Моделирование и анализ информационных систем. Т. 25, №6 (2018), с. 726–733
Modeling and Analysis of Information Systems. Vol. 25, No 6 (2018), pp. 726–733

Тезаурусы
Thesauri

DOI: 10.18255/1818-1015-2018-6-726-733

УДК 004.912

Векторное представление слов с семантическими отношениями: экспериментальные наблюдения

Каряева М. С.[1], **Браславский П. И.**, **Соколов В. А.**[1]

Поступила в редакцию 1 сентября 2018
После доработки 20 ноября 2018
Принята к публикации 25 ноября 2018

Аннотация. Возможность идентификации семантической близости между словами сделала модель word2vec широко используемой в NLP-задачах. Идея word2vec основана на контекстной близости слов. Каждое слово может быть представлено в виде вектора, близкие координаты векторов могут быть интерпретированы как близкие по смыслу слова. Таким образом, извлечение семантических отношений (отношение синонимии, родо-видовые отношения и другие) может быть автоматизировано. Установление семантических отношений вручную считается трудоемкой и необъективной задачей, требующей большого количества времени и привлечения экспертов. Но среди ассоциативных слов, сформированных с использованием модели word2vec, встречаются слова, не представляющие никаких отношений с главным словом, для которого был представлен ассоциативный ряд. В работе рассматриваются дополнительные критерии, которые могут быть применимы для решения данной проблемы. Наблюдения и проведенные эксперименты с общеизвестными характеристиками, такими как частота слов, позиция в ассоциативном ряду, могут быть использованы для улучшения результатов при работе с векторным представлением слов в части определения семантических отношений для русского языка. В экспериментах используется обученная на корпусах Флибусты модель word2vec и размеченные данные Викисловаря в качестве образцовых примеров, в которых отражены семантические отношения. Семантически связанные слова (или термины) нашли свое применение в тезаурусах, онтологиях, интеллектуальных системах для обработки естественного языка.

Ключевые слова: векторное представление слов, word2vec, семантические отношения, тезаурус, гипонимы, гиперонимы, синонимы

Для цитирования: Каряева М. С., Браславский П. И. Соколов В. А., "Векторное представление слов с семантическими отношениями: экспериментальные наблюдения", *Моделирование и анализ информационных систем*, **25**:6 (2018), 726–733.

Об авторах: Каряева Мария Сергеевна, orcid.org/0000-0003-4466-1735, аспирант
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: mari.karyaeva@gmail.com

Браславский Павел Исаакович, orcid.org/0000-0002-6964-458X, канд. техн. наук, доцент,
Уральский федеральный университет,
г. Екатеринбург, ул. Мира, 19, 620002 Россия, e-mail: pbras@yandex.ru

Соколов Валерий Анатольевич, orcid.org/0000-0003-1427-4937, доктор физ.-мат. наук, профессор,
Ярославский государственный университет им. П.Г. Демидова,
ул. Советская, 14, г. Ярославль, 150003 Россия, e-mail: sokolov@uniyar.ac.ru

Благодарности:
[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов №16-07-01180 и №16-06-00497

Введение

Изменение метода распознания семантических ролей на основе нейронной сети [1], разработанного в 2013 году Томасом Миколовым, привело к разработке подхода векторного представления слов word2vec[1], который является на сегодняшний день одним из самых распространенных методов семантического моделирования при работе с текстовой информацией. Векторное представление слов применяется в широком спектре задач [2], [3], [4], [5] обработки естественного языка. На вход алгоритму word2vec подается корпус текстов, после обучения модели с определенными параметрами на выходе формируются векторные представления слов, которые отражают их семантику. Для оценки полученных пар часто используют метрику – косинусная мера близости, которая является оптимальной для отображения семантического сходства слов.

Представление слов в виде векторов позволяет применять математические операции. В большинстве примеров можно встретить вычитание векторов, когда результат вычисления vec('Madrid') - vec('Spain') + vec('France') будет ближе к vec('Paris'), чем к другим векторам из распределения. Таким образом, разница векторов может быть использована для поиска семантических отношений между словами.

Word2vec не возвращает напрямую семантические отношения между словами. В ассоциативном ряду, который может быть возвращен в качестве близких слов к запрашиваемому (главному) слову, отражаются слова, которые часто употребляются рядом в контексте. Бесспорно, в ассоциативном ряду встречаются синонимы, антонимы, гипонимы, гиперонимы, холонимы, меронимы, ассоциации и другие типы, которые могут быть определены как семантические отношения.

