), которая также учитывает синтаксис посредством одной “головы” механизма внимания. В исследовании (Park 2019) была представлена глубокая архитектура, основанная на представлениях внимания и выборочном

подключении (selective connection). Модуль внимания в целом помогает улавливать корреляцию между словами в последовательности.

Наконец, выбор эмбеддингов является еще одним важным этапом решения задачи SRL. В частности, в последних исследованиях популярны предобученные контекстуализированные эмбеддинги BERT (Devlin et al. 2018) и ELMo (Peters et al. 2018), которые дают словам разные вектора на основе их семантики в контексте. Авторы (Larionov et al. 2019) провели эксперименты на корпусе FrameBank с различными типами эмбеддингов (word2vec, FastText, ELMo, BERT) и доказали преимущество эмбеддингов, генерируемых глубокими предобученными языковыми моделями. Более того, BERT превзошел эмбеддинги ELMo и RuBERT для “известных” предикатов. В статье (Shi, Lin 2019) различные модели на основе BERT описаны в рамках span-based и dependency-based представлений без использования синтаксических признаков.

3. Данные

Одной из задач этой работы является разметка ролей на материале корпуса СинТагРус в качестве внешнего теста и практического применения парсера. Для обучения модели, автоматически приписывающей семантические роли, был использован корпус FrameBank со слоем автоматической синтаксической разметки зависимостей посредством парсера SyntaxNet, разработанный в работе (Shelmanov, Devyatkin 2017). Часть этих данных была также использована в качестве валидационной и тестовой выборки.

3.1. Данные для обучения

Корпус FrameBank+SyntaxNet находится в открытом доступе в формате JSON[1]. Каждый элемент словаря представлен списком предложений, некоторые из которых имеют слой семантической разметки, то есть размеченную предикатно-аргументную структуру. Всего предложений в корпусе 322357, из которых только 7952 обеспечены семантическими ролями. Размеченные предложения являются основным материалом для исследования.

[1] http://nlp.isa.ru/framebank_parser/data/

Среди аннотированных предложений 7357 конструкций предикатно-ориентированные, в остальных наличие предиката не подразумевается. Большинство предикатов (7216) выражено глаголами, 79 относятся к именам существительным, 43 выражены прилагательными, причастиями и деепричастиями. Всего в корпусе 609 уникальных лемм предикатов.

Данные достаточно разрежены, так как в изначальном корпусе в предложении могла быть размечена максимум одна предикатно-аргументная конструкция и, кроме того, большинство токенов в последовательности не являются аргументами обозначенного предиката. Здесь и далее в работе “предикат” будет включен в группу семантических ролей для удобства описания и построения модели. Только 10% токенов (21449 из 215437) имеют семантическую роль, и из них 34% присвоена роль предиката, 7% — роль агенса, 6% — роль пациенса, 5% — роль темы в предложении.

Кроме семантической разметки представлен слой морфологической и автоматически добавленной синтаксической информации. Описание каждого токена включает словоформу, лемму слова, морфологические признаки (падеж, число, время, лицо и др.), часть речи, синтаксический ранг участника, семантические ограничения на заполнение слота конструкции (Кашкин, Ляшевская 2013), а также зависимостные признаки — индекс вершины и характер синтаксической зависимости (модификатор, союз, пунктуация и т.п.).

3.2. Данные для тестирования

Данные, использованные в качестве теста, — это глубоко аннотированный корпус текстов русского языка СинТагРус (Дяченко и др. 2015; Droganova et al. 2018) с эталонной синтаксической разметкой в формате Universal Dependencies (UD). Рассматриваемые данные разбиты на обучающую, валидационную и тестовую выборки для удобства использования и находятся в открытом доступе в формате CoNLL-U[2].

Для настоящего исследования в качестве примера был взят тестовый подкорпус СинТагРуса, состоящий из 6491 предложения и 117521 токена.

[2] https://github.com/UniversalDependencies/UD_Russian-SynTagRus

Набор признаков в корпусе по большей мере совпадает с корпусом FrameBank: словоформа, лемма, морфологические признаки, часть речи, индекс вершины и тип синтаксических отношений. Кроме того, в корпусах формата CoNLL-U подразумевается признак XPOS, обозначающий специфичные для языка части речи, но для русского языка такие обозначения нерелевантны. Также дополнительным признаком является наличие пробела после токена. Учитывая структуру корпуса FrameBank, использованного в качестве обучающего материала для модели, два последних описанных признака не учитывались при разметке семантических ролей, так как отсутствовали при обучении.

4. Метод

В работе (Гавришина 2020) была предпринята первая попытка автоматической разметки семантических ролей на материале FrameBank. В качестве классификатора был использован CatBoost Classifier (Prokhorenkova et al. 2018), ориентированный на категориальные признаки, который размечал каждый токен без использования всей последовательности и информации о контексте. По результатам предсказания, роль агенса определялась верно с вероятностью 82%, роль пациенса — с вероятностью 63%, тема — с вероятностью 50%. В исследовании была также подтверждена гипотеза о важности синтаксических признаков для производительности модели.

У пословного подхода есть свои недостатки: приписывание одинаковых ролей нескольким аргументам в пределах одного фрейма, неразличение актантов и сирконстантов и отсутствие адаптации к сложным клаузам (адвербиальные клаузы, однородные сказуемые), что приводит к приписыванию семантических ролей составляющим, которые не могут их иметь.

В большинстве исследований, посвященных разметке семантических ролей (Marquez et al. 2005; Daza, Frank 2018; Li et al. 2020), а также, например, частеречной (Zhang et al. 2018; Giuseppe 2020) и синтаксической разметке (Li et al. 2018), используется не пословный, а последовательный подход (sequence-based), ориентированный на входную последовательность целиком с опорой на контекст.

