УДК 004.41 + 004.02

АНАЛИЗ МОДЕЛЕЙ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ В ЗАДАЧЕ РАЗМЕТКИ СЕМАНТИЧЕСКИХ РОЛЕЙ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

Л. М. Кадермятова[1], Е. В. Тутубалина[2]

Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

[1]lkadermy@gmail.com, [2]ElVTutubalina@kpfu.ru

Аннотация

Изучено влияние использования векторных представлений слов на качество установления семантических ролей в русскоязычных текстах. Задача установления семантических ролей в русскоязычных текстах получила широкое распространение после выхода на свет корпуса FrameBank. Были исследованы модели векторных представлений слов word2vec, fastText и ELMo (Embeddings from Language Models). Анализировались метрики качества микро- и макро-F1 как оценочные показатели результатов автоматической разметки актантов. Был проведен ряд экспериментов, демонстрирующих, что модели ELMo, основанные на токенах предикатно-аргументных конструкций, показывают больший прирост качества по сравнению со всеми остальными моделями, в том числе, в сопоставлении с моделями ELMo, обученными на леммах, как по величине микро-F1, так и по величине макро-F1.

Ключевые слова: *машинное обучение, обработка естественного языка, векторные представления слов, семантические роли.*

ВВЕДЕНИЕ

Автоматическая разметка семантических ролей – одна из техник парсинга текстов на естественных языках, которая позволяет выделять предикаты и аргументы в структурах предложений. Предикаты поясняют основной смысл ситуации, происходящей в тексте. С точки зрения большинства семантических теорий, к предикатам относятся глаголы, отглагольные существительные и др. Аргументами являются выражения, которые поясняют ситуацию более подробно, раскрывают детали. Задача автоматической разметки ролей состоит в том, чтобы найти актантов, т. е. участников ситуации, описанной предикатом, и приписать им семантические роли. Семантический анализ находит широкое применение в различных задачах обработки естественного языка, включая вопросно-ответные системы [15, 16], извлечение информации [17], информационный поиск [18], машинный перевод [19] и др.

Данная статья является продолжением исследования [10], посвященного семантическому анализу русскоязычных текстов. Ранее в работе [10] был предложен подход к автоматической разметке семантических ролей в русскоязычных текстах корпуса FrameBank[1] [4], проведены эксперименты на различных моделях эмбеддингов.

В ходе текущей работы было добавлено 12 моделей векторных представлений слов, исследовано их влияние на качество автоматической разметки актантов на примере текстов корпуса FrameBank.

Был проведен ряд экспериментов, демонстрирующих, что модели ELMo, основанные на токенах предикатно-аргументных конструкций, показывают больший прирост качества по сравнению со всеми остальными моделями, в том числе, в сопоставлении с моделями ELMo, обученными на леммах, как по величине микро-F1, так и по величине макро-F1. Модели, основанные на векторных представлениях fastText, показывают в среднем лучшие результаты по отношению к моделям word2vec применительно к русскоязычному корпусу FrameBank.

1. ОБЗОР ЛИТЕРАТУРЫ

[1] https://github.com/olesar/framebank

Существует огромное количество работ, посвященных автоматической обработке английского языка (например, [1–3]), однако тема обработки актантов русского языка долгое время не исследовалась ввиду отсутствия корпуса тестовых и тренировочных данных. Задача установления семантических ролей в русскоязычных текстах получила широкое распространение после выхода в свет корпуса FrameBank. В отличие от проекта FrameNet [8], FrameBank опирается не на понятие фрейма, а на грамматику конструкций. Также FrameBank более сфокусирован на морфосинтаксических шаблонах, что обусловлено структурой русского языка.

В работе [20] обсуждены возможности применения корпуса FrameBank к задаче автоматической разметке семантических ролей в русскоязычных текстах, рассмотрены теоретические вопросы соотношения семантических классов глаголов, семантических ролей и семантических ограничений на заполнение валентностей.

В статье [21] обсуждены подходы к оценке парсеров для автоматической разметки актантов. Исследованы статистические критерии дистрибуции ролей в словаре конструкций и расположение ролей на графе для того, чтобы сопоставить ответ системы и ответ золотого стандарта.