Наиболее популярным и эффективным методом для извлечения семантических отношений были [7] и остаются [8], [9], [10] лексико-синтаксические шаблоны. Наша работа не перекрывает полученные результаты использованием векторных представлений слов, в данном исследовании мы стремимся поделиться наблюдениями и закономерностями, которые могут быть полезны при работе с word2vec.

1. Связанные работы

Обучение модели с использованием word2vec строится по принципу наличия семантических отношений между словами в схожих контекстах. Получение связанных слов с оценкой может послужить базой для поиска связей между терминами при автоматическом создании тезауруса.

Исследование [6], посвященное задачам семантической близости, показало, что морфология русского языка не является препятствием для обучения модели с использованием word2vec на русских корпусах, в частности был рассмотрен Национальный корпус русского языка (НКРЯ).

В основу исследования [12] положена идея поиска пар гипоним-гипероним на основе векторного представления: сначала разности векторов кластеризуются, после этого для каждого кластера обучается отдельная проекция вектора на основе обуча-

[1] https://code.google.com/p/word2vec/

ющей выборки, полученной из тезауруса. В результате пара слов, соответствующих векторам, может быть классифицирована как “род-вид”. Данная работа послужила фундаментом для поиска семантических отношений с применением word2vec. В работе [13] дополнительно к применению концепции [12] было изучено негативное влияние примеров на прогнозирование наличия родо-видовых отношений.

Кроме родо-видовых отношений, с помощью word2vec возможно извлечение синонимии. В работе [14] были проведены эксперименты по автоматическому поиску синонимов в сфере медицины. Данное исследование не только демонстрирует жизнеспособность извлечения синонимов с использованием word2vec, но и подтверждает применимость векторного представления слов в предметных областях.

2. Векторное представление слов

Примеры ассоциативных рядов (выборка) с указанием метрики близости представлены в Таблице 1. Полужирным шрифтом выделены слова-ассоциаты, которые связаны с главным словом отношением род-вид. Среди слов-ассоциатов встречаются слова, характеризующие главное слово: голос хрипловатый, спокойный и т.д. Таким образом, существует вероятность благоприятного результата при использовании векторного представления слов для извлечения пар-кандидатов, имеющих семантические отношения с главным словом, в данном случае родо-видовые.

Таблица 1. Примеры векторных представлений
с оценкой близости (косинусная мера)
Table 1. Some examples of pairs with semantic relations
and their scores of cosine similarity

голос		оружие		лук	
хрипловатый	0.80	огнестрельное	0.71	луком	0.72
голосок	0.76	стрелковое	0.66	**порей**	0.65
звучал	0.72	**ружье**	0.62	колчан	0.65
баритон	0.72	**пистолет**	0.61	**репчатый**	0.62
бас	0.64	метательное	0.61	чеснок	0.59
женский	0.61	**копье**	0.58	шинкованный	0.58
спокойный	0.56	**лучеметы**	0.52	зелень	0.57

Кандидат из ассоциативного ряда с главным словом, для которого был автоматически сгенерирован ассоциативный ряд, не является симметричным относительно позиции кандидата в ассоциативном ряду. Другими словами, если заранее предположить, что между X и Y есть семантические отношения, а затем для слова X сгенерировать ассоциативный ряд и найти в нем позицию слова Y, то данная позиция не будет равна той позиции, которая установлена при проведении поиска кандидата X в ассоциативном ряду, построенном для Y как главного слова.

Власть – диктатура (31-е место в ассоциативном ряду),
Диктатура – власть (6-е место в ассоциативном ряду).

Следовательно, при поиске семантических несимметричных отношений необходимо определить вертикаль главного слова и кандидата. Например, для родовидовых отношений определить – гипонимом или гиперонимом будет являться главное слово.

В данной работе для экспериментов с родо-видовыми отношениями главное слово является видом, а искомый кандидат ассоциативного ряда – родом.

3. Данные

3.1. Флибуста

Обучение модели word2vec на корпусе Флибусты показало следующие результаты: сгенерировано 931 896 векторных представлений в виде главного слова и слов-ассоциатов с оценкой близости, которая представлена косинусной мерой между векторами главного слова и рассматриваемого ассоциата. Среди слов-ассоциатов встречаются как слова окружения, характеризующие главное слово, так и слова, имеющие семантические отношения с главным словом.

3.2. Викисловарь

Викисловарь[2] (общий объем – более 173 тыс. словарных входов) как лексикографический проект с прямым указанием семантических свойств слов (в том числе ссылками на гипонимы и гиперонимы) подходит для тестирования данных и получения автоматической оценки извлеченных пар-кандидатов. Для проведения экспериментов использовался дамп Викисловаря, содержащий 59 582 родо-видовых пар.