В основе настоящей работы лежит многослойная нейронная сеть, включающая слой эмбеддингов и двунаправленный слой долгой краткосрочной

памяти (BiLSTM). Ниже будут подробно описаны примененная предобработка обучающего корпуса, выбор лексических эмбеддингов и собственно архитектура модели и ее вариации.

4.1. Предобработка

В нашем исследовании предполагается, что обученная модель размечает семантические роли в предложении без информации о предикате, которая могла бы значительно улучшить качество аннотации. Учитывая невозможность использования данных о предикате, в качестве замены этого признака была представлена информация о вершине каждого токена — лемма вершины и часть речи. Выбор этих признаков обусловлен тем, что семантическая роль аргумента напрямую связана с лексической и морфологической характеристикой вершины. Так, если в обучающей выборке присутствует глагол “говорить” в качестве предиката, модель может более точно распознать семантику зависимых аргументов — “говорящий” и “содержание высказывания”. Кроме того, часть речи вершины полезна для исключения большинства претендентов на роль: зависимые наречия или прилагательного будут маркированы гораздо реже, чем зависимые глагола.

Еще один добавленный признак, описывающий отношения зависимого и вершины, — это относительное расположение зависимого. Каждому токену присвоено абсолютное число, равное расстоянию до вершины, с минусом, в случае если зависимое находится в постпозиции, и с плюсом, если в препозиции. У вершины всего предложения этот признак равен 0.

Изначальная репрезентация семантических ролей в корпусе отражает языковую реальность, в связи с чем распределение ролей не сбалансировано — есть определенно более частотные роли (агенс, пациенс, тема) и более редкие (например, “заместитель”, “малефициант”, а также двойные роли “источник звука - пациенс”, “агенс - говорящий” и др.). Так как малопредставленные роли могут только отрицательно повлиять на качество модели, токены в обучающей выборке с меткой роли, представленной меньше чем в 20 примерах, были обозначены как не имеющие роли.

В связи с тем, что подход основан на формате последовательности и для

нейросети необходимо изначально задать конкретную длину входного вектора, за это значение была взята максимальная длина предложения из внешнего тестового корпуса СинТагРус — 123 токена. Таким образом, предложения длиной более 123 токенов (около 80 предложений) были удалены из обучающих данных.

Кроме того, в исходном наборе данных корпуса FrameBank представлены признаки, отсутствующие во внешнем тесте, которые, следовательно, не могли быть использованы для обучения нейросети. В число опущенных признаков вошли признак *feat*, обозначающий список морфологических характеристик в неудобном формате (был оставлен признак *feat_p,* содержащий те же показатели в формате UD), признаки *sem* и *sem2*, содержащие лексико-семантические теги, и признак *pred* с индексом предиката в системе FrameBank.

Полученный после предобработки формат данных изображен на Рисунке 1. В таблице представлены следующий признаки: *FullIndex* (индекс предложения в корпусе FrameBank), *index* (индекс токена в предложении), *lemma* (лемма токена), *POS* (часть речи токена), *feat_p* (морфологические признаки), *link_name* (тип синтаксических отношений), *lemma_parent* (лемма вершины), *pos_parent* (часть речи вершины), *parent* (расположение относительно вершины) и целевая переменная *Role* (семантическая роль).

FullIndex	index	lemma	POS	feat_p	link_name	lemma_parent	pos_parent	parent	Role
102819_8	1	в	ADP	_	case	итог	NOUN	1	_
102819_8	2	итог	NOUN	Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing	nmod	опоздать	VERB	8	_
102819_8	3	за	ADP	_	case	год	NOUN	2	_
102819_8	4	20	NUM	_	nummod	год	NOUN	1	_
102819_8	5	год	NOUN	Animacy=Inan|Case=Gen|Gender=Masc|Number=Plur	nmod	опоздать	VERB	5	_
102819_8	6	упрямство	NOUN	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing	nmod	год	NOUN	-1	_
102819_8	7	и	CCONJ	_	cc	упрямство	NOUN	-1	_
102819_8	8	косность	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	conj	упрямство	NOUN	-2	_
102819_8	9	мы	PRON	_	nsubj	опоздать	VERB	1	субъект перемещения
102819_8	10	опоздать	VERB	Aspect=Perf|Mood=Ind|Number=Plur|Tense=Past|Ve...	root	_	_	0	предикат
102819_8	11	на	ADP	_	case	жизнь	NOUN	3	_
102819_8	12	тысяча	NOUN	Animacy=Inan|Case=Acc|Gender=Fem|Number=Plur	nmod	жизнь	NOUN	2	срок
102819_8	13	человеческий	ADJ	Case=Gen|Degree=Pos|Number=Plur	amod	жизнь	NOUN	1	_
102819_8	14	жизнь	NOUN	Animacy=Inan|Case=Gen|Gender=Fem|Number=Plur	nmod	опоздать	VERB	-4	_
102819_8	15	.	PUNCT	_	punct	опоздать	VERB	-5	_
57483_4	1	"	PUNCT	_	punct	как	SCONJ	1	_
57483_4	2	как	SCONJ	_	cc	продолжать	VERB	5	_
57483_4	3	я	PRON	_	nsubj	рад	ADJ	1	_
57483_4	4	рад	ADJ	Degree=Pos|Gender=Masc|Number=Sing|Variant=Brev	advmod	продолжать	VERB	3	_
57483_4	5	"	PUNCT	_	punct	рад	ADJ	-1	_
57483_4	6	,	PUNCT	_	punct	рад	ADJ	-2	_
57483_4	7	продолжать	VERB	Aspect=Imp|Gender=Masc|Mood=Ind|Number=Sing|Te...	root	_	_	0	_
57483_4	8	Корсаков	NOUN	Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing	nsubj	продолжать	VERB	-1	_
57483_4	9	"	PUNCT	_	punct	что	SCONJ	1	_
57483_4	10	что	SCONJ	_	mark	умереть	VERB	4	_
57483_4	11	ты	PRON	_	nsubj	умереть	VERB	3	_
57483_4	12	еще	ADV	Degree=Pos	advmod	не	PART	1	_
57483_4	13	не	PART	_	neg	умереть	VERB	1	_
57483_4	14	умереть	VERB	Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|T...	advcl	Корсаков	NOUN	-6	_

Рис. 1. Формат представления данных.