И. Кузнецов написал диссертацию по семантическому анализу русскоязычных текстов [9]. В своем исследовании И. Кузнецов сделал вывод, что различие моделей с использованием только синтаксических свойств и комбинаций семантико-синтаксических свойств невелика. При учете форм глагола, лемм предиката, падежей и только синтаксических свойств доля корректно классифицированных объектов составляет 76.1%, при комбинации семантико-синтаксических свойств она возрастает до 76.4%.

В дальнейших проектах, в связи с развитием искусственного интеллекта, ученые начали широко применять машинное обучение и нейронные сети для задач автоматической разметки актантов [10, 13]. К достоинствам нейронных сетей можно отнести их возможность обработки низкоуровневых представлений и атомарных признаков слов, им не требуется множество характеристик предикатов и аргументов для определения ролей актантов.

В работе [10] нейронные сети используются для определения семантических ролей текстов корпуса FrameBank. Рассмотрены две модели – с «известными» и «неизвестными» предикатами, одна из них обучается на известных предикатах, в то время как другая модель обучается на векторных представлениях лемм предикатов. Согласно статье [10], модель с «неизвестными» предикатами показала лучшие результаты метрик микро- и макро-F1 по сравнению с моделью с «известными» предикатами на наборе с векторными представлениями слов, основанными на ELMo (показатель микро-F1 был выше на 10%, макро-F1 – на 8%).

2. РАЗМЕЧЕННЫЙ КОРПУС FRAMEBANK НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА (НКРЯ)[2]

FrameBank объединяет в себе словарь лексических конструкций русского языка и размеченный корпус их реализаций в текстах НКРЯ. Конструкции включают предикатно-аргументные структуры глаголов, существительных, прилагательных, наречий и предикативов, а также идиомы, в которых часть элементов фиксирована, а часть представляет собой переменные (т. н. конструкции «малого синтаксиса»).

Ядро системы FrameBank составляют 2200 частотных русских глаголов и ассоциированных с ними конструкций и корпусных примеров. Словарь русских глагольных конструкций представляет каждую конструкцию как шаблон, в котором указаны: морфосинтаксические характеристики элементов конструкции, синтаксический ранг участника, экспликация (роль) участника, семантические ограничения на заполнение слота конструкции.

Пример	Предикат	Аргумент	Семантическая роль
Продавец режет сыр	Режет	Продавец	Агенс
Продавец режет сыр	Режет	Сыр	Пациенс
Он говорит правду	Говорит	Он	Говорящий
На полу лежал человек	Лежал	На полу	Место

Таблица 1. Разметка семантических ролей на примерах корпуса FrameBank

3. МЕТОДИКА ИССЛЕДОВАНИЯ

[2] http://www.ruscorpora.ru/

Был использован предпроцессинг, описанный в статье [10]. Тексты корпуса FrameBank предварительно обрабатывались токенайзером, разделялись на предложения, анализировались POS-таггером, лемматизатором, синтаксическим парсером. В результате препроцессинга данных установлено, что корпус содержит 52751 конструкцию с 21 уникальными семантическими ролями. Количество предикатов уменьшилось с 803 до 643, поскольку для некоторых предикатов существовало менее 10 примеров.

Процесс автоматической разметки семантических ролей состоит из следующих этапов: идентификация предиката, извлечение аргументов, классификация аргументов (присвоение аргументам семантических ролей), глобальная оптимизация через методы целочисленного программирования. Следует отметить, что четвертый этап необходим для исключения присвоения одинаковых семантических ролей нескольким актантам одного предиката.

На шаге идентификации предиката все глаголы и отглагольные формы маркируются согласно POS-тегам токенов предложений. Отглагольные существительные не относятся к предикатам, поскольку они отсутствуют в корпусе FrameBank.

При определении аргументов внутри предложения каждому маркированному предикату сопоставляются соответствующие аргументы, анализируется дерево синтаксических зависимостей с заранее проставленными правилами. Аргументами являются одиночные токены (существительные, имена собственные, местоимения).

