К вопросу о сходстве мер ассоциации применительно к задаче автоматического извлечения глагольных коллокаций

М.В. Хохлова

Санкт-Петербургский государственный университет

m.khokhlova@spbu.ru

Аннотация

Извлечение коллокаций представляет собой одну из актуальных задач в обработке естественного языка, результаты которой важны и востребованы в разных областях прикладной лингвистики.
Наше исследование посвящено сравнению пятнадцати статистических мер, проведенному на подмножестве русскоязычного корпуса «Тайга». Были рассмотрены автоматически извлеченные глагольные коллокации. В ходе экспериментов изучались отличия между статистическими метриками, а также была предпринята попытка найти наиболее эффективную из них для русского языка. Вычислялся коэффициент ранговой корреляции Спирмена между мерами, а также извлеченные словосочетания оценивались относительно данных словаря, то есть проводилось сравнение между полученными автоматически и «вручную» словосочетаниями. Полученные результаты показывают, что некоторые меры показывают сходство и относительную взаимозаменяемость.

Ключевые слова: глагольные коллокации, корпус текстов, статистика, словари, меры ассоциации, оценка

Библиографическая ссылка: Хохлова М.В. К вопросу о сходстве мер ассоциации применительно к задаче автоматического извлечения глагольных коллокаций // Компьютерная лингвистика и вычислительные онтологии. Выпуск 3 (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. Сборник научных трудов). — СПб: Университет ИТМО, 2019. С. 9 –18. DOI: 10.17586/2541-9781-2019-3-9-18

1. Введение

Изучению лексической сочетаемости при помощи статистических методов посвящено большое количество работ. Первые попытки оценить силу связи между элементами словосочетания были высказаны в сборнике статей [1], в то время как в работе [2] статистический аппарат описывался применительно к задаче извлечения коллокаций. Наиболее полно большое число так называемых мер ассоциации (статистических метрик, которые чаще всего используются при автоматическом извлечении словосочетаний) представлено в монографии [3]. На русскоязычном материале подобные исследования проводились разными авторами (см., например, некоторые работы [4-6]). По мнению ряда исследователей (см., например, [7]), мы можем говорить о лексических и эмпирических коллокациях. Первый тип словосочетаний носит лингвистический характер, в то время как сочетания второго типа являются результатом работы некоторой программы или алгоритма. Цель нашего исследования заключается в том, чтобы показать, существует ли корреляция между этими двумя типами языковых явлений на примере глагольных словосочетаний.

2. Методология

В статье нами представлены результаты автоматического извлечения глагольных коллокаций при помощи статистического аппарата. В ходе исследования решались две задачи: во-первых, насколько статистические меры могут быть взаимозаменяемыми, во-вторых, насколько меры эффективны при извлечении глагольных словосочетаний типа «глагол +объект» на русском языке. В отличие от предыдущих работ, которые также проводились на русскоязычном материале, было исследовано 15 мер ассоциации.

В качестве материала было использовано подмножество корпуса «Тайга», которое содержит около 2,1 млн токенов (1,8 млн слов) [8]. Небольшой объем данных был выбран в связи с тем, что пока нет достаточного количества примеров (золотого стандарта) для оценки. Всего было извлечено 43808 сочетаний глаголов с существительными в форме объекта, например, *задействовать нейросеть*, *начинать жизнь*, *увеличивать шанс* и др. На следующем этапе данные были обработаны при помощи инструмента UCS [7]. Нами были выбраны следующие метрики для автоматического извлечения словосочетаний: коэффициенты логарифмического правдоподобия (LL), взаимной информации (MI), MI2, MI3, t-score, Дайса (Dice), Фишера (Fisher), Жаккарда (Jaccard), минимальной чувствительности (MS), Пуассона (Poisson), хи-квадрат, gmean, odds ratio и z-score. Также была использована абсолютная частота встречаемости слов. В основе исследуемых коэффициентов лежат разные принципы, например, точные и асимптотические статистические тесты или эвристические комбинации наблюдаемых и ожидаемых частот. Полное описание может быть найдено в работе [9]. Эксперименты предполагали сравнение между парами мер для определения того, в какой степени они выдают одинаковые результаты.

