Сопоставительный анализ статистических мер на примере частеречных преференций сочетаемости существительных

М. В. Хохлова

Санкт-Петербургский государственный университет

m.khokhlova@spbu.ru

Аннотация

Применение квантитативных методов к корпусному материалу позволяет исследователям количественно оценить получаемые данные. Наряду с наиболее часто используемыми мерами для вычисления силы связанности в рамках словосочетаний, такими как MI, t-score или log-likelihood, существуют и иные коэффициенты, которые реже упоминаются в исследованиях, посвященных автоматическому выделению коллокаций. В статье представлен обзор некоторых из данных метрик, в том числе обсуждаются их основные характеристики. Производится их сравнение на материале биграмм для некоторых высокочастотных русских существительных.

Ключевые слова: статистический аппарат, меры ассоциации, биграммы, MI, LL, MI3, MS, t-score

1. Введение

В последнее время в связи с возросшей потребностью в автоматизированных системах большое внимание уделяется вопросу, связанному с автоматическим выделением словосочетаний сочетаний в текстах. Существуют различные статистические метрики для оценки сочетаемости слов. Ряд мер получил название мер ассоциации, или ассоциативных мер. Они позволяют вычислять силу связи между элементами словосочетаний и основываются на частотах данных словосочетаний и входящих в них отдельных слов. Таким образом, может быть вычислена некоторая устойчивость, присущая лексическим единицам, позволяющая их расположить на условной шкале: от свободных сочетаний до фразеологизированных структур. Всего существует более 80 мер, позволяющих оценить силу связанности словосочетаний [1].

В статье мы рассмотрим как часто используемые меры, так и редко описываемые статистические метрики применительно к русскоязычному материалу и сравним результаты по автоматическому выделению биграмм. Изложение построено следующим образом. В разделе 2 дается обзор статистического аппарата, в разделе 3 приводится методика исследования и в последней части обсуждаются результаты.

2. Статистический аппарат

В рамках исследования были отобраны 7 мер: MI, log-likelihood (LL), t-score, MI^3, minimum sensitivity (MS), logDice и MI.log-f. Наиболее часто в литературе, посвященной вычислению силы связанности, упоминаются первые три меры. Их подробный обзор дан в ряде работ (см., например, [2]).

Особенностью меры MI (или коэффициента взаимной информации) является то, что она позволяет найти в корпусе редкие словосочетания. Таким образом, вес каждой отдельной коллокации тем больше, чем реже она встречается. Поэтому в случаях, когда частота сочетания мала, использование данной формулы может привести к неправильным результатам. Чтобы решить эту проблему, в ряде работ были предложены модификации данной меры (примерами являются следующие меры).

В работе [3, с. 171–172] эмпирически выводится формула MI^3, в которой величина f(n,c) возводится в куб (как оказалась, данная степень позволяет улучить результаты):

$$MI^3 = \log_2 \frac{f^3(n,c) \times N}{f(n) \times f(c)},$$

где:

n — ключевое слово; c — коллокат; f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе; N — общее число словоформ в корпусе.

Исследователями мера MI^3 оценивается неоднозначно. Так, в монографии [4, с. 83] указывается, что она завышает значения для высокочастотных слов наряду с LL и t-score. Также подчеркивается, что мера успешно справляется с проблемой редких словосочетаний, характерной для MI [5, с. 274], хотя по своему поведению присущая ей кривая значений напоминает график для меры Дайса [6] [7].

Следующая мера MI.log-f также является вариантом взаимной информации и была введена в работе [8]:

$$MI.\log\text{–}f = MI \times \ln(f(n,c)+1)$$

В работе [9] была предложена мера *logDice*. Она является модификацией меры Дайса [6] [7] и призвана устранить ее недостаток, связанный с выдачей повышенных значений для редко встречающихся биграмм:

$$\log Dice = 14 + \log_2 \frac{2f(n,c)}{f(n) + f(c)}$$

Мера MS (minimum sensitivity) была предложена в работе [10]. Ее значения колеблются в диапазоне от 0 до 1.

$$MS = \min\{\frac{f(n,c)}{f(n,c)+f(n,\bar{c})}, \frac{f(n,c)}{f(n,c)+f(\bar{n},c)}\}$$

Как отмечается в работе [2], данная мера показала наилучшие результаты по сравнению с другими, и была реализована в ряде систем.

