Методы машинного обучения применительно к задаче выделения глагольных и атрибутивных коллокаций

М.В. Хохлова, Е.В. Еникеева

Санкт-Петербургский государственный университет

m.khokhlova@spbu.ru, protoev@gmail.com

Аннотация

В статье представлены результаты применения алгоритмов машинного обучения к задаче автоматического выявления глагольных и атрибутивных коллокаций. Изучение сочетаемости показало, что дистрибуционные модели могут быть успешно использованы для моделирования отношений внутри словосочетаний. Словосочетание признается значимым, если его векторное представление близко к векторному представлению заглавного слова. Нами были использованы следующие методы оценки коллокаций на основе машинного обучения и векторных представлений текстов: базовый метод, метод аналогии и линейного преобразования. Автоматически выделенные словосочетания сравнивались с данными, приведенными в лексикографических источниках (в толковых словарях и словарях сочетаемости, всего было рассмотрено пять источников), которые образовали так называемый золотой стандарт. Результаты показали, что рассматриваемые методы успешно используются для извлечения словосочетаний, в том числе находят те, которые не отражены в словарях. Данные примеры могут претендовать на лексикографическое описание, хотя и не приведены в источниках и нуждаются в дополнительной экспертной проверке. Поэтому необходимо дополнительно провести сравнение использованных алгоритмов с другими статистическими метриками и увеличить количество словосочетаний, которые привлечены в качестве золотого стандарта.

Ключевые слова: глагольные коллокации, атрибутивные коллокации, машинное обучение, алгоритмы, корпусы текстов, русский язык

Библиографическая ссылка: Хохлова М.В., Еникеева Е.В. Методы машинного обучения применительно к задаче выделения глагольных и атрибутивных коллокаций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 4 (Труды XXIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2020, Санкт-Петербург, 17 – 20 июня 2020 г. Сборник научных статей). — СПб.: Университет ИТМО, 2020. С. 54-60. DOI: 10.17586/0000-0000-2020-4-54-60

Введение

Исследование сочетаемости не теряет своей актуальности на протяжении последних десятилетий. Определение лексических конструкций и их последующий анализ весьма важны для современных прикладных задач: создание словарей тональности, расширение поисковых запросов, автоматический перевод и др. Описание развития статистических методов, применяемых к нахождению в корпусах текстов словосочетаний разных типов, приведено в работе [1]. Для автоматического извлечения сочетаний используются разные статистические методы, в том числе ставшие широко распространенными в последние годы методы машинного обучения, которые нашли применение при решении

лингвистических задач в связи с появлением как больших текстовых данных, так и технических возможностей.

В статье рассматриваются результаты экспериментов по автоматическому выделению и последующей оценке атрибутивных и глагольных словосочетаний при помощи моделей машинного обучения на русскоязычных коллекциях текстов большого объема.

1. Обзор методов

Традиционные методы извлечения лексических конструкций используют лексико-грамматические шаблоны для поиска возможных примеров, которые затем ранжируются в соответствии с одной из статистических мер, отражающих степень связи между двумя признаками. Для вычисления совместной встречаемости необходимо учитывать совместную частоту встречаемости признаков (фиксированного и переменного компонента конструкции) в корпусе, что не дает возможности извлекать реализации конструкций, не встречающиеся в корпусе.

Современные методы, основанные на машинном обучении, решают, в частности, эту проблему, используя методики сглаживания данных. Популярные в последнее время модели языка, основанные на нейронных сетях (англ. a neural probabilistic language model), позволяют оценить вероятность не встреченной в корпусе последовательности: этот подход основан на получении информации о более широком контексте слова, чем в классических статистических моделях, для предсказания следующего слова в цепочке.

Векторное представление слов (англ. word embeddings) стало еще одним методом машинного обучения, который нашел применение при автоматическом извлечении словосочетаний. Векторные представления слов используются для представления связей между словами: синонимии, гипонимии и семантических аналогий [2]. Группа моделей, которые представляют слова, носит название word2vec и описана в работе [там же]. В отличие от традиционной дистрибутивной семантики word2vec основан на алгоритме машинного обучения, согласно которому слово может быть предсказано на основе его контекста (архитектура непрерывного мешка слов, англ. continuous bag-of-words) или наоборот, контекст в зависимости от слова (архитектура skip gram).

Исследование лингвистической композициональности [3; 4] показало, что дистрибутивные представления могут быть использованы для моделирования отношений в рамках словосочетаний довольно простым образом. Например, атрибутивное словосочетание признается значимым, если его векторное представление (вычисленное как сумма составляющих векторов) близка к вектору опорного слова. Таким образом, словосочетания могут быть отранжированы по их приемлемости согласно функции близости (например, косинусного сходства, англ. cosine similarity):

$$sim(noun, adj) = cos(noun + adj, noun)$$

Но для нормализованных векторов функции sim(noun, adj) и cos(noun, adj) будут одинаково монотонны, то есть для оценки степени приемлемости словосочетания достаточно оценить близость входящих в него слов. Такой подход мы будем использовать в качестве базовой оценки (англ. baseline).