Остается открытым вопрос, каким образом интерпретировать количественные показатели этих мер, так как в их основе лежат не только разные гипотезы, но и сами значения метрик существенно отличаются друг от друга. Нами были проанализированы ранги найденных коллокаций, которые были упорядочены по значениям мер.

Для того чтобы оценить результаты, также необходимо ответить на следующий вопрос: какое словосочетание следует рассматривать как лексическое (условно говоря, «правильное»). Здесь следует обратиться или к носителям языка, или к верифицированным источникам, что представляет собой определенную проблему. Существует большое количество словарей, однако далеко не все подходят для решения этой задачи. Насколько нам известно, не существует словарей коллокаций большого объема (так, словарь [10] небольшой). Что касается толковых словарей, то в них представлена неполная информация. Словарь, в котором были бы представлены данные о сочетаемости, должен быть довольно большим, чтобы обеспечить «золотой стандарт» (в особенности этот вопрос стоит остро, когда речь идет об обработке корпусов текстов большого объема)[1]. Словари, в которых представлены устойчивые словосочетания и фразеологические единицы, не могут считаться надежным источником материала из-за присутствия низкочастотных слов и редких языковых явлений. Для нашего исследования мы использовали Словарь устойчивых глагольно-именных словосочетаний в русском языка [12], предназначенный для пользователей, изучающих русский язык. Он содержит более 5000 словосочетаний. После обработки словарных данных, исключения предложных конструкций и представления словосочетаний с видовыми парами как отдельных единиц был сформирован список из 7790 единиц (например, *оправдывать (оправдать) ожидания*, *питать уважение*, *читать лекцию* и т.д.). Словосочетания были лемматизированы при помощи программы MyStem [13].

[1] Словарь [11] отвечает данному требованию, однако работа над ним еще не завершена.

Нами были проведены два эксперимента. Целью первого было попарное сравнение мер между собой при помощи коэффициента ранговой корреляции Спирмена (например, MI и t-score). Результаты важны для оценки эффективности мер и определения их общих черт. В ходе второго эксперимента производились оценка относительно словаря. Мы сравнили списки полученные списки биграмм со словарем, чтобы найти коллокации, которые были описаны лексикографами и, следовательно, могут быть использованы для оценки статистических мер. Была протестирована следующая гипотеза: коллокации, зафиксированные в словарях, должны присутствовать в списках и обладать более высоким рангом.

3. Результаты

3.1. Попарное сравнение мер ассоциации

Нами были проанализированы списки выданных разными мерами ассоциации биграмм и вычислен коэффициент ранговой корреляции Спирмена (*rs*) для того, чтобы оценить сходство между мерами. Значения коэффициента варьируются в диапазоне от -1 до +1, указывая на отрицательную или положительную взаимосвязь соответственно. Также мы использовали частоту совместной встречаемости (абсолютная частота, *freq*) для того, чтобы сравнить ранжирование согласно сложным и простой (условного *baseline*) мерам. Результаты первого эксперимента представлены в табл. 1. Рис. 1 дает графическое представление о матрице корреляции (интенсивность цвета возрастает по мере увеличения значения коэффициента *rs* по модулю).

Некоторые меры обнаруживают высокую корреляцию между собой (коэффициент *rs* равен 1 и, следовательно, имеет место одинаковое ранжирование): 1) коэффициенты Фишера и Пуассона; 2) коэффициенты Жаккарда и Дайса; 3) хи-квадрат, MI2, gmean, z-score. Таким образом, можно выдвинуть предположение о том, что меры внутри одной группы обладают общими чертами и, таким образом, имеют схожее поведение.

Меры Жаккарда и Дайса — два коэффициента, которые часто рассматриваются как схожие ввиду самих формул. Полученные результаты доказывают, что они являются полными эквивалентами, когда речь заходит о рангах. Интересно также, что значение *rs* также является высоким (1,0) для коэффициентов Пуассона и Фишера. В наших экспериментах также несколько других мер показали сильную корреляцию: хи-квадрат и z-score принадлежат к группе асимптотических гипотетических тестов, а MI2 является чистой эвристической статистикой. Квадраты значений z-score равны значениям хи-квадрата, и, следовательно, ранжируют коллокации похожим образом. Как было отмечено в [7], коэффициент gmean использует среднее геометрическое и похож по своему поведению на метрику MI. Это утверждение подтверждается, мы обнаруживаем, что коэффициент даже показывает более сильную корреляцию с мерой MI2.