3.3. НКРЯ

Частотные списки Национального корпуса русского языка[3] были использованы для определения частоты встречаемости слов.

4. Экспериментальные результаты

4.1. Определение границы

В работе [15] приводится сравнительный анализ метрик близости с целью векторного представления слов для поиска семантических отношений. Наилучший результат показала косинусная мера близости. Изучение результатов распределения косинусной меры близости может быть полезно с точки зрения сужения границ ассоциативного ряда для увеличения вероятности нахождения кандидата родо-видовых отношений. В данном эксперименте в качестве данных были использованы пары из Википедии с родо-видовыми отношениями, для них были сформированы векторные

[2] https://ru.wiktionary.org/
[3] http://www.ruscorpora.ru/

представления и определена косинусная мера близости. Главное слово представлено видом, а слово из ассоциативного списка – родом.

На рис. 1 изображено распределение, позволяющее детектировать границы, в которых могут быть найдены семантические отношения. Анализируя распределение можно с уверенностью сказать, что вероятность встретить кандидата в ассоциативном ряду с высокой или низкой косинусной мерой близости очень низка. Основная концентрация потенциальных кандидатов будет расположена в пределах 0.53–0.63.

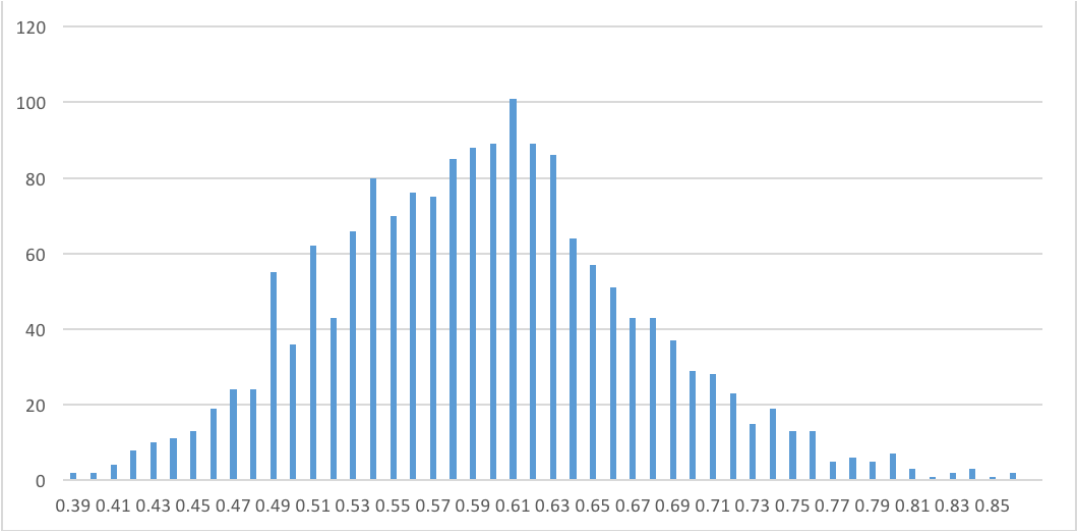

Рис. 1. Распределение значения косинусной меры
по парам вид-род из Викисловаря (1 688 пар)
Fig. 1. Cosine similarity score distribution for hyponym/hypernym pairs
(1 688 pairs) from Wiktionary

4.2. Ассоциативный ряд

Был проведен эксперимент, который показал, что *расстояние от главного слова до слова-кандидата является наименьшим, если главное слово представлено видом, а слово-кандидат – родом.* Данное расстояние измеряется не косинусной мерой, а количеством слов-ассоциатов, упорядоченных по косинусной мере и стоящих выше потенциального кандидата.

Для примера рассмотрим пару *мясо-конина.* Для слова *мясо* составим пронумерованный ряд слов-ассоциатов в порядке убывания косинусной меры. Таким образом, в ряду слов-ассоциатов *конина* окажется на 64-м месте. Аналогично для главного слова *конина* в ряду слов-ассоциатов найдем *мясо*, которое оказалось на 5-й позиции от главного слова.

Мясо – жареное, . . . , говядина, баранина, вареное, копченое, . . . , нарезанное, зажаренное, вяленое, курятина, полусырое, . . . , ягнятина, . . . , конина, . . .

Конина – говядина, свинина, баранина, конину, мясо, козлятина, . . .