Базовые аргументы выделяются из синтаксического дерева в соответствии с правилами, учитывающими POS-теги токенов и прямые связи синтаксических зависимостей, имеющих корни-предикаты.

На этапе классификации аргументов происходит обучение нейронной модели с использованием библиотеки tensorflow[3]. В модель включаются леммы предикатов, векторные представления аргументов, а также лексические и морфосинтаксические признаки:

1. различные типы морфологических характеристик аргументов и предикатов: падеж, валентность, глагольная форма и т. д.;

[3] https://www.tensorflow.org/

2. соответствующая позиция аргумента в предложении по отношению к предикату;

3. предлог аргумента, извлеченного из синтаксического дерева в предложении относительно предиката;

4. синтаксическая связь, которая соединяет токен-аргумент с его родителем в синтаксическом дереве;

5. леммы аргументов и предикатов.

Входные параметры – (i) векторные представления предикатов, (ii) векторные представления аргументов, (iii) разреженные категориальные признаки по отдельности проходят первый слой активации – усеченное линейное преобразование ReLU (Rectified Linear Unit) [14]. Конканетированные выводы первого слоя затем проходят через другой слой ReLU и обрабатываются с помощью softmax-функции. Перед функцией активации, на каждом слое проходит нормализация по мини-батчам. Для совершенствования работы модели используются следующие слои: скрытый слой категориальных признаков размерности 400, скрытый слой векторных представлений слов размерности 100 и слой конкатенированных векторов размерности 400. Выбирается дропаут, равный 0.3. Результатом работы модели является вектор вероятностей для каждой семантической роли в перечне.

Четвертый этап глобальной оптимизации необходим для исключения присвоения одинаковых семантических ролей нескольким актантам одного предиката. В единичной предикатно-аргументной структуре каждая семантическая роль должна быть определена только единожды, и каждый аргумент должен иметь только одну роль.

Датасет был разделен на 2 части: 80% предикатов с примерами использовались для тренировки модели, 20% – для валидации.

Все эти шаги были проведены с использованием результатов работы [10], ресурса IsaNLP SRL FrameBank[4].

4. ОПРЕДЕЛЕНИЕ МЕТРИК КАЧЕСТВА МИКРО- И МАКРО-F1

[4] https://github.com/IINemo/isanlp_srl_framebank/

Точность отражает долю слов, которым была проставлена корректная семантическая роль относительно всех слов, которым система проставила данную семантическую роль.

Полнота системы есть доля слов, которым системой была проставлена корректная семантическая роль, относительно всех слов, которым должна была быть проставлена эта роль в тестовой выборке.

Мера F1 есть гармоническое среднее между точностью и полнотой.

Для подсчета микро-F1 точность и полнота усредняются по всем классам, а затем вычисляется итоговая метрика. При макро-усреднении сначала вычисляется итоговая метрика для каждого класса, а затем результаты усредняются по всем классам.

5. ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ В РАБОТЕ МОДЕЛЕЙ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ

Были использованы следующие дистрибутивно-семантические модели векторных представлений слов ресурсов RusVectores[5] [11] и DeepPavlov[6] [12]:

1. word2vec_nkrya_cbow_300d[7]: RusVectores, обучена на НКРЯ на 270 миллионах слов, алгоритм «непрерывного мешка слов», 300-мерная размерность векторов;

2. word2vec_nkrya_wiki_skipgram_300d[8]: RusVectores, обучена на НКРЯ и Википедии[9] за декабрь 2018 года на 788 миллионах слов, алгоритм skip-gram, 300-мерная размерность векторов;

3. word2vec_tayga_skipgram_300d[10]: RusVectores, обучена на веб-корпусе русского языка Тайга[11], снабженного морфологической и синтаксической разметкой на 5 миллиардах слов, алгоритм skip-gram, 300-мерная размерность векторов;

[5] http://rusvectores.org/
[6] https://deeppavlov.ai/
[7] http://vectors.nlpl.eu/repository/20/180.zip
[8] http://vectors.nlpl.eu/repository/20/182.zip
[9] https://ru.wikipedia.org/
[10] http://vectors.nlpl.eu/repository/20/185.zip
[11] http://www.webcorpora.ru/