Как и ожидалось, частота совместной встречаемости показала самую низкую корреляцию с другими мерами за исключением t-score. Статистика t-score близко связана с наблюдаемыми частотами и, следовательно, значение *rs* (0,38) подтверждает утверждение, что данная пара имеет корреляцию. Отрицательные значения (и, следовательно, отрицательная корреляция) были обнаружены между частотой совместной встречаемости и MI и odds ratio.

В работах других исследователей MI часто описывается как мера, которая чувствительна к низким частотам, имеющая тенденцию приписывать им высокие значения данным, выдавая среди наиболее значимых низкочастотные словосочетания. Также нами обнаружены другие значения *rs*. Значения около 0 говорят об отсутствии корреляции.

Пуассон и Фишер показывают также очень низкие значения коэффициента корреляции с freq. MI демонстрирует очень сильную корреляцию с MI2 и MI3 (что может быть объяснено эмпирическим характером данных мер), с хи-квадратом, gmean, odds ratio и z-score.

Таблица 1. Корреляция между мерами ассоциации для глагольных словосочетаний

	freq.	**MI**	**MI²**	**MI³**	**LL**	**t-score**	**Dice**	**Fisher**	**Jaccard**	**Poisson**	**chi-squared**	**MS**	**gmean**	**odds.ratio**	**z.score**
freq.	1.00	-0.17	-0.03	0.09	0.15	0.38	0.03	0.18	0.03	0.18	-0.05	0.03	-0.03	-0.13	-0.05
MI	-0.17	1.00	0.97	0.91	0.85	0.61	0.80	0.83	0.80	0.84	0.98	0.74	0.97	0.94	0.98
MI²	-0.03	0.97	1.00	0.98	0.94	0.75	0.84	0.93	0.84	0.93	**1.00**	0.78	**1.00**	0.91	**1.00**
MI³	0.09	0.91	0.98	1.00	0.98	0.84	0.84	0.97	0.84	0.97	0.97	0.79	0.98	0.84	0.97
LL	0.15	0.85	0.94	0.98	1.00	0.87	0.78	0.99	0.78	0.99	0.94	0.72	0.94	0.82	0.94
t-score	0.38	0.61	0.75	0.84	0.87	1.00	0.67	0.90	0.67	0.90	0.73	0.63	0.75	0.56	0.73
Dice	0.03	0.80	0.84	0.84	0.78	0.67	1.00	0.80	**1.00**	0.81	0.83	0.99	0.84	0.62	0.83
Fisher	0.18	0.83	0.93	0.97	0.99	0.90	0.80	1.00	0.80	**1.00**	0.92	0.75	0.93	0.77	0.92
Jaccard	0.03	0.80	0.84	0.84	0.78	0.67	**1.00**	0.80	1.00	0.81	0.83	0.99	0.84	0.62	0.83
Poisson	0.18	0.84	0.93	0.97	0.99	0.90	0.81	**1.00**	0.81	1.00	0.92	0.76	0.93	0.78	0.92
chi-squared	-0.05	0.98	**1.00**	0.97	0.94	0.73	0.83	0.92	0.83	0.92	1.00	0.78	**1.00**	0.92	**1.00**
MS	0.03	0.74	0.78	0.79	0.72	0.63	0.99	0.75	0.99	0.76	0.78	1.00	0.78	0.55	0.78
gmean	-0.03	0.97	1.00	0.98	0.94	0.75	0.84	0.93	0.84	0.93	**1.00**	0.78	1.00	0.91	**1.00**
odds.ratio	-0.13	0.94	0.91	0.84	0.82	0.56	0.62	0.77	0.62	0.78	0.92	0.55	0.91	1.00	0.91
z.score	-0.05	0.98	**1.00**	0.97	0.94	0.73	0.83	0.92	0.83	0.92	**1.00**	0.78	**1.00**	0.91	1.00