Несмотря на то, что были достигнуты значительные результаты, методы редко используются для кодирования синтагматических отношений. Авторы [5] используют word2vec для предсказания коллокатов из словаря Macmillan Collocations Dictionary. Некоторые эксперименты по русскоязычным коллокациям были описаны в [6; 7]. Так, в работе [6] был протестирован данный метод на материале русского языка, показавший точность до 0,9.

2. Методология исследования

Как было отмечено выше, для работы с алгоритмами машинного обучения необходим довольно большой объем текстовых данных, а также нужны примеры эталонных данных, относительно которых будут оцениваться результаты.

В ходе работы над проектом коллокации нами извлекались в два этапа из Интернет-корпуса Aranea Russicum Maximum [8]. В качестве эталонных коллокаций, т.е. золотого стандарта, использовались словосочетания ряда словарей (толкового и специализированных): Толковый словарь русского языка в 4-х томах (МАС) [9], Словарь коллокаций [10], Словарь глагольной сочетаемости непредметных имен русского языка [11], Словарь русской идиоматики [12] и Словарь устойчивых глагольно-именных словосочетаний русского языка [13]. На их основе был создан список эталонных конструкций.

На первой стадии нами были разработаны правила, которые описывали русскоязычные словосочетания и были применены к извлечению данных. Были рассмотрены следующие синтаксические модели: 1) прилагательное + существительное; 2) глагол + существительное; 3) существительное + глагол. Второй этап включал использование модели word2vec для оценки извлеченных словосочетаний. Векторные представления были получены с помощью данного инструмента, обученного на текстах Национального корпуса русского языка (НКРЯ [14]). Список образцов лексических конструкций собран по данным синтаксически размеченного подкорпуса НКРЯ – СинТагРус.

Нами были использованы простые векторные операции к векторными представлениям слова, чтобы найти семантические аналогии, как это описано в работе [15]. Результирующий вектор «queen – woman + man» (слова представлены векторами) схож с векторным представлением слова «king».

Мы используем этот подход, чтобы смоделировать синтагматические отношения, а именно, найти наиболее близкий коллокат для данного слова и тип коллокации [6]. Набор коллокаций C, собранный на материале небольшого сбалансированного корпуса разных жанров, используется как пример аналогии (как queen – woman в примере выше).

Приемлемость коллокации w1+w2, где w1 является заглавным словом, вычисляется следующим образом:

$$S(w_1, w_2) = \underset{(c_1, c_2) \in C}{\operatorname{argmax}} ([c_1 - c_2 + w_2], w_1)$$

Всего были рассмотрены 3 метода оценки коллокаций на основе машинного обучения и векторных представлений текстов:

— базовый метод (англ. baseline), основанный на сходстве векторов слов, составляющих коллокацию;
— метод аналогии, основанный на предположении, что разность векторов слов, составляющих одну коллокацию, должна быть близка к разности векторов в эталонных коллокациях;
— метод линейного преобразования: из векторных представлений формируются пространства главных слов словосочетания и коллокатов. Затем на основании обучающей выборки строится матрица линейного преобразования между этими пространствами; для новых коллокаций оценивается близость произведения матрицы преобразования на вектор ключевого слова к вектору коллоката.

3. Результаты

В табл. 1 приведены данные по количеству биграмм после фильтрации, соответствующих использованным морфосинтаксическим-шаблонам, а также примеры наиболее устойчивых сочетаний в лемматизированном виде (то есть слова записаны в начальной форме).

Таблица 1. Количество извлеченных биграмм

Шаблон	Количество биграмм	Примеры
прилагательное + существительное	150642	докторский диссертация; населённый пункт; смертный казнь
существительное + глагол	103221	поезд курсировать; судно затонуть; войско двинуться
глагол + существительное	138281	гол забивать; победа одерживать; внимание обращать

Для оценки использовались стандартные меры – точность (precision), полнота (recall), F-мера (F-mean). Последняя является средним гармоническим двух предыдущих. Поскольку результатом автоматического выделения коллокаций является список словосочетаний, ранжированный по оценке одного из предложенных алгоритмов M, были проведены эксперименты по подбору порога значений M, который позволял бы отделить коллокации от свободных словосочетаний. Стоит подчеркнуть, что оценивались именно точность и полнота полученного списка коллокаций, а не их ранжирование (поскольку золотой стандарт не был отранжирован).