4. fastText_tayga_300d[12]: RusVectores, обучена на веб-корпусе русского языка Тайга на 5 миллиардах слов, модель fastText, 300-мерная размерность векторов;

5. fastText_wiki_lenta_300d[13]: DeepPavlov, обучена на Википедии и новостном портале Lenta[14], модель fastText, 300-мерная размерность векторов;

6. fastText_twitter_300d[15]: DeepPavlov, обучена на русскоязычном портале Твиттер[16], модель fastText, 300-мерная размерность векторов;

7. elmo_nkrya_wiki18_tokens_1024d[17]: RusVectores, обучена на НКРЯ и Википедии за декабрь 2018 года на 788 миллионах токенов, модель ELMo, 1024-мерная размерность векторов;

8. elmo_nkrya_wiki18_lemmas_1024d[18]: RusVectores, обучена на НКРЯ и Википедии за декабрь 2018 года на 788 миллионах лемм, модель ELMo, 1024-мерная размерность векторов;

9. elmo_tayga_lemmas_2048d[19]: RusVectores, обучена на веб-корпусе русского языка Тайга на 5 миллиардах лемм, модель ELMo, 2048-мерная размерность векторов;

10. elmo_wiki_tokens_1024d[20]: DeepPavlov, обучена на Википедии на 386 миллионах токенов, модель ELMo, 1024-мерная размерность векторов;

11. elmo_wmtnews_tokens_1024d[21]: DeepPavlov, обучена на новостном параллельном корпусе WMT News[22] на 946 миллионах токенов, модель ELMo, 1024-мерная размерность векторов;

---

[12] http://vectors.nlpl.eu/repository/20/187.zip

[13] http://files.deeppavlov.ai/embeddings/ft_native_300_ru_wiki_lenta_nltk_word_tokenize.bin

[14] https://lenta.ru/

[15] http://files.deeppavlov.ai/embeddings/ft_native_300_ru_twitter_nltk_word_tokenize.bin

[16] https://twitter.com/?lang=ru

[17] http://vectors.nlpl.eu/repository/20/195.zip

[18] http://vectors.nlpl.eu/repository/20/196.zip

[19] http://vectors.nlpl.eu/repository/20/199.zip

[20] http://files.deeppavlov.ai/deeppavlov_data/elmo_ru-wiki_600k_steps.tar.gz

[21] http://files.deeppavlov.ai/deeppavlov_data/elmo_ru-news_wmt11-16_1.5M_steps.tar.gz

[22] http://www.statmt.org/

12. elmo_twitter_tokens_1024d[23]: DeepPavlov, обучена на русскоязычном портале Твиттер на 810 миллионах токенов, модель ELMo, 1024-мерная размерность векторов**.**

6. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Модель	Микро-F1	Макро-F1
word2vec_nkrya_cbow_300d	81.7	82.0
word2vec_nkrya_wiki_skipgram_300d	82.1	83.3
word2vec_tayga_skipgram_300d	83.3	82.9

Таблица 2. Показатели качества моделей, обученных на векторных представлениях слов word2vec

Качественные показатели метрики микро-F1 увеличиваются по мере перехода от одной модели word2vec к другой. На основе данных результатов можно сделать предположение о том, что на это повлияло количество примеров в корпусах. Также можно отметить, что процентные соотношения метрик микро-F1 растут пропорционально количеству слов в корпусах. К примеру, разница в показателе микро-F1 между word2vec_nkrya_cbow_300d и word2vec_nkrya_wiki_skipgram_300d составляет 0.4% при увеличении количества слов с 270 до 788 миллиона, в то же время разность между word2vec_nkrya_wiki_skipgram_300d и word2vec_tayga_skipgram_ 300d уже составляет 1.3% при росте количества слов с 788 миллионов до 5 миллиардов.

Соответствующие значения таблицы 2 также демонстрируют, что модели word2vec, обученные с помощью алгоритма skip-gram, имеют наилучшие результаты по сравнению с моделью word2vec, обученной на алгоритме непрерывного мешка слов. Архитектура skip-gram использует текущее слово, чтобы определить окружающие его слова. Алгоритм непрерывного мешка слов предсказывает текущее слово, исходя из контекста. Вероятно, в тестовых данных редко встречались

[23]http://files.deeppavlov.ai/deeppavlov_data/elmo_ru-twitter_2013-01_2018-04_600k_steps.tar. gz

повторяющиеся слова, поэтому модель, обученная на skip-gram, показала себя лучше.