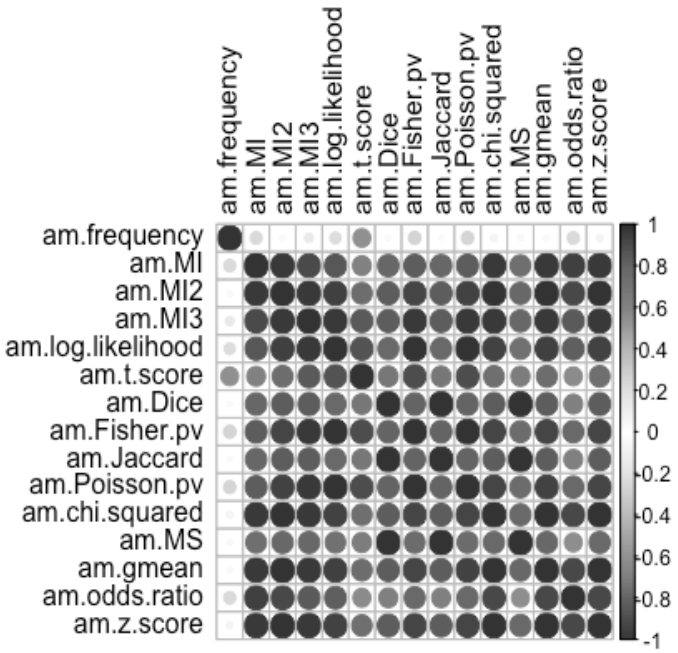

Рис. 1. Матрица корреляции между мерами ассоциации для глагольных словосочетаний

Однако, чем больше степень (в формулах для MI2 и MI3), тем менее выражена корреляция с метрикой MI. Значения rs для t-score с другими мерами варьируются в диапазоне от 0,38 до 0,90. Это может свидетельствовать о том, что коэффициент показывает «усредненное» поведение, частично ранжируя результаты так, как и иные меры.

3.2. Анализ глагольных коллокаций

В рамках следующего эксперимента мы выделили коллокации разными мерами и оценили результаты относительно словаря [12]. Мы проанализировали также вручную коллокации из первых 100 выданных словосочетаний. Результаты показывают, что списки содержат не только «правильные» коллокации, но и также бессмысленные комбинации и те словосочетания, которые могут быть рассмотрены как коллокации, но не приводятся в словаре. По нашему мнению, словосочетания из данной группы могут быть полезны для лексикографических целей. В табл. 2 приводятся результаты для точности при вычислении мер относительно словарной и экспертной оценок.

Таблица 2. Точность статистических мер

	Топ 100 (экспертная оценка)	**Топ 100**	**Топ 500**	**Топ 1000**	**Топ 2000**	**Все данные**
freq.	40,0	36,0	23,4	17,2	13,4	11,2
MI	0,00	0,0	0,0	0,2	0,3	1,5
MI2	0,00	0,0	0,0	0,4	0,9	1,9
MI3	41,0	37,0	8,0	5,1	4,0	4,9
LL	**48,0**	41,0	25,0	18,4	11,9	10,1
t-score	43,0	39,0	24,8	18,4	**13,9**	**11,3**
Dice	0,0	0,0	0,0	0,6	1,1	2,0
Fisher	**48,0**	41,0	**25,2**	**18,6**	12,4	10,5
Jaccard	0,0	0,0	0,0	0,6	1,1	2,0
Poisson	**48,0**	**42,0**	**25,2**	18,4	12,4	10,4

	Топ 100 (экспертная оценка)	Топ 100	Топ 500	Топ 1000	Топ 2000	Все данные
chi-squared	3,0	0,0	0,0	0,4	0,8	1,8
MS	3,0	0,0	0,0	0,6	1,2	2,0
gmean	3,0	0,0	0,0	0,4	0,9	1,9
odds.ratio	3,0	1,0	0,2	0,2	0,6	1,5
z.score	3,0	0,0	0,0	0,3	0,9	1,8

Частота совместной встречаемости

Частота совместной встречаемости показала достаточно высокие результаты и в качестве наиболее частотных указала коллокации, которые присутствуют в словаре: проводить испытание, играть роль, обращать внимание, делать вывод, получать информацию. Среди не зафиксированных в словаре словосочетаний, которые при этом могли бы представлять определенный лексикографический интерес (точность, вычисленная на 100 словосочетаний, возрастает до 40%): выполнять полет, совершать полет, определять местоположение.

