Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации

В.П. Захаров, М.В. Хохлова

Санкт-Петербургский государственный университет
vz1311@yandex.ru, khokhlova.marie@gmail.com

Аннотация

Статья представляет результаты исследования по выделению терминологических словосочетаний на основе различных статистических мер: t-score, MI, MI^3, min. sensitivity, log-likelihood, logDice, и MI.log_f. Приводятся результаты экспериментов с помощью системы Sketch Engine.

1. Введение

Огромное место в лингвистике, в терминоведении, занимают специализированные подъязыки. Их особенности по синтаксису, семантике, формальным характеристикам таковы, что требуют отдельных исследований. Однако тексты специализированных подъязыков недостаточно полно отражены в корпусах русского языка и требуют создания для них специальных корпусов как эмпирической базы для анализа. Специальные тексты во многом терминологичны, и поэтому представляет интерес разработка корпусно-ориентированных методов автоматизированного выделения терминов и терминологических сочетаний (коллокаций) на основе корпусов текстов [1].

На основе статей по компьютерной лингвистике (в первую очередь доклады на конференциях «Диалог» и «Корпусная лингвистика») был создан корпус по компьютерной лингвистике общим объемом 2,6 млн словоупотреблений.

Большая часть терминов — лексических единиц терминосистемы и тезауруса предметной области — представляет собой словосочетания. Поэтому встает задача выработки автоматических (полуавтоматических) методов выявления таких сочетаний по корпусам текстов. Многие из этих методов базируются на вероятностно-статистических основаниях [2]. Статистический аппарат, применяемый в системах работы с корпусами текстов, позволяет пользователям ранжировать выделенные термины и терминосочетания по количественным параметрам и задавать численные пороговые значения, что повышает достоверность получаемых данных и создает параметрически настраиваемую систему.

Проблема сочетаемости в лингвистике исследована давно и подробно. Однако внедрение в лингвистику компьютерных технологий принесло новые методы изучения и освоения проблемы сочетаемости. Количественные показатели силы синтагматической связи, вычисленные на основе мер ассоциации и статистических данных, получаемые на больших корпусах, позволяют достаточно эффективно выделять устойчивые терминологические словосочетания.

Существует ряд работ, посвященных автоматическому выявлению устойчивых словосочетаний на англоязычном материале (в английской терминологии, collocations или multiword expressions) [3], но для русского языка картина скорее обратная.

2. Меры ассоциации

Наше исследование выполнено с помощью специализированной комплексной корпусной системы Sketch Engine (http://www.sketchengine.co.uk) [4]. В данном исследовании мы используем механизм вычисления и выдачи коллокаций – сочетаний заданного слова с другими - с количественным указанием силы связи, которая рассчитывается на основе мер ассоциации.

В указанной системе доступны следующие меры: t-score (t-test), MI (mutual information), MI^3, MI-log-prod (MI.log_f), minimum sensitivity, log-likelihood ratio, Dice (logDice), MI.log_f.

Сегодня в корпусной лингвистике наиболее популярны меры t-score, MI и log-likelihood.

MI^3 и MI-log-prod являются модификациями популярной меры MI, придающие дополнительный вес частоте совместной встречаемости элементов коллокации.

Мера minimum sensitivity (минимальная чувствительность) представляет собой минимум из двух отношений (отношение совместной частоты к частотам ключевого слова и коллоката), и отдает предпочтение тем коллокациям, частота совместной частоты которых среди всех коллокаций в наибольшей степени приближается к частоте ключевого слова или коллоката [5].

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Все меры в своих формулах учитывают частоту составных элементов коллокации и частоту совместной встречаемости (по-разному). При этом меры minimum sensitivity, log-likelihood и Dice не учитывают объем корпуса.

3. Эксперименты

Для эксперимента были выбраны 10 частотных слов-терминов: *слово, язык, текст, значение, система, словарь, глагол, анализ, объект, отношение*.

Исследовалась сочетаемость указанных слов, а именно, для них выделялись коллокации в интервале [-2;0] и [0;+2] относительно заданного (ключевого) слова с использованием всех вышеуказанных мер ассоциации.

Sketch Engine может выдавать список коллокатов сразу по всем мерам, ранжированный по какой-либо одной заданной мере. Ниже приводится пример такого списка с указанием частоты каждого коллоката в корпусе и значения силы связи между ключевым словом и коллокатом по каждой мере ассоциации (рис. 1, список ранжирован по MI.log_f).