3. Методика исследования

Наше исследование было проведено на материале корпуса новостных текстов Интернет-издания «Фонтанка», содержащего 1,2 млн. словоупотреблений.

Нами были рассмотрены биграммы для десяти высокочастотных существительных, отобранных по словарю [11]: *год, человек, время, дело, жизнь, день, рука, работа, слово* и *место*. Были проанализированы 100 левосторонних нелемматизированных биграмм для каждого из приведенных существительных, которые были выделены при помощи 7 вышеописанных статистических мер и отсортированы по убыванию значения каждой из мер. Таким образом, было получено 7000 биграмм, далее были рассмотрены части речи найденных словоформ. Предварительно не была проведена фильтрация списков, так как планировалось проверить, какие биграммы выделяются в «сыром» виде. В ходе эксперимента сравнивались списки биграмм, полученные при помощи разных мер, т.е. результативность статистических метрик.

4. Результаты

Ниже приведены результаты эксперимента. В таблице 1 указывается количество левосторонних соседей разных части речи, выделенных для каждого существительного описанными мерами.

Таблица 1. Распределение морфологических тегов (в процентах)

	Существи-тельное	**глагол**	**Числитель-ное**	**Местоиме-ние**	**предлог**	**Прилага-тельное**	**союз**	**частица**	**наречие**	**пунктуация**
MI	15,7	29,8	12,1	21,8	4,3	13,3	0,4	0,4	0,4	1,8
t-score	16,1	29,6	11,5	21,3	4,2	13,3	0,6	0,4	0,7	2,3
MI3	23,4	25,2	9,6	17,0	5,1	13,5	1,0	1,0	1,2	3,0
MI.log-f	15,8	30,1	11,5	21,4	4,1	13,3	0,6	0,4	0,6	2,2
LL	24,3	24,2	10,1	17,5	4,7	13,9	0,8	0,7	1,0	2,8
MS	20,5	26,0	9,9	19,5	4,9	14,1	0,9	0,6	0,7	2,9
logDice	20,8	25,4	10,3	19,0	4,9	13,9	0,9	0,8	1,0	3,0

4.1 Мера MI

Отличительной особенностью меры MI оказалось то, что по сравнению с другими мерами она выделила наибольшее количество сочетаний с числительными и местоимениями. Это можно объяснить наличием в корпусе биграмм с лексемами «год», «человек», «дело», «день», «слово» и «место», в которых числительные имеют небольшие частоты (что подтверждает утверждение о том, что MI завышает значения для низкочастотных единиц). Например, «первого места», «третьего дня» или «2017 году». Также мерой было выдано наименьшее число существительных по сравнению с другими статистическими метриками и сочетания с ними были зафиксированы в конце списка (т.е. значения MI были невысокими), например, «неимением места».

4.2 Мера t-score

Данная мера показывает высокие значения для сочетаний с глаголами, числительными и местоимениями. Наибольшее количество биграмм с глаголами (так были размечены причастия) было зафиксировано для лексемы «время» («потребуется время»). Наряду с мерой MI t-score демонстрирует небольшое количество сочетаний с существительными. Исключением является лексема «жизнь», для которой модель существительное + «жизнь» оказалась самой частотной («образа жизни»).

4.3 Мера MI3

По сравнению с другими мерами (за исключением меры logDice) было продемонстрировано наибольшее количество сочетаний со знаками препинания, союзами, частицами, наречиями и предлогами. Полученные результаты подтверждают наблюдения, описанные в [4]. Наименьшее число биграмм по сравнению с другими мерами характерно для числительных и местоимений, что демонстрирует противоположный результат по сравнению с мерой MI. Так же в отличие от взаимной информации зафиксировано большое число сочетаний с существительными (23,4% vs 15,7%).