Оценка результатов была произведена относительно собранного золотого стандарта. Было сделано сравнение двух списков словосочетаний, соответствующих определенному шаблону (например, прилагательное + существительное), – золотой стандарт G и результат автоматического выделения A. В таблице 2 представлены результаты экспериментов.

Таблица 2. Оценка результатов экспериментов с использованием методов машинного обучения

Метод 1 (baseline)			
	сущ+прил	гл+сущ(косв)	сущ0+гл
Точность	0,1717	0,0613	0,0115
Полнота	0,6268	0,2898	0,4636
F-мера	0,2695	0,0952	0,0225
Метод 2 (метод аналогии)			
	сущ+прил	гл+сущ(косв)	сущ0+гл
Точность	0,1229	0,0531	0,0296
Полнота	0,4070	0,2220	0,2202
F-мера	0,1888	0,0858	0,0521
Метод 3 (метод линейного преобразования)			
	сущ+прил	гл+сущ(косв)	сущ0+гл
Точность	0,0480	0,0398	0,0339
Полнота	0,3459	0,1234	0,2033
F-мера	0,0843	0,0602	0,0580

Ниже приведены примеры словосочетаний, которые были извлечены автоматическими методами (изначально они являются лемматизированными), но при этом отсутствуют в рассмотренных словарях: «кандидатская диссертация», «фондовая биржа», «злокачественная опухоль», «смертная казнь»; «оканчивать школу», «надевать платье», «снимать кинофильм», «уничтожать противника»; «осадки выпадают», «войско двинулось», «избиратель голосует», «Бог смилостивился», «рассудок помутился».

4. Обсуждение

Следует отметить, что значения точности не показательны, поскольку, во-первых, используемые словари сочетаемости составлялись на основе других корпусов и ограничены по своему объему, во-вторых, среди данных золотого стандарта содержатся фразеологизмы и термины, которые имеют низкие частоты, а в-третьих, не учитываются правильно выделенные коллокации, отсутствующие в словарях. По критерию полноты, в отличие от экспериментов по предсказанию значений лексических функций [6], методы 2 и 3 показывают не такие высокие результаты по сравнению с baseline. Возможно, это связано с тем, что отношения внутри коллокаций (пусть и одного синтаксического типа) менее регулярны.

Заключение

Впоследствии будут рассмотрены иные источники, а также будут извлечены словосочетания в рамках других синтаксических моделей. Дополнительно планируется сопоставить полученные результаты с данными экспериментов по извлечению словосочетаний при помощи статистических мер ассоциации.

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-2513.2018.6 «Исследование методов автоматического извлечения лексических конструкций на основе машинного обучения».

Литература

[1] Хохлова М.В. Статистический подход применительно к исследованию сочетаемости: от мер ассоциации к машинному обучению // Структурная и прикладная лингвистика. 2019. Вып. 13. С. 106–122.

[2] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // Advances in neural information processing systems. 2013. P. 3111–3119.

[3] Vecchi E. M., Baroni M., Zamparelli R. (Linear) maps of the impossible: capturing semantic anomalies in distributional space // Proceedings of the Workshop on Distributional Semantics and Compositionality. Association for Computational Linguistics. 2011. P. 1–9.

[4] Kochmar E., Briscoe T. Capturing anomalies in the choice of content words incompositional distributional semantic space // RANLP. 2013. P. 365–372.

[5] Rodríguez-Fernández S., Anke L., Carlini R., Wanner L. Semantics-driven recognition of collocations using word embeddings // Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany. 2016. P. 499–505.

[6] Enikeeva E., Mitrofanova O. Russian Collocation Extraction based on Word Embeddings: Computational Linguistics and Intellectual Technologies // Papers from the Annual International Conference «Dialogue» (2017). Computational Linguistics: Practical Applications. Moscow: RSUH. 2017. Issue 16, Vol. 1. P. 52–64.

[7] Enikeeva E., Popov A. Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings // Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, 2018. P. 799–808.

[8] Benko V. Aranea Yet Another Family of (Comparable) Web Corpora // Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257-264.

[9] Толковый словарь русского языка: В 4-х т. /под ред. А.П. Евгеньевой. М., 1981–1984.
[10] Борисова Е. Г. Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. М.: 1995. 148 с.
[11] Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. Словарь глагольной сочетаемости непредметных имен русского языка. М., 2008. [Электронный ресурс] URL: http://dict.ruslang.ru/ abstr_noun.php (дата обращения: 01.05.2020).
[12] Кустова Г. И. Словарь русской идиоматики. Сочетания слов со значением высокой степени. М., 2008. [Электронный ресурс] URL: http://dict.ruslang.ru/magn.php (дата обращения: 01.05.2020)
[13] Дерибас В. М. Устойчивые глагольно-именные словосочетания русского языка. М.: Русский язык, 1979. 256 с.
[14] Национальный корпус русского языка. [Электронный ресурс] URL: http://ruscorpora.ru (дата обращения: 01.05.2020)
[15] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // ICLR: Proceeding of the International Conference on Learning Representations Workshop Track, Arizona, USA, 2013. P. 1301–3781.