Для обработки и кодирования категориальных признаков и лемм были составлены соответствующие словари индексов — словарь лемм (lemma2id), частей речи (pos2id), типов синтаксических отношений (rank2id), морфологических признаков (morph2id), положения относительно вершины (rel2id) и целевых семантических ролей (role2id). Затем все описанные признаки, включая лемму и часть речи вершины, при помощи словарных соответствий были преобразованы в векторы чисел. Последовательности, содержащие меньше токенов, чем установленная максимальная длина предложения (123 токена), были продлены при помощи паддинга.

4.2. Эмбеддинги

В работе рассматривались три различных предобученных дистрибутивно-семантических модели для получения эмбеддингов лемм. Модель, обученная на НКРЯ и Википедии (размер корпуса 788 млн слов, размерность вектора 300), и модель, обученная на русскоязычных новостях (2,6 млрд слов,

размерность вектора 300), находятся в свободном для скачивания доступе, благодаря проекту RusVectores[3] (Kutuzov, Andreev 2015). Третья модель[4] мультилингвальная, разработанная исследователями компании Google и основанная на сверточной нейронной сети, возвращает эмбеддинги размерностью 512 и может быть адаптирована под 16 различных языков, включая русский (Yang et al. 2019).

Матрицы весов, полученные при помощи описанных языковых моделей, подгружались в эмбеддинг-слои для леммы конкретного токена и леммы его вершины.

Весь корпус был поделен на обучающую, валидационную и тестовую выборки в соотношении 60:20:20, то есть модель была обучена на 4770 предложениях.

4.3. Особенности модели

Входной слой модели был разделен на несколько частей, так как объединение векторов категориальных признаков и лексических эмбеддингов лемм в один длинный вектор не оптимально в сравнении с отдельными входами для каждого признака (Shelmanov, Devyatkin 2017). Для проверки гипотезы о важности информации о вершине, в первой простейшей модели (см. рис. 2) не учитывались эмбеддинги леммы и части речи вершины и показатель расположения относительно вершины, и входов было всего четыре — для леммы (lemma), части речи (POS), морфологических признаков (morph) и синтаксического ранга (rank). В последующих вариациях архитектуры было добавлено еще три входа на первом слое для описанных признаков вершины.

[3] https://rusvectores.org/ru/models/

[4] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

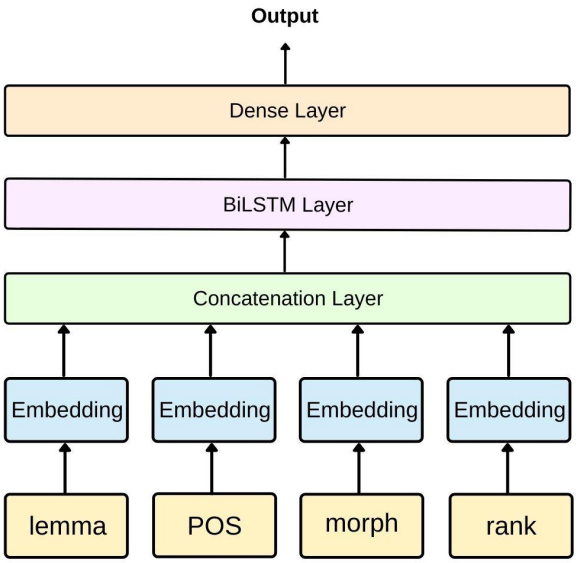

Рис. 2. Архитектура простейшей модели.

Основой архитектуры модели послужил слой двунаправленной долгой краткосрочной памяти (Bidirectional long short-term memory, BiLSTM). Изначально система LSTM — разновидность рекуррентных нейронных сетей (Recurrent neural networks, RNN), учитывающих предыдущие вычисления на каждом шаге — была предназначена для обучения долговременным зависимостям, то есть для эта архитектура невосприимчива к длительности временных разрывов между шагами модели. Двунаправленный LSTM — усовершенствованный вариант сети, который использует не только “прошлую” информацию, но и “будущую”, то есть рассматривает одновременно предшествующий контекст и последующий.

4.4. Архитектура

Первая нейронная сеть принимает на вход четыре признака, описанных выше (см. рис. 2), и обрабатывает их на уровне эмбеддинг-слоя, преобразуя целые числа (индексы) в плотные векторы фиксированного размера. В случае леммы токена эмбеддинг-слой преобразует индексы в векторы, используя веса из предобученной языковой модели. Затем полученные векторы каждого признака конкатенируются и подаются на вход BiLSTM слою. Внутри BiLSTM слоя применяется рекуррентный дропаут (recurrent dropout), который позволяет исключать из сети случайные нейроны с вероятностью 20% во избежание

переобучения (чрезмерной адаптации сети к обучающим примерам) не только на входе и выходе, но и на горизонтальных связях между рекуррентными модулями. Полученные векторы проходят через полносвязный выходной слой (Dense), который обеспечивает связь каждого нейрона со всеми входными нейронами, с активацией “softmax”, чтобы преобразовать векторы в вероятности классов.

В качестве оптимизатора применяется алгоритм Adam (adaptive moment estimation), позволяющий учитывать инерцию сдвига градиента, а также вторые моменты градиентов. Функции потерь для обучения модели — категориальная кросс-энтропия (перекрестная энтропия) — соответствует характеру задачи, так как она измеряет расхождение между двумя вероятностными распределениями (на выходе модели мы получаем векторы вероятности распределения классов) и адаптирована под мультиклассовую классификацию.