Меры MI, MI2, MI3

Коэффициентом взаимной информации показаны худшие результаты при выделении словосочетаний для всех наборов данных. Несмотря на то, что в качестве материала были использованы тексты довольно хорошего качества, результаты снова подтвердили, что мера MI имеет тенденцию завышать значения для редких словосочетаний: опечаток, ошибок и др. Среди первых 500 словосочетаний ни одно не было зафиксировано в словаре, из 100 биграмм также ни одна не была отмечена как «правильная». Данный результат также интересен тем, что мера MI, во-первых, может использоваться как «детектор» ошибок (для поиска ошибок в текстах) или, наоборот, редких терминологических словосочетаний, а во-вторых, может демонстрировать хотя и низкую точность, но обнаруживать редкие осмысленные словосочетания на большом объеме данных. Результаты для меры MI2 почти полностью совпадают с теми, которые были продемонстрированы мерой MI. Результаты для меры MI3 резко отличаются от тех данных, что были получены для предыдущих двух мер. Резко выросла точность (для словарной и экспертной оценок до 37% и 41% соответственно), хотя также были найдены опечатки (например, *голосовать зазапрет, публикевать журнал* и др.).

Мера LL

Коэффициент логарифмического правдоподобия продемонстрировал один из лучших результатов для разных объемов данных. Следующие коллокации были найдены среди результатов: развивать скорость, проводить эксперимент, оказывать влияние, поражать цель, вносить вклад и др. Среди не зафиксированных в словаре коллокаций были отмечены следующие: *отслеживать перемещение, издавать звук, развивать тягу* и др.

Мера T-score

Мера t-score также демонстрирует высокие результаты, причем самые большие показатели точности были получены для списка из 2000 словосочетаний и всего объема данных. Среди наиболее частотных словосочетаний были выделены следующие: проводить испытание, развивать скорость, получать название, вести разработку, делать вывод и др. Были найдены словосочетания с высокочастотными существительными, что позволяет сделать вывод о том, что t-score выделяет чаще всего наиболее употребительные словосочетания.

Коэффициенты Дайса и Жаккарда

Данные меры показали абсолютно одинаковые результаты, точность варьировалась от 0% до 2% (т.е. практически не были найдены словарные коллокации). Результаты оказались схожими с мерой MI, также были выявлены словосочетания с ошибками и опечатками.

Коэффициенты Фишера и Пуассона

Мера Фишера показывает один из самых высоких результатов для точности на материале 500 и 1000 словосочетаний. Мера Пуассона, хотя и показала полную корреляцию с мерой Фишера, тем не менее продемонстрировала более успешные результаты для точности при оценке 100 первых словосочетаний относительно словаря.

Хи-квадрат, MS, gmean, odds.ratio

Хи-квадрат обнаружила низкие результаты при сравнении со словарем, однако при экспертном анализе были найдены словосочетания, которые могли бы претендовать на описание в словаре (например, *вспенивать шампунь, переждать непогоду, приумножать благосостояние, отозвать иск*). Меры MS, gmean, z-score показали в целом схожие результаты. Odds.ratio демонстрирует низкие результаты, выявляя только одно словосочетание, зафиксированное в словаре: приковать внимание. В остальном также большая часть результатов содержит ошибки и опечатки, которым приписываются большие значения меры, хотя результаты отличаются от других мер, перечисленных в данной группе.