4.4 Мера MI.log-f

Для всех 10 лексем данная мера выделила максимальное число биграмм с глаголами (30,1%), при этом был зафиксирован небольшой процент сочетаний со служебными словами. Частотное распределение частей речи, полученное для данной меры, схоже со значениями для метрики MI.

4.5 Мера LL

Мера LL выделила наибольшее количество существительных по сравнению с другими мерами («конец года») Почти четверть найденных согласно данной мере сочетаний включает эту часть речи. Зафиксировано такое же количество биграмм с глаголами, что отличает данную меру среди остальных. Так, для лексемы «время» 37% биграмм построены по модели глагол + «время».

4.6 Мера MS

Зафиксировано максимальное число биграмм с прилагательными («выборного года»), которые оказались в верхней части списков. Также было найдено большое количество биграмм со знаками пунктуации. Отметим, что в силу специфики формулы ряд биграмм получил одинаковые значения меры, что затрудняло ранжирование результатов и их последующую обработку.

4.7 Мера logDice

Наряду с мерой MI3 было выделено наибольшее количество сочетаний со знаками препинания. Также две меры ведут себя схожим образом: одинаково ранжируют по частоте найденные модели биграмм.

Заключение

Результаты показывают, что все меры выделяют глагол в качестве наиболее вероятного синтагматического партнера для 10 рассмотренных существительных. Меры t-score, MI и MI.log-f показывают одинаковое ранжирование частей речи по частоте встречаемости, то же относится к метрикам MI^3 и logDice. Однако внутри самих списков для каждого существительного распределение биграмм отличается.

Можно предположить, что выделяются группы статистических мер, которые имеют схожее поведение и, как следствие, выдают одинаковые результаты. С одной стороны, это может свидетельствовать об относительной взаимозаменяемости метрик внутри групп, а с другой стороны, позволяет нам выдвинуть гипотезу о том, что для улучшения результатов полезно комбинировать меры (их абсолютные значения или ранги) из разных групп, «охватывая» таким образом разные словосочетания.

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-5274.2016.6 «Исследование статистических закономерностей сочетаемости лексических единиц».

Литература

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

[2] 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 25.05.2017)

[3] Oakes M. Statistics for corpus linguistics. Edinburgh: Edinburgh University Press, 1998.

[4] Kosem I. Interrogating a Corpus. In The Oxford Handbook of Lexicography, edited by Ph. Durkin. Oxford University Press, 2015. P. 74–93.

[5] Pazienza M.T., Pennacchiotti M., Zanzotto F. M. Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In Knowledge Mining

Proceedings of the NEMIS 2004 Final Conference, edited by Spiros Sirmakessis. Springer, 2005. P. 255–280.
[6] Smadja F. Retrieving collocations from text: Xtract. Computational Linguistics, 19 (1), 1993. Pp. 143–177.
[7] Dias G., Guillore S., Lopes J. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. In Proceedings of of 6ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN), 12–17 July, Cargèse, France, 1999. Pp. 333–339.
[8] Kilgariff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. In Proceedings of Euralex Lorient, France, July, 2004. P. 105-116.
[9] Rychly, P. A lexicographer-friendly association score. In Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008, 2008. Pp. 6–9.
[10] Pedersen T., Bruce R. What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX, 1996.
[11] Ляшевская О.А., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

A Comparative Analysis of the Association Measures based on the Analysis of Part-of-Speech Collocational Preferences of the Russian Nouns

M. V. Khokhlova

St. Petersburg State University

Application of quantitative methods to corpora data allows researchers to evaluate the represented results. Along with the most frequently implemented association measures such as mutual information, t-score or log-likelihood there are a number of other coefficients that are rarely described in papers dealing with automatic collocation extraction. In our survey we dwell on several of the coefficients paying attention to their characteristics and compare extracted bigrams for high frequent Russian nouns.

Keywords: statistical tools, measures of association, bigrams, MI, LL, MI3, minimum sensitivity, t-score.