Во второй модели, для сравнения, было добавлено еще три входа на первом слое для признаков, относящихся к вершине, — леммы вершины, части речи вершины и расположения относительно вершины. Эта модификация нужна для проверки гипотезы о важности информации о вершине для разметки семантических ролей.

В третьей модели (см. рис. 3) после эмбеддинг-слоя для каждого входа был добавлен полносвязный слой (Dense) с вероятностью дропаута 0.4, чтобы лучше подстроиться под все типы признаков отдельно до их объединения и избежать переобучения. Одним из важных параметров слоя является активация ReLU (Rectified linear unit), которая наиболее часто используется в глубоком обучении и позволяет работать с множеством последовательно соединенных слоев.

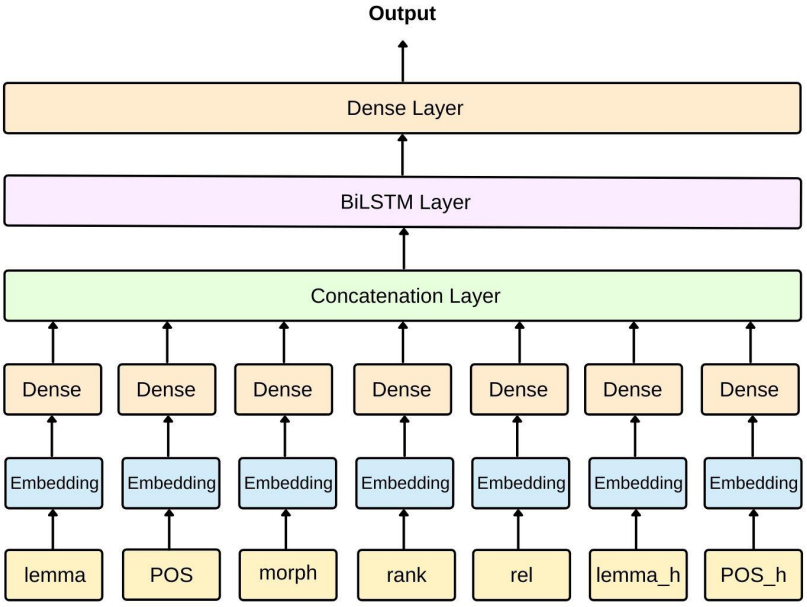

Рис. 3. Архитектура модели с полносвязным слоем до конкатенации.

Четвертая модель отличается от третьей тем, что был добавлен еще один полносвязный слой после BiLSTM перед выходным слоем. После него для регуляризации был применен дропаут с вероятностью 0.5, так как возрастание числа связей в сети может негативно влиять на производительность модели.

4.5. Метрики

Для оценки качества модели, в первую очередь, были использованы классические метрики точности: Accuracy, Precision и Recall. Accuracy дает самое базовое представление о работе классификатора, представляя собой долю примеров, по которым модель приняла правильное решение. Так как в нашем исследовании рассматривается задача многоклассовой классификации, была использована категориальная разновидность описанной метрики, адаптированная для one-hot кодированных признаков. Precision (точность) — это показатель “надежности” ответов модели, то есть доля примеров, действительно принадлежащих к конкретному классу, относительно всех примеров, которые модель отнесла к этому классу. Recall (полнота) — это показатель способности модели найти как можно больше вхождений конкретного класса, то есть доля найденных примеров относительно всех примеров этого класса.

В связи с тем, что имеющие семантическую роль аргументы составляют

только 10% от всей выборки, описанные выше метрики могут давать нерепрезентативно высокие результаты, благодаря превалирующему количеству токенов, не имеющих роли. Другими словами, если модель распознает и правильно определит все токены без семантической роли, базовая точность модели уже составит около 90%. Для большей наглядности качества была написана метрика Labeled Accuracy, которая при подсчете точности учитывает только те позиции, которым модель присвоила класс, отличный от "_" (отсутствие роли) и от "pad" (индекс паддинга).

С целью получения более конкретного представления о результатах были разработаны метрики оценки трех наиболее популярных семантических классов — предиката, агенса и пациенса. Для каждой из ролей создана функция оценки точности (precision) и полноты (recall), которые рассматривают только те позиции, которым модель предсказывает конкретный класс, или те, которые в действительности принадлежат этому классу, соответственно. Описанные дополнительные метрики не используются во время обучения модели, они представлены только на тестовых данных.

5. Сравнение моделей

5.1. Архитектуры

Здесь и далее будут рассматриваться результаты обучения четырех моделей:

1. Модель с четырьмя входами без учета информации о вершине и одним Dense-слоем.
2. Модель с входами для семи признаков и одним Dense-слоем.
3. Модель с добавлением Dense-слоя на каждом входе до конкатенации.
4. Модель с добавлением еще одного Dense-слоя после BiLSTM перед выходным слоем.

Как и предполагалось ранее, классические метрики, вычисляющие долю правильно предсказанных классов, показывают очень высокие результаты. Это связано с тем, что 90% данных составляют токены, не имеющие семантическую роль, то есть модель, предсказывающая только класс "_" (маркер отсутствия роли), имела бы сравнительно высокую производительность. С другой стороны,

эти значения в том числе помогают оценить тенденцию качества с изменением моделей.

Оценка качества моделей с точки зрения точности и полноты представлена в Таблице 1, а результаты работы модели относительно точности и полноты предсказания наиболее популярных семантических ролей указаны в Таблице 2.

Модели	**Accuracy**	**Precision**	**Recall**	**Labeled Accuracy**
Модель 1	0.9054	0.9123	0.8997	0.6936
Модель 2	0.9088	0.9252	0.8947	0.5997
Модель 3	0.9172	0.9395	0.8990	0.6588
Модель 4	0.9156	0.9409	0.9008	0.6812

Таблица 1. Значения основных метрик на тестовой выборке для моделей.