4. Дискуссия

Анализ показывает, что, несмотря на разные принципы, лежащие в основе мер ассоциации, они демонстрируют в определенной степени одинаковые результаты (ранжирование может быть одинаковым или почти одинаковым). Можно заметить, что точность относительно словарных данных довольно низкая. Это связано с небольшим объемом самих данных, которые использовались в качестве эталона. Согласно закону Ципфа, большое количество лексических единиц имеет низкие частоты и, следовательно, в большом корпусе присутствует значительное число слов, встречающихся только один раз. В случае статистических метрик при ранжировании им присваивается одинаковый ранг (если меры тяготеют к выделению низкочастотных словосочетаний). Результаты показывают, что целый ряд мер (MI2, MS, хи-квадрат, gmean, Дайса и Жаккарда) ставят наверх среди первых 100 словосочетаний hapax legomena, и поэтому выделяемые биграммы совпадают.

Результаты также показывают, что словарные и экспертные данные коррелируют — чем выше результат, зафиксированный относительно лексикографического источника, тем также больше точность, оцениваемая экспертом. Топ-100 словосочетаний, выданных при помощи частоты совместной встречаемости, коэффициента логарифмического правдоподобия, мер MI3, t-score, Фишера и Пуассона включает большее число «правильных» коллокаций, чем списки, выданные иными мерами. Это может свидетельствовать о том, что словарные коллокации является по своей природе частотными и могут быть извлечены только при помощи мер, обнаруживающих умеренную или сильную корреляцию с частотой совместной встречаемости. Результаты также подтвердили первоначальную гипотезу: словарные коллокации обладают большими значениями статистических мер и, соответственно, низкими рангами (т.е. сконцентрированы в верхних частях списков).

5. Заключение

Результаты демонстрируют, что частота совместной встречаемости, MI3, t-score, меры Фишера и Пуассона выявляют значимые коллокации, которые встречаются относительно часто. В большинстве случаев это наиболее надежные меры. Наш подход имеет ограничение, так как каждый словарь обладает ограниченным объемом и не предоставляет полноценного описания сочетаемости. Общее количество коллокаций, используемое для оценки, оказывается недостаточным. Таким образом, необходимо и дальше увеличивать золотой стандарт, чтобы иметь возможность провести оценку на больших данных. Также

важно дать оценку мерам на материале больших корпусов текстов и сравнить их между собой. Предварительные результаты проведенных нами экспериментов показывают, что меры на большом объеме данных ведут себя по-другому. Также корпусные данные должны быть по возможности избавлены от ошибок, так как большинство мер обнаруживают чувствительность к разного рода выбросам.

Результаты показывают относительную взаимозаменяемость мер и могут быть использованы в дальнейшей работе по квантитативным методам и для их последующей оценки. Возможное решение для дальнейшего улучшения результатов — использование техник для комбинирования рангов, например, более сложных рангов, включающих разные меры.

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] Stevens M.E., Giuliano V.E., Heilprin, L.B. (eds.) Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation. Washington 1964. Vol. 269. National Bureau of Standards Miscellaneous Publication, 1965.

[2] Berry-Rogghe G. The Computation of Collocations and their Relevance in Lexical Studies // The Computer and Literary Studies. Edinburgh: Edinburgh University Press, 1973. P. 103-112.

[3] Pecina P. Lexical Association Measures. Collocation Extraction. Prague: Institute of Formal and Applied Linguistics, 2009.

[4] Хохлова М.В. Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Под ред. А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой. Хельсинки, 2008. С. 343–357.

[5] Zakharov V. Automatic Collocation Extraction: Association Measures Evaluation and Integration // Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”. V. 1. Computational Linguistics: Practical Applications. Moscow: RSUH, 2007. P. 396–407.

[6] Pivovarova L., Kormacheva D., Kopotev M. Evaluation of collocation extraction methods for the Russian language // Quantitative Approaches to the Russian Language (ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki). London, New York: Routledge, 2018. P. 137–157.

[7] Evert S. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004. URL: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (дата обращения: 27.05.2019).

[8] Taiga Corpus. URL: https://tatianashavrina.github.io/taiga_site/ (дата обращения: 27.05.2019).

[9] Computational Approaches to Collocations. URL: http://collocations.de (дата обращения: 27.05.2019).

[10] Борисова Е.Г. Слово в тексте. Словарь коллокаций (устойчивых сочетаний) русского языка с англо-русским словарем ключевых слов. М., 1995.

[11] Апресян Ю.Д. (ред.) Активный словарь русского языка. Т. 1-3. М.: Языки славянской культуры, 2014–2017.