Для подтверждения гипотезы был проведен следующий эксперимент. Имена собственные были удалены из пар дампа Викисловаря, таким образом, работа производилась с 39,5 тыс. пар. В результате поиска данных пар в векторном представлении были получены 1 688 пары. Из них:

- 52% пар, где позиция рода в списке слов-ассоциатов меньше позиции вида в списке слов-ассоциатов;
- 36% пар, где позиция рода в списке слов-ассоциатов больше позиции вида в списке слов-ассоциатов;
- 12% пар, где позиция рода в списке слов-ассоциатов равна позиции вида в списке слов-ассоциатов.

Результат второго места (36%) можно объяснить наличием вида с высокой частотностью, как, например, в паре *'мясо-говядина'*. Тем не менее, таких часто встречающихся пар оказывается не так много.

4.3. Частотность слов

Информация о частоте встречаемости слов является одной из ключевых характеристик в задачах компьютерной лингвистики. Опираясь на успешный опыт применения статистики использования слов [16], [17], мы рассмотрели зависимость – частотность рода и вида среди пар Викисловаря, которые имеют векторное представление. Примеры пар род-вид с частотностью:

Игра – 5 815*, прятки – 306, жмурки – 122.*
Мясо – 4 900*, говядина – 231, свинина – 157, конина – 29, курятина – 25.*

88% составляют пары, где частотность рода больше, чем частотность вида, например: *песня (29 220) – баллада (900)*. Остальные 12% составляют пары, где частотность вида больше, чем частотность рода. Такие пары можно отнести к исключениям, поскольку они отображают: слабую родо-видовую связь: *директор (27 442) – начальник (40 352)*; смещение рода к виду или вида к роду: *помещение (14 659) – коридор (17 507)*

На рис. 2 приведены графики распределения частоты от встречаемости отдельно для рода и вида.

5. Заключение

Работа с векторным представлением слов предоставляет огромный спектр исследований как с точки зрения лингвистики, так и статистики. Поиск закономерностей и эвристик, полученных таким образом, позволит улучшать качество извлекаемых сущностей. Представленные в данной работе заключения могут служить дополнительными критериями для устранения избыточных кандидатов при извлечении семантических отношений.

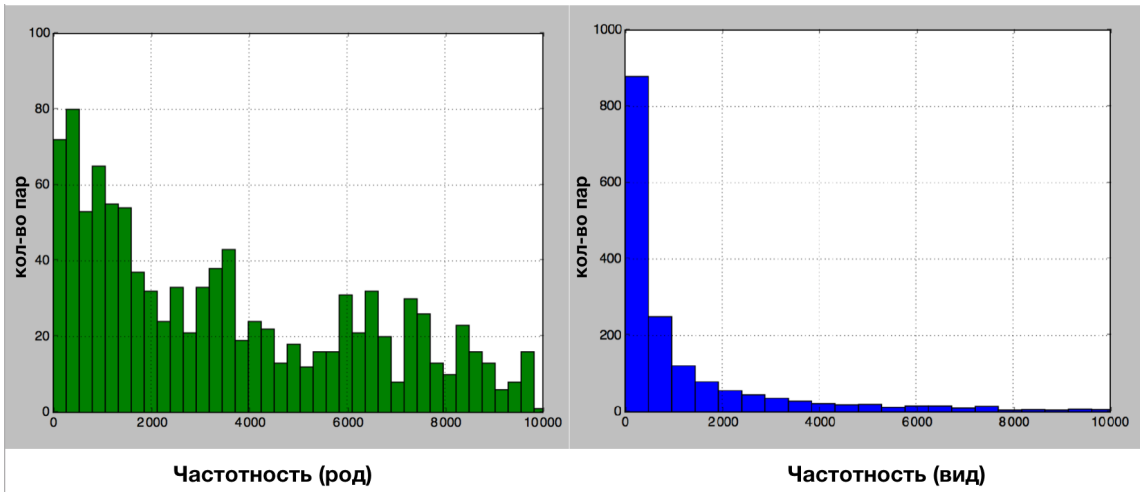

Рис. 2. Распределение частоты от встречаемости (количества пар)
Fig. 2. Word frequency distribution and their occurrence

Список литературы / References

[1] Mikolov T., Yih W., Zweig G., "Linguistic Regularities in Continuous Space Word Representations", *HLT-NAACL*, 2013, 746–751.

[2] Sienčnik S.K., "Adapting word2vec to named entity recognition", *Proceedings of the 20th nordic conference of computational linguistics*, 2015, 239–243.

[3] Lilleberg J., Zhu Y., Zhang Y., "Support vector machines and word2vec for text classification with semantic features", *Cognitive Informatics & Cognitive Computing*, IEEE 14th International Conference, 2015, 136–140.