Модели	**Предикат**		**Агенс**		**Пациенс**	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Модель 1	0.7063	0.3221	0.25	0.0035	0.0	0.0
Модель 2	0.6351	0.4968	0.4884	0.0734	0.4839	0.06
Модель 3	0.7265	0.6646	0.4516	0.1958	0.4133	0.124
Модель 4	0.7512	0.6744	0.3529	0.0839	0.2475	0.100

Таблица 2. Точность и полнота частотных классов для моделей.

Можно заметить, что первая модель, не учитывающая признаки вершины, не предсказала ни одной роли пациенса верно, из чего можно сделать вывод, что еще менее частотные роли были представлены не лучше.

По общим результатам второй модели, в которой были добавлены входы для признаков, зависящих от вершины, видно, что существенно снизился показатель разработанной метрики Labeled Accuracy, не учитывающей токены без роли. Вероятно, эта закономерность связана с тем, что информация о лемме вершины избыточна и только “запутывает связи” в обученной модели, тем самым усложняя нахождение аргументов, или с тем, что первая модель слишком большому количеству вхождений предсказывала отсутствие роли, уменьшая тем самым долю

примеров, на которых считается кастомная метрика. Последнее объяснение подтверждается, если обратить внимание на то, что вторая модель гораздо лучше научилась предсказывать роль пациенса (точность 48%), а также стала чаще находить предикаты и агенсы на тестовых предложениях. По этим данным можно судить о подтверждении гипотезы о важности информации о вершине для усовершенствования модели. Даже если общая метрика точности (Labeled Accuracy) снизилась, вероятность определения и надежности предсказания для отдельных классов значительно возросла. Остальные общие классические метрики несколько выросли, что в том числе указывает на их не оптимальную репрезентативность.

Третья модель, в которой каждый вход отдельно обрабатывался при помощи полносвязного слоя, показала более высокие значения всех описанных метрик качества по сравнению с предыдущей моделью. Этот рост может быть связан с ростом внимания к каждому признаку в отдельности до их слияния, что позволяет модели подстроиться под различные параметры в изолированном режиме. По значениям точности и полноты избранных классов также видно, насколько увеличилось “покрытие” модели, то есть способность найти как можно больше вхождений рассматриваемого класса.

При добавлении еще одного полносвязного слоя после рекуррентного слоя общие метрики в среднем улучшились, а результаты отдельных классов в среднем снизились. В этом случае важно указать, что на тестовых данных третья модель распознала 1879 ролей в сравнении с четвертой моделью, отметившей только 1286 ролей, то есть вкупе с этими показателями третья модель выглядит более оптимально. Вероятно, добавление второго Dense-слоя повлекло за собой переобучение, не отразившееся на основных метриках, но усугубившее ситуацию с разметкой конкретных ролей.

5.2. Предобученные эмбеддинги

В этом разделе будет рассмотрено влияние смены предобученных языковых моделей на качество обучения нейронной сети. Стоит отметить, что статистические модели word2vec, обученные на НКРЯ+Википедия и на новостных источниках, не полностью покрывают словарь лемм обучающей выборки. Так, в

первой модели не было найдено около 3450 лемм, а во второй, учитывая специфику новостной лексики, — 4340. Однако предполагалось, что характер предложений в тестовом корпусе СинТагРус в большинстве лексически совпадает со словарем второй модели. Третья же модель, предложенная Google, основывается на контексте и для любой последовательности символов возвращает ненулевой вектор, так что говорить о “незнакомых” модели леммах в этом случае неуместно.

Для оценки предобученных языковых моделей, каждая из трех матриц весов была представлена в эмбеддинг-слоях на входах для лемм (собственно токена и его вершины) в третьей модели, так как эта архитектура показала относительно лучшее качество среди остальных, а также распознала наибольшее количество семантических ролей.

Сравнить первую и вторую модели непросто в связи с тем, что каждая из них опережает другую по нескольким параметрам. Однако если в большей степени ориентироваться на разработанную метрику Labeled Accuracy, то модель, обученная на новостных источниках, уступает первой модели. Это может быть связано с тем, что, как было указано, во второй модели большее количество лемм отсутствует, или с тем, что данные корпуса все же не полностью принадлежат новостному жанру и скорее пересекаются со словарем НКРЯ (который также включает некоторые новостные источники).

Результаты работы модели с мультилингвальными эмбеддингами не оправдали ожиданий, оказавшись в среднем самыми низкими из исследуемых. Такой эффект может быть вызван тем, что ориентированные на русский язык модели с RusVectores лучше представляют соответствующие тексты в отличие от мультилингвального векторного пространства, обученного для 16 языков. Скорее всего, эмбеддинги Google более полезны и важны для решения кросс-лингвистических задач.

Модели	**Accuracy**	**Precision**	**Recall**	**Labeled Accuracy**
НКРЯ + Википедия	0.9156	0.9409	0.9008	0.6812
Новостные	0.9173	0.9486	0.8957	0.6178

источники				
Эмбеддинги Google	0.9078	0.9300	0.8933	0.6722

Таблица 3. Качество модели с использованием разных предобученных эмбеддингов.

Модели	**Предикат**		**Агенс**		**Пациенс**	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
НКРЯ + Википедия	0.7512	0.6744	0.3529	0.0839	0.2475	0.100
Новостные источники	0.6973	0.8506	0.331	0.1643	0.2775	0.232
Эмбеддинги Google	0.6798	0.4877	0.375	0.0105	0.000	0.000

Таблица 4. Точность и полнота частотных классов с использованием разных предобученных эмбеддингов.

6. Анализ результатов

В этой главе будут описаны и проанализированы результаты работы третьей модели на материале корпуса СинТагРус, так как эта модель лучше распознала наиболее частотные классы, а также предсказала больше семантических ролей в целом, чем модель 4.