[12] Дерибас В.М. Устойчивые глагольно-именные сочетания русского языка. М.: Русский язык, 1983.

[13] MyStem. URL: https://tech.yandex.ru/mystem/ (дата обращения: 27.05.2019).

On Similarity between Association Measures for Automatic Extraction: a Case Study of Verbal Collocations

M. Khokhlova

St. Petersburg State University

Collocation extraction has gained much attention in natural language processing, its results are important in various areas of applied linguistics. The research is a comparison of fifteen association measures based on a subset of Russian “Taiga” corpus. The paper studies automatically extracted Verb-Noun collocations. The aim of experiments is two-fold. First, to examine the difference between statistical measures and second to find out which measure is most efficient for Russian data. The former assumes calculation of Spearman’s rank correlation coefficient, whereas the latter implies evaluation of extracted lists against a Russian-language dictionary, i.e. to identify automatically extracted and manually collected collocations. The results are far not straightforward; we can distinguish groups of measures that demonstrate relative interchangeability. More so, produced bigrams can be of interest to lexicographers and may therefore enrich dictionaries.

Keywords: verbal collocations, text corpus, statistics, dictionaries, association measures, evaluation

Reference for citation: Khokhlova M. On Similarity between Association Measures for Automatic Extraction: a Case Study of Verbal Collocations // Computer Linguistics and Computing Ontologies. Vol. 3 (Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019). - St. Petersburg: ITMO University, 2019. P. 9 – 18. DOI: 10.17586/2541-9781-2019-3-9-18

Reference

[1] Stevens M.E., Giuliano V.E., Heilprin, L.B. (eds.) Proceedings of the Symposium on Statistical Association Methods for Mechanized Documentation, Washington 1964. V.269 of National Bureau of Standards Miscellaneous Publication, 1965.

[2] Berry-Rogghe G. The Computation of Collocations and their Relevance in Lexical Studies // The Computer and Literary Studies. Edinburgh: Edinburgh University Press, 1973. P. 103-112.

[3] Pecina P. Lexical Association Measures. Collocation Extraction. Prague: Institute of Formal and Applied Linguistics, 2009.

[4] Khokhlova M.V. Eksperimental’naja proverka metodov vydelelnija kollokacij [Evaluation of Methods for Collocation Extraction]. In Slavica Helsingiensia 34. Instrumentarij rusistiki: Korpusnye podhody. Eds. A. Mustajoki, M.V. Kopotev, L.A.Birjulin, J.J. Protasova. Helsinki, 2008. P. 343–357.

[5] Zakharov V. Automatic Collocation Extraction: Association Measures Evaluation and Integration. In Computational Linguistics and Intellectual Technologies: Papers from the Annual conference “Dialogue”. V.1. Computational Linguistics: Practical Applications. Moscow: RSUH, 2017. P. 396–407.

[6] Pivovarova L., Kormacheva D., Kopotev M. Evaluation of collocation extraction methods for the Russian language. In Quantitative Approaches to the Russian Language (ed. by M. Kopotev, O. Lyashevskaya, A. Mustajoki). London, New York: Routledge, 2018. P. 137–157.

[7] Evert S. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2004. URL: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf (дата обращения: 27.05.2019).

[8] Taiga Corpus. URL: https://tatianashavrina.github.io/taiga_site/ (дата обращения: 27.05.2019).
[9] Computational Approaches to Collocations. URL: http://collocations.de (дата обращения: 27.05.2019).
[10] Borisova E.G. Slovo v tekste. Slovar' kollokatsiy (ustoychivykh cochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov [A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords]. Moscow, 1995.
[11] Apresyan Ju. D. (ed.) Aktivnyy slovar' russkogo yazyka [Active Dictionary of the Russian Language]. Vol. 1-3. M.: Yazyki slavyanskoy kul'tury, 2014–2017.
[12] Deribas V.M. Ustojchivyje glagol'no-imennyje slovosochetanija russkogo jazyka [Verb-Noun Collocations in Russian]. Moscow: Russian language, 1983.
[13] MyStem. URL: https://tech.yandex.ru/mystem/ (дата обращения: 27.05.2019).