Модель	Микро-F1	Макро-F1
fastText_tayga_300d	82.9	82.3
fastText_wiki_lenta_300d	84.0	84.7
fastText_twitter_300d	83.3	83.3

Таблица 3. Показатели качества моделей, обученных на векторных представлениях слов fastText

В наших экспериментах модели векторных представлений слов fastText ресурса DeepPavlov получили лучшие результаты по сравнению с моделью векторных представлений слов fastText RusVectores. Модель fastText_twitter_300d превосходит модель fastText_tayga_300d по метрике макро-F1 на 1%, а fastText_wiki_lenta_300d – на 2.4%.

Рассмотрим модели векторных представлений слов fastText от DeepPavlov и сравним их друг с другом. Модель fastText_wiki_lenta_300d имеет лучшие показатели метрик микро- и макро-F1 в сопоставлении с моделью fastText_twitter_300d. Метрика микро-F1 выше на 0.7%, макро-F1 – на 1.4%. Большая часть текстов fastText_wiki_lenta_300d относится к научному и публицистическому стилям, модель fastText_twitter_300d предназначена для применения в текстах разговорного стиля, в свою очередь, FrameBank основан на данных НКРЯ, что объясняет полученные результаты.

Анализируя значения таблиц 2 и 3, можно сделать вывод, что модели, основанные на векторных представлениях слов fastText, имеют средние значения показателей микро- и макро-F1 выше, чем модели word2vec. Модель fastText учитывает символьные n-граммы, то есть подстроки фиксированной длины, что играет немаловажную роль при обучении русскоязычных корпусов.

Модель	Микро-F1	Макро-F1
elmo_nkrya_wiki18_tokens_1024d	85.8	86.0
elmo_nkrya_wiki18_lemmas_1024d	81.0	81.2

elmo_tayga_lemmas_2048d	81.7	81.2
elmo_wiki_tokens_1024d	86.0	86.9
elmo_wmtnews_tokens_1024d	87.2	87.7
elmo_twitter_tokens_1024d	86.9	86.8

Таблица 4. Показатели качества моделей, обученных на векторных представлениях слов ELMo

Рассматривая данные таблицы 4, нельзя не отметить сильную дифференциацию показателей метрик качества микро- и макро- F1 моделей векторных представлений ELMo, на вход которым подавались леммы (elmo_nkrya_wiki18_lemmas_1024d, elmo_tayga_lemmas_2048d) и токены (elmo_nkrya_wiki18_tokens_1024d, elmo_wiki_tokens_1024d, elmo_wmtnews_tokens_ 1024d, elmo_twitter_tokens_1024d). На примере модели ELMo можно сделать вывод, что векторные представления токенов предикатно-аргументных структур имеют лучшие показатели качества в сопоставлении с системами, основанными на леммах применительно к корпусу FrameBank. Вероятно, такие результаты объясняются разнообразием морфем, морфологических форм слова в русском языке.

Наивысшие значения метрик качества микро- и макро-F1 среди моделей векторных представлений слов ELMo, на вход которым подавались токены, получила модель elmo_wmtnews_tokens_1024d ресурса DeepPavlov. На основе данных результатов можно сделать предположение, что на это повлияло количество примеров в корпусах. Модель elmo_wmtnews_tokens_1024d содержит наибольшое количество примеров (946 миллионов). Несмотря на относительную близость количества токенов в моделях elmo_wmtnews_tokens_1024d (946 миллинов токенов) и elmo_twitter_tokens_1024d (810 миллионов токенов), модель elmo_wmtnews_tokens_1024d продемонстрировала результаты лучше (разница в показателе микро-F1 составляет 0.3%, макро-F1 – 0.9%). Это обусловлено спецификой корпусов и стилями речи их текстов. Тексты корпуса, взятого с портала Твиттер, относятся к разговорному стилю, WMT News – к научно-публицистическому. Тестируемый корпус FrameBank был основан на НКРЯ, что объясняет полученные результаты.