Всего в тестовом корпусе содержится 6941 предложение (117327 токенов). В результате применения третьей модели к данным корпуса было размечено 1879 ролей, в том числе 1351 предикат, 162 агентивных роли, 46 пациентивных, 78 субъектов перемещения, 77 говорящих, 36 субъектов психологического состояния, 70 тем. Сразу стоит отметить, что, очевидно, количество найденных предикатов сильно меньше их предполагаемого реального количества в рамках тестовой выборки. Более того, встречаются предложения, в которых модель распознала несколько предикатов, то есть количество неразмеченных предложений больше 5500.

Судя по результатам разметки, модель в большей мере подстраивается под лексику (эмбеддинги), чем под категориальные признаки. Эта тенденция заметна, например, если посмотреть на разброс синтаксических меток в пределах одного

класса. Так, роль предиката на материале корпуса присвоена токенам с синтаксической разметкой “0:root” (вершина предложения, корень), “conj” (конъюнкт), “xcomp”, “ccomp” (придаточная клауза), “advcl”, “acl” (клаузальный модификатор), среди которых есть достаточно неочевидные и комплексные конструкции. Более того, в корпусе есть примеры ошибочной разметки, которая может быть вызвана близкими по смыслу предикатами. В примере (1) аргумент “кондукторов” отмечен как причина, что, вероятно, вызвано близким значением предиката “бояться”, который действительно имел бы аргумент с ролью причины в качестве прямого дополнения.

(1) *Свое лицо он знал наизусть и никогда не пугал**[предикат]** остатками грима кондукторов**[причина]** последних ночных трамваев.*

Положительным результатом можно считать практически полное отсутствие клауз, в которых были бы найдены семантические роли аргументов при отсутствующей роли предиката. С другой стороны, большое количество предложений с размеченным предикатом не имеют никаких других ролей в предикатно-аргументной структуре.

Если рассматривать роль агенса — самую частотную роль в обучающем корпусе, то, судя по выходным данным, лучше всего модель предсказывает местоимения (*я, он, мы*) и имена собственные (см. пример (2)), особенно находящиеся в начале клаузы. Нужно отметить, что модель умеет размечать роль агенса и на относительных местоимениях (см. пример (3)).

(2) *Илья**[агенс]** Ильич вытер**[предикат]** платком лоб.*

(3) *Это подметили русские бражники, то есть беспутные люди, пропойцы, которые**[агенс]** в корчмах пьют.*

Существенную часть предсказанных классов составляют такие семантические роли, как “говорящий”, “субъект психологического состояния”, “содержание высказывания”. Высокое качество работы модели в отношении этих ролей может быть связано с их ярко выраженными лексическими особенностями, которые помогают модели научиться находить эти классы, основываясь на лемме предиката (вершины). Примеры (4), (5) и (6) иллюстрируют описанную положительную закономерность.

(4) *Именно ему я**[говорящий]** задал**[предикат]** главный интересующий меня вопрос.*

(5) *Потенциальных многодетных родителей**[субъект психологического состояния]** смущает**[предикат]** то**[причина]**, что деньги нельзя потратить на лечение ребенка или покупку садового участка.*

(6) *Прошлой осенью Юрий**[агенс]** Лужков призвал**[предикат]** вернуть**[содержание высказывания]** на Лубянку Железного Феликса.*

Пример (5) также доказывает меньшую значимость морфологических признаков в сравнении с лексической информацией, так как субъект психологического состояния модель определяет даже в роли синтаксического объекта.

Примечательно, что роль темы в большинстве примеров определяется также за счет леммы вершины — она практически везде находится в контексте предикатов “являться”, “наблюдать”, “появляться”, “остаться” и подобных им фазовых глаголов (см. примеры (7) и (8)).

(7) *Загородный дом можно сдать в аренду — и у вас появляется**[предикат]** весьма приличный доход**[тема]**.*

(8) *А тем времени вокруг проекта продолжались**[предикат]** научные и околонаучные споры**[тема]**.*

Следует также отметить, что роль пациенса присваивается аргументам, в конструкции которых уже размечен агенс (см. пример (9)) или которые являются частью пассивной конструкции без субъекта (см. пример (10)), и лишь изредка модель распознает пациенс, не маркируя агенс при его наличии.

(9) *Схватив работающие пилы, спортсмены**[агенс]** начинают очень быстро отрезать**[предикат]** у бревен**[пациенс]** ровные диски.*

(10) *От холода слона**[пациенс]** можно спасти**[предикат]**, если дать ему напиться густого красного вина.*

В целом, если говорить только о точности распознавания найденных аргументов, модель показывает удовлетворительные результаты, но в то же время увеличение показателя полноты и адаптация модели к количеству возможных ролей в пределах одной клаузы (сейчас соотношение ролей по предложениям

сильно варьируется) оказали бы положительное влияние на качество распознавания.

7. Заключение

В настоящем исследовании представлен метод распознавания семантических ролей для русского языка, основанный на нейронной сети с рекуррентным слоем долгой краткосрочной памяти (LSTM). На материале корпуса FrameBank сопоставлены и проанализированы четыре различных архитектуры многослойной модели. По результатам исследования частично доказана гипотеза о важности информации о вершине в формате отдельных признаков для повышения качества работы модели. Кроме того, использована и подтверждена идея разделения первого слоя на несколько разных входов по типу признаков (Shelmanov, Devyatkin 2017) и их дальнейшей раздельной обработки до конкатенации.

В работе также представлены различные методы оценки работы модели, адаптированные под характер задачи — большинство токенов не должны быть маркированы, то есть метрикам нужно ориентироваться только на “смысловые” семантические роли.