[4] Ling W. et al., "Two/too simple adaptations of word2vec for syntax problems", *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, 1299–1304.

[5] Najafabadi M.M. et al., "Deep learning applications and challenges in big data analytics", *Journal of Big Data*, **2** (2015), 1.

[6] Kutuzov A., Andreev I., "Texts in, meaning out: neural language models in semantic similarity task for Russian", 2015, https://arxiv.org/abs/1504.08183.

[7] Hearst M. A., "Automatic acquisition of hyponyms from large text corpora", *Proceedings of the 14th conference on Computational linguistics – Association for Computational Linguistics*, **2** (1992), 539–545.

[8] Klaussner C., Zhekova D., "Lexico-syntactic patterns for automatic ontology building", *Proceedings of the Second Student Research Workshop associated with RANLP*, 2011, 109–114.

[9] Maedche A., Pekar V., Staab S., "Ontology learning part one—on discovering taxonomic relations from the web", *Web Intelligence*, 2003, 301–319.

[10] Snow R., Jurafsky D., Ng A. Y., "Learning syntactic patterns for automatic hypernym discovery", *Advances in Neural Information Processing Systems*, 2005, 1297–1304.

[11] Panchenko A., et al., "Human and Machine Judgements for Russian Semantic Relatedness", *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016*, (Yekaterinburg, Russia, April 7–9, 2016, Revised Selected Papers), 2017, 221–235.

[12] Fu R., et al., "Learning semantic hierarchies via word embeddings", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, **1** (2014), 1199–1209.

[13] Ustalov D., Arefyev N., Biemann C., Panchenko A., "Negative sampling improves hypernymy extraction based on projection learning", 2017, https://arxiv.org/pdf/1707.03903.pdf.

[14] Wang C., Cao L., Zhou B., “Medical Synonym Extraction with Concept Space Models”, 2015, https://arxiv.org/pdf/1506.00528.pdf.

[15] Rei M., Briscoe T., “Looking for hyponyms in vector space”, *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014, 68–77.

[16] Turney P., Pantel P., “From frequency to meaning: Vector space models of semantics”, *Journal of artificial intelligence research*, **37** (2010), 141–188.

[17] Matsuo Y., Ishizuka M., “Keyword extraction from a single document using word co-occurrence statistical information”, *International Journal on Artificial Intelligence Tools*, **13**:1 (2004), 157–169.

Karyaeva M. S., Braslavski P. I., Sokolov V. A., "Word Embedding for Semantically Relative Words: an Experimental Study", *Modeling and Analysis of Information Systems*, **25**:6 (2018), 726–733.

DOI: 10.18255/1818-1015-2018-6-726-733

Abstract. The ability to identify semantic relations between words has made a word2vec model widely used in NLP tasks. The idea of word2vec is based on a simple rule that a higher similarity can be reached if two words have a similar context. Each word can be represented as a vector, so the closest coordinates of vectors can be interpreted as similar words. It allows to establish semantic relations (synonymy, relations of hypernymy and hyponymy and other semantic relations) by applying an automatic extraction. The extraction of semantic relations by hand is considered as a time-consuming and biased task, requiring a large amount of time and some help of experts. Unfortunately, the word2vec model provides an associative list of words which does not consist of relative words only. In this paper, we show some additional criteria that may be applicable to solve this problem. Observations and experiments with well-known characteristics, such as word frequency, a position in an associative list, might be useful for improving results for the task of extraction of semantic relations for the Russian language by using word embedding. In the experiments, the word2vec model trained on the Flibusta and pairs from Wiktionary are used as examples with semantic relationships. Semantically related words are applicable to thesauri, ontologies and intelligent systems for natural language processing.

Keywords: word embedding, word2vec, semantic relations, thesaurus, hyponymy, hypernymy, synonymy

On the authors:
Maria Karyaeva, orcid.org/0000-0003-4466-1735, graduate student,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150003, Russia, e-mail: mari.karyaeva@gmail.com

Pavel Braslavski, orcid.org/0000-0002-6964-458X, PhD, Docent,
Ural Federal University,
19 Mira str., Ekaterinburg 620002, Russia, e-mail: pbras@yandex.ru

Valery A. Sokolov, orcid.org/0000-0003-1427-4937, Doctor, Professor,
P.G. Demidov Yaroslavl State University,
14 Sovetskaya str., Yaroslavl 150003, Russia, e-mail:sokolov@uniyar.ac.ru

Acknowledgments:
[1] The reported study was funded by RFBR according to the research projects №16-07-01180 и №16-06-00497