Applying Machine Learning Methods to Verbal and Noun Phrases Extraction

M. Khokhlova, E. Enikeeva

St. Petersburg State University

The paper presents the results of applying machine learning algorithms to the task of automatic evaluation of verbal and noun collocations. The study of collocability showed that distribution models can be successfully used to model relations within phrases. A phrase is considered to be significant if its vector representation is close to the vector representation of the headword. We used the following methods for evaluating collocations based on machine learning and word embeddings: baseline, the method of analogy and linear transformation. Automatically selected phrases were compared with the data provided in lexicographic sources (in explanatory dictionaries and collocation dictionaries, five resources were considered in total), which formed the so-called gold standard. The results showed that the methods under consideration are successfully used to extract phrases, including those that are not reflected in the dictionaries. These examples can gain lexicographers' attention, although they are not given in the resources and need additional expert evaluation. Therefore, it is necessary to further compare the algorithms with other statistical metrics and increase the number of phrases in the gold standard.

Keywords: verbal collocations; noun collocations; machine learning; algorithms; text corpora; Russian language

Reference for citation: Khokhlova M.V., Enikeeva E.V. Applying Machine Learning Methods to Verbal and Noun Phrases Extraction // Computer Linguistics and Computing Ontologies. Vol. 4 (Proceedings of the XXIII International Joint Scientific Conference «Internet and Modern Society», IMS-2020, St. Petersburg, June 17-20, 2020). - St. Petersburg: ITMO University, 2020. P. 54 – 60. DOI: 10.17586/0000-0000-2020-4-54-60

References

[1] Hohlova M.V. Statisticheskij podhod primenitel'no k issledovaniyu sochetaemosti: ot mer associacii k mashinnomu obucheniyu // Strukturnaya i prikladnaya lingvistika. 2019. Vyp. 13. S. 106–122. [In Russian].

[2] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality // Advances in neural information processing systems. 2013. P. 3111–3119.
[3] Vecchi E.M., Baroni M., Zamparelli R. (Linear) maps of the impossible: capturing semantic anomalies in distributional space // Proceedings of the Workshop on Distributional Semantics and Compositionality. Association for Computational Linguistics. 2011. P. 1–9.
[4] Kochmar E., Briscoe T. Capturing anomalies in the choice of content words incompositional distributional semantic space // RANLP. 2013. P. 365–372.
[5] Rodríguez-Fernández S., Anke L., Carlini R., Wanner L. Semantics-driven recognition of collocations using word embeddings // Proceedings of the 2016 Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany. 2016. P. 499–505.
[6] Enikeeva E., Mitrofanova O. Russian Collocation Extraction based on Word Embeddings: Computational Linguistics and Intellectual Technologies // Papers from the Annual International Conference «Dialogue» (2017). Computational Linguistics: Practical Applications. Moscow: RSUH. 2017. Issue 16, Vol. 1. P. 52–64.
[7] Enikeeva E., Popov A. Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings // Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, 2018. P. 799–808.
[8] Benko V. Aranea Yet Another Family of (Comparable) Web Corpora // Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257-264.
[9] Tolkovyj slovar' russkogo yazyka: V 4-h t. /pod red. A.P. Evgen'evoj. M., 1981–1984.
[10] Borisova E.G. Slovo v tekste. Slovar' kollokacij (ustojchivyh slovosochetanij) russkogo yazyka s anglo-russkim slovarem klyuchevyh slov. M.: 1995. 148 s. [In Russian].
[11] Biryuk O.L., Gusev V.YU., Kalinina E.YU. Slovar' glagol'noj sochetaemosti nepredmetnyh imen russkogo yazyka. M., 2008. [Elektronnyj resurs] URL: http://dict.ruslang.ru/abstr_noun.php (data obrashcheniya: 01.05.2020). [In Russian].
[12] Kustova G.I. Slovar' russkoj idiomatiki. Sochetaniya slov so znacheniem vysokoj stepeni. M., 2008. [Elektronnyj resurs] URL: http://dict.ruslang.ru/magn.php (data obrashcheniya: 01.05.2020) [In Russian].
[13] Deribas V.M. Ustojchivye glagol'no-imennye slovosochetaniya russkogo yazyka. M.: Russkij yazyk, 1979. 256 s. [In Russian].
[14] Nacional'nyj korpus russkogo yazyka. [Elektronnyj resurs] URL: http://ruscorpora.ru (data obrashcheniya: 01.05.2020)
[15] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // ICLR: Proceeding of the International Conference on Learning Representations Workshop Track, Arizona, USA, 2013. P. 1301–3781.