На тестовых данных проведен анализ ошибок и выявлены закономерности предсказания определенных ролей. Полнота разметки ролей достаточно низкая, во многих предложениях отсутствует предикатно-аргументная структура. С другой стороны, надежность распознавания предикатов высокая. Большинство специфичных ролей, имеющих яркие отличительные черты с точки зрения лексики (например, говорящий и субъект ментального состояния), модель предсказывает качественно. Нужно также отметить, что практически отсутствуют примеры конструкций, в которых были бы маркированы аргументы без размеченного предиката. В целом, если опираться только на те семантические роли, которые распознала модель, качество работы можно назвать удовлетворительным.

Дальнейшее развитие метода может идти в нескольких направлениях:

1. Применение механизма внимания (или “само-внимания”) может значительно улучшить качество модели, если адаптировать его под рассматриваемый формат данных и особенности предсказываемых классов.

Основная идея механизма заключается в том, чтобы сформировать матрицу весов важности и сообщать модели, на какие признаки стоит обратить больше внимания. В случае “само-внимания” (Self-attention) модель также учитывает другие слова входной последовательности на каждом шаге, то есть смотрит на контекст. Механизм зачастую дополняет такие нейронные сети, как RNN (рекуррентные нейронные сети), LSTM (сеть долгой краткосрочной памяти), GRU (Gated recurrent unit).

2. Подбор различных комбинаций признаков и параметров модели могут также привести к повышению точности работы сети за счет поиска и удаления не таких важных признаков или выделения при помощи весов особенно важной информации. Однако эти модификации, скорее всего, не смогут радикально поднять качество, но могут способствовать медленному росту в сочетании с другими механизмами.
3. Из-за малого объема обучающего корпуса модели может быть тяжело “выучить” малопредставленные роли, а также в целом значения признаков, характерные для тех или иных предикатно-аргументных конструкций. Вероятно, специализированная аугментация данных, учитывающая общий характер предложений и семантической разметки, облегчит задачу.

Разработанные модели могут быть полезны для разметки семантических ролей на стадии предобработки в задачах, ориентированных на семантические зависимости, таких как суммаризация текстов и диалоговые системы. Кроме того, полученный в результате работы корпус с синтаксическим и семантическим слоями разметки может быть использован в качестве материала для решения других NLP задач для русского языка.

Результаты распознавания ролей на внешних тестовых данных СинТагРус и код, обрабатывающий данные и запускающий модели, доступны по ссылке: https://github.com/kategavrishina/srl-modeling.

Список литературы

Гавришина 2020 — Гавришина Е. Эталонный корпус русского языка для универсального семантического парсинга. 2020. (рукопись).

Дяченко и др. 2015 — Дяченко П.В., Иомдин Л.Л., Лазурский А.В., Митюшин Л.Г., Подлесская О.Ю., Сизов В.Г., Фролова Т.И., Цинман Л.Л. Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) //*Сборник «Национальный корпус русского языка: 10 лет проекту». Труды Института русского языка им. В.В. Виноградова*. Москва, 2015. Вып. 6. С. 272–299.

Кашкин, Ляшевская 2013 — Кашкин Е. В., Ляшевская О. Н. Семантические роли и сеть конструкций в системе FrameBank *//Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции Диалог.* 2013. P. 297-311.

Маркова 2013 — Маркова М. В. Автоматическая семантическая разметка предложений английского языка : дис. 2013.

Abzianidze et al. 2017 — Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., Van Noord, R., Ludmann, P., Bos, J. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. //*arXiv preprint arXiv:1702.03964*. 2017.

Andor et al. 2016 — Andor D., Alberti C., Weiss D., Severyn A., Presta A., Ganchev K., Petrov S., and Collins M. Globally normalized transition-based neural networks //*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016. P. 2442–2452.

Anisimovich et al. 2012 — Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A. Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: //*Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*. 2012. P. 90–103.

Baker et. al 1998 — Baker C. F., Fillmore C. J., Lowe J. B. The berkeley framenet project //*Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1998. P. 86–90.

Bobrow et al. 2007 — Bobrow, D. G., Cheslow, B., Condoravdi, C., Karttunen, L., King, T. H., Nairn, R., Zaenen, A. PARC's bridge and question answering system. *Proc. of the GEAF 2007 Workshop. CSLI Studies in Computational Linguistics Online.* 2007.

Bos et al. 2017 — Bos, J., Basile, V., Evang, K., Venhuizen, N. J., & Bjerva, J. The groningen meaning bank. *Handbook of linguistic annotation.* 2017. P. 463-496.

Collobert et al. 2011 — Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch //*Journal of machine learning research.* V. 12, №. Aug, 2011. P. 2493–2537.

Das et al. 2014 — Das, D., Chen, D., Martins, A. F., Schneider, N., Smith, N. A. Frame-semantic parsing //*Computational linguistics*. V. 40, №. 1, 2014. P. 9–56.

Daza, Frank 2018 — Daza A., Frank A. A sequence-to-sequence model for semantic role labeling *//arXiv preprint arXiv:1807.03006*. 2018.

Devlin et al. 2018 — Devlin, J., Chang, M. W., Lee, K., Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding //*arXiv preprint arXiv:1810.04805.* 2018.

Droganova et al. 2018 — Droganova, K., Lyashevskaya, O., Zeman, D. Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks. //*Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), No. 155.* 2018. P. 52-65.

Fellbaum 1998 — Fellbaum C. A semantic network of English verbs //*WordNet: An electronic lexical database. T. 3.* 1998. P. 153-178.

Fillmore 1968 — Fillmore C. J. The case for case //*Universals in Linguistic Theory* / Ed. by Emmon Bach, Robert T. Harms. 1968. P. 1–88.

Foland, Martin 2015 — Foland, W., Martin, J. H. Dependency-based semantic role labeling using convolutional neural networks. *//Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics.* 2015. P. 279-288.

Fonseca. Rosa 2013 — Fonseca, E. R., Rosa, J. L. G. A two-step convolutional neural network approach for semantic role labeling. //*The 2013 International Joint Conference on Neural Networks (IJCNN)* 2013. P. 1-7. IEEE.