ЗАКЛЮЧЕНИЕ

В статье рассмотрены различные модели векторных представлений слов, проанализировано влияние использования той или иной модели эмбеддингов на качество автоматической разметки семантических ролей в русскоязычных актантах. Был проведен ряд экспериментов, демонстрирующих, что модели ELMo, основанные на токенах предикатно-аргументных конструкций, показывают больший прирост качества по сравнению со всеми остальными моделями, в том числе, в сопоставлении с моделями ELMo, обученными на леммах, как по величине микро-F1, так и по величине макро-F1.

Исследование выполнено за счет гранта Российского научного фонда (проект № 19-71-10056).

СПИСОК ЛИТЕРАТУРЫ

1. Christensen J., Mausam, Soderland S., and Etzioni O. (2011), An analysis of open
information extraction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture, pp. 113–120.

2. Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005), Ann Arbor, MI.

3. Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 473–483.

4. Olga Lyashevskaya and Egor Kashkin. 2015. Framebank: a database of russian lexical constructions. In International Conference on Analysis of Images, Social Networks and Texts, pages 350–360.

5. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

6. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

7. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2227–2237.

8. Baker C. F., Fillmore C. J., and Lowe J. B. (1998), The Berkeley FrameNet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Volume 1, pp. 86–90.

9. Ilya Kuznetsov. 2016. Automatic semantic role labelling in Russian language, PhD thesis (in Russian). Ph.D. thesis, Higher School of Economics.

10. Shelmanov A., Smirnov I., Larionov D., Chistova E. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates // Proceedings of Recent Advances in Natural Language Processing, pages 619–628, Varna, Bulgaria, Sep 2–4, 2019.

11. Andrey Kutuzov and Elizaveta Kuzmenko, 2017. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, pages 155–161. Springer.

12. Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialoguesystems. In Proceedings of ACL 2018, System Demonstrations, pages 122–127.

13. Shelmanov A., Devyatkin D. Semantic role labeling with neural networks for texts in Russian // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2017). — Vol. 1. — 2017. — P. 245–256.

14. Agarap, A. F. 2018. Deep Learning using Rectified Linear Units (ReLU), Neural and Evolutionary Computing, Vol. 1.

15. Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP 2015), pages 643–653, 2015.

16. Wen Tau Yih, Matthew Richardson, Chris Meek, Ming Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pages 201–206, 2016.

17. Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading. Association for Computational Linguistics, Los Angeles, California, pages 52–60.

18. GS Osipov, IV Smirnov, and IA Tikhomirov. 2010. Relational-situational method for text search and analysis and its applications. Scientific and Technical Information Processing, 37(6):432–437.

19. Liu, D., Gildea, D., 2010. Semantic role features for machine translation. Proc. 23rd Int. Conf. on Computational Linguistics, p.716–724.

20. Kashkin, E.V., Lyashevskaya, O.N.: Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set' konstrukcij v sisteme FrameBank] (in Russian). In: Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog", vol. 12-1, pp. 297–311. RSUH, Moscow (2013)

21. Lyashevskaya O. N., Kashkin E. V. Evaluation of frame-semantic role labeling in a case-marking language // Papers from the Annual International Conference "Dialogue" (2014). — 2014. — P. 350–365.

ANALYSIS OF WORD EMBEDDINGS FOR SEMANTIC ROLE LABELING OF RUSSIAN TEXTS

Leysan Kadermyatova[1], Elena Tutubalina[2]

Higher Institute of Information Technology and Intelligent Systems, Kazan Federal University

[1]lkadermy@gmail.com, [2]ElVTutubalina@kpfu.ru

Abstract

Currently, there are a huge number of works dedicated to semantic role labeling of English texts [1–3]. However, semantic role labeling of Russian texts was an unexplored area for many years due to the lack of train and test corpora. Semantic role labeling of Russian Texts was widely disseminated after the appearance of the FrameBank corpus [4]. In this approach, we analyzed the influence of the word embedding models on the quality of semantic role labeling of Russian texts. Micro- and macro- F1 scores on word2vec [5], fastText [6], ELMo [7] embedding models were calculated. The set of experiments have shown that fastText models averaged slightly better than word2vec models as applied to Russian FrameBank corpus. The higher micro- and macro- F1 scores were obtained on deep tokenized word representation model ELMo in relation to classical shallow embedding models.