Gildea, Jurafsky 2002 — Gildea D., Jurafsky D. Automatic labeling of semantic roles //*Computational linguistics*. V. 28, №. 3, 2002. P. 245–288.

Giuseppe 2020 — Giuseppe C. A gradient boosting-Seq2Seq system for Latin PoS tagging and lemmatization //*Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages.* 2020. P. 119-123.

Guiglea, Moschitti 2006 — Guiglea, A.-M., Moschitti, A. Semantic role labeling via FrameNet,VerbNet and PropBank. //*Proceedings of Coling-ACL 2006.* 2006. P. 929–936.

Hovy et al. 2016 — Hovy E., Marcus M., Palmer M., Ramshaw L., Weischedel R. OntoNotes: the 90% solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers.* 2016. P. 57–60.

Jiang, Ng 2006 — Jiang, Z. P., & Ng, H. T. Semantic role labeling of NomBank: A maximum entropy approach. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* 2006. P. 138-145.

Johansson, Nugues 2008 — Johansson, R., Nugues, P. Dependency-based semantic role labeling of PropBank. //*Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.* 2008. P. 69-78.

Kingsbury et al. 2002 — Kingsbury, P., Palmer, M., Marcus, M. Adding semantic annotation to the penn treebank. //*Proceedings of the human language technology conference.* 2002. P. 252-256.

Kipper et al. 2006 — Kipper, K., Korhonen, A., Ryant, N., Palmer, M. Extending VerbNet with Novel Verb Classes. //*LREC.* 2006. P. 1027-1032.

Kutuzov, Andreev 2015 — Kutuzov A., Andreev I. Texts in, meaning out: neural language models in semantic similarity task for Russian //arXiv preprint arXiv:1504.08183. 2015.

Larionov et al. 2019 — Larionov, D., Shelmanov, A., Chistova, E., & Smirnov, I. Semantic role labeling with pretrained language models for known and unknown predicates. //*Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019).* 2019. P. 619-628.

Li et al. 2018 — Li, Z., Cai, J., He, S., Zhao, H. Seq2seq dependency parsing *//Proceedings of the 27th International Conference on Computational Linguistics.* 2018. P. 3203-3214.

Li et al. 2020 — Li, Z., Zhao, H., He, S., Cai, J. Syntax Role for Neural Semantic Role Labeling //*arXiv preprint arXiv:2009.05737*. 2020.

Lyashevskaya, Kashkin 2015 — Lyashevskaya O., Kashkin E. FrameBank: a database of Russian lexical constructions //International Conference on Analysis of Images, Social Networks and Texts. – Springer, Cham, 2015. – C. 350-360.

Marcheggiani, Titov 2017 — Marcheggiani, D., Titov, I. Encoding sentences with graph convolutional networks for semantic role labeling. //*arXiv preprint arXiv:1703.04826*. 2017.

Marquez et al. 2005 — Marquez, L., Comas, P., Giménez, J., Catala, N. Semantic role labeling as sequential tagging //*Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*. 2005. P. 193-196.

Meyers et al. 2004 — Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., Grishman, R. The NomBank project: An interim report. //*Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*. 2004. P. 24-31.

Moeller et al. 2020 — Moeller, S., Wagner, I., Palmer, M., Conger, K., Myers, S. The Russian PropBank //*Proceedings of The 12th Language Resources and Evaluation Conference.* 2020. P. 5995– 6002.

Palmer et al. 2005 — Palmer M., Gildea D., Kingsbury P. The proposition bank: An annotated corpus of semantic roles //*Computational linguistics.* V. 31, №. 1, 2005. P. 71–106.

Park 2019 — Park, J. Selectively Connected Self-Attentions for Semantic Role Labeling. //*Applied Sciences, 9*(8). 2019. P. 1716.

Peters et al. 2018 — Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. Deep contextualized word representations. *//arXiv preprint arXiv:1802.05365.* 2018.

Prokhorenkova et al. 2018 — Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. CatBoost: unbiased boosting with categorical features //*Advances in neural information processing systems.* 2018. P. 6638–6648.

Punyakanok et al. 2008 — Punyakanok V., Roth D., Yih W. The importance of syntactic parsing and inference in semantic role labeling //*Computational Linguistics*. V. 34, №. 2, 2008. P. 257–287.

Shelmanov, Devyatkin 2017 — Shelmanov A. O., Devyatkin D. A. Semantic role labeling with neural networks for texts in Russian //*Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" 2017*. V. 16, 2017. P. 245–256.

Shelmanov, Smirnov 2014 — Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts //*Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog*. V. 13, №. 20, 2014. P. 607–620.

Shi, Lin 2019 — Shi, P., Lin, J. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*. 2019.

Sokirko 2001 — Sokirko, A. A short description of Dialing Project. *Retrieved from http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html*. 2001.

Strubell et al. 2018 — Strubell, E., Verga, P., Andor, D., Weiss, D., McCallum, A. Linguistically-informed self-attention for semantic role labeling. //*arXiv preprint arXiv:1804.08199*. 2018.

Vaswani et al. 2017 — Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. Attention is all you need //*arXiv preprint arXiv:1706.03762*. 2017.

Xue, Palmer 2004 — Xue, N., Palmer, M. Calibrating Features for Semantic Role Labeling. //*EMNLP*. 2004. P. 88-94.

Yang et al. 2019 — Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Kurzweil, R. Multilingual universal sentence encoder for semantic retrieval //*arXiv preprint arXiv:1907.04307*. 2019.

Zhang et al. 2018 — Zhang M., Yu N., Fu G. A simple and effective neural model for joint word segmentation and POS tagging *//IEEE/ACM Transactions on Audio, Speech, and Language Processing. T. 26. №. 9*. 2018. P. 1528-1538.