Keywords: *machine learning, ML-model, natural language processing, word embedding, semantic role labeling.*

REFERENCES

1. Christensen J., Mausam, Soderland S., and Etzioni O. (2011), An analysis of open
information extraction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture, pp. 113–120.

2. Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005), Ann Arbor, MI.

3. Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 473–483.

4. Olga Lyashevskaya and Egor Kashkin. 2015. Framebank: a database of russian lexical constructions. In International Conference on Analysis of Images, Social Networks and Texts, pages 350–360.

5. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

6. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

7. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2227–2237.

8. Baker C. F., Fillmore C. J., and Lowe J. B. (1998), The Berkeley FrameNet project. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Volume 1, pp. 86–90.

9. Ilya Kuznetsov. 2016. Automatic semantic role labelling in Russian language, PhD thesis (in Russian). Ph.D. thesis, Higher School of Economics.

10. Shelmanov A., Smirnov I., Larionov D., Chistova E. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates // Proceedings of Recent Advances in Natural Language Processing, pages 619–628, Varna, Bulgaria, Sep 2–4, 2019.

11. Andrey Kutuzov and Elizaveta Kuzmenko, 2017. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models, pages 155–161. Springer.

12. Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialoguesystems. In Proceedings of ACL 2018, System Demonstrations, pages 122–127.

13. Shelmanov A., Devyatkin D. Semantic role labeling with neural networks for texts in Russian // Computational Linguistics and Intellectual Technologies. Papers from

the Annual International Conference "Dialogue" (2017). — Vol. 1. — 2017. — P. 245–256.

14. Agarap, A. F. 2018. Deep Learning using Rectified Linear Units (ReLU), Neural and Evolutionary Computing, Vol. 1.

15. Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP 2015), pages 643–653, 2015.

16. Wen Tau Yih, Matthew Richardson, Chris Meek, Ming Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pages 201–206, 2016.

17. Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading. Association for Computational Linguistics, Los Angeles, California, pages 52–60.

18. GS Osipov, IV Smirnov, and IA Tikhomirov. 2010. Relational-situational method for text search and analysis and its applications. Scientific and Technical Information Processing, 37(6):432–437.

19. Liu, D., Gildea, D., 2010. Semantic role features for machine translation. Proc. 23rd Int. Conf. on Computational Linguistics, p.716–724.

20. Kashkin, E.V., Lyashevskaya, O.N.: Semantic roles and construction net in Russian FrameBank [Semanticheskie roli i set' konstrukcij v sisteme FrameBank] (in Russian). In: Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog", vol. 12-1, pp. 297–311. RSUH, Moscow (2013)

21. Lyashevskaya O. N., Kashkin E. V. Evaluation of frame-semantic role labeling in a case-marking language // Papers from the Annual International Conference "Dialogue" (2014). — 2014. — P. 350–365.

СВЕДЕНИЯ ОБ АВТОРАХ

КАДЕРМЯТОВА Лейсан Маратовна – магистрант Высшей школы информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, инженер по тестированию программного обеспечения.

Leysan Maratovna Kadermyatova – postgraduate student of the Higher School of Information Technologies and Intelligent Systems at Kazan Federal University, QA Engineer.

email: lkadermy@gmail.com

ТУТУБАЛИНА Елена Викторовна – кандидат физико-математических наук, старший научный сотрудник Высшей школы информационных технологий и интеллектуальных систем Казанского федерального университета. Сфера научных интересов – машинное обучение, обработка естественного языка, медицинская информатика

Elena Victorovna TUTUBALINA – candidate of physico-mathematical sciences, senior researcher of the Higher School of Information Technologies and Intelligent Systems at Kazan Federal University. Research interests include natural language processing, machine learning, medical informatics.

email: ElVTutubalina@kpfu.ru

Материал поступил в редакцию 2 апреля 2020 года