Collocation candidates

Page 1 Go Next >

	Freq	T-score	MI	MI3	log likelihood	min. sensitivity	logDice	MI.log_f
p/n ключевой	416	20.278	7.436	24.837	3862.485	0.040	10.290	44.864
p/n порядок	243	15.353	6.046	21.896	1631.382	0.024	9.472	33.236
p/n знаменательный	66	8.066	7.138	19.227	568.207	0.006	7.700	30.014
p/n служебный	91	9.443	6.626	19.642	696.959	0.009	8.148	29.961
p/n опорный	64	7.922	6.685	18.685	496.597	0.006	7.650	27.905
p/n многозначный	63	7.855	6.599	18.554	479.444	0.006	7.627	27.447
p/n иной	189	13.380	5.225	20.350	1032.402	0.018	9.065	27.418
p/n вводный	46	6.723	6.826	17.873	368.467	0.004	7.181	26.280
p/n фонетический	95	9.542	5.573	18.712	567.022	0.009	8.173	25.435
p/n со	147	11.565	4.439	18.839	637.919	0.014	8.649	22.184
p/n реестровый	14	3.728	8.095	15.710	162.259	0.001	5.478	21.923
p/n значение	376	17.892	3.694	20.803	1260.115	0.037	9.446	21.913
p/n употребление	136	11.088	4.346	18.520	572.452	0.013	8.539	21.380
p/n отдельный	101	9.613	4.523	17.839	449.592	0.010	8.182	20.917
p/n пара	78	8.480	4.652	17.223	361.154	0.008	7.851	20.326

Рис. 1. Коллокаты для слова *слово*

4. Оценка

Всего нами было проанализировано 220 биграмм (при учете контекста слева [-2;0] и справа [0;+2]).

Анализ правого контекста [0;+2] показал большое количество комбинаций со знаками пунктуации, специальными символами и незначимыми словами, поэтому в дальнейшем подробно анализировался контекст в интервале [-2;0] (первое и второе слово слева).

Были выделены типовые структурные схемы номинативных конструкций с указанными терминами. Результаты поиска биграмм для левого контекста были сгруппированы вручную по следующим грамматическим моделям: 1) Adj + N; 2) N + N; 3) V + N; 4) другие. Например, для слова *словарь*: 1) Adj + N — *толковый словарь*, *семантический словарь*, *электронный словарь*; 2) N + N2 — *составление словаря*, *единица словаря*, *статья словаря*; 3) V + N —*использовать словарь*, *создать словарь*, *пополнять словарь*.

Процент биграмм, сгруппированных по вышеозначенным моделям следующий: 1) Adj + N — 65%; 2) N + N2 — 20%; 3) V + N — 5%; 4) другие (комбинации со знаками пунктуации, специальными символами и др.) — 10%. Таким образом, еще раз подтверждается известное утверждение, что большая часть двучленных терминологических сочетаний – это прилагательное + существительное.

Среди найденных биграмм можно выделить две группы. Первая группа включает терминологические сочетания, которые должны быть включены в терминологический словарь, напр.: *поверхностный словарь*, *кластерный анализ*. Вторая группа представлена высокочастотными коллокациями, переходными между терминами и свободными сочетаниями: *терминологическая система*, *опорное слово, порядок слов* и др. Обе группы должны быть учтены при описании терминосистемы компьютерной лингвистики.

Далее были проанализированы «терминологические» коллокаты относительно разных мер. Нас интересовало, как часто тот или другой коллокат представлен в выдаче по разным мерам и какой он там имеет ранг. Фрагмент такого анализа представлен в Табл. 1.

Часто указывалось, что MI дает высокие показатели тем биграммам, в состав которых входят низкочастотные составляющие (редкие слова) или даже ошибки. Поэтому эта мера из данного анализа была исключена. Отсюда максимальное число мер, на основе которых было которые выделено одно и то же сочетание, шесть (вместо 7).

Биграммы, показанные в таблице, были найдены 5 или 6 мерами, их ранги по данным мерам приведены в третьем столбце, четвертый столбец показывает количество мер, пятый – частота коллоката в корпусе. Внутри каждой биграммы данные отсортированы по рангу.

Таблица 1. Левые коллокаты для слова *слово*, найденные 6 или 5 мерами

Коллокат	Статистическая мера	Ранг	Количество мер	Частота в корпусе
значение	**min.sensivity**	**2**	**6**	**376**
	t-score	2		376
	logDice	3		376
	log-likelihood	3		376
	MI3	3		376
	MI.log_f	12		376
ключевой	**logDice**	**1**	**6**	**416**
	log-likelihood	1		416
	MI.log_f	1		416
	MI3	1		416
	min.sensivity	1		416
	t-score	1		416
отдельный	**t-score**	**3**	**6**	**101**
	logDice	13		101
	log-likelihood	13		101
	MI.log_f	14		101
	MI3	17		101
	Min.sensivity	22		101
порядок	**logDice**	**2**	**6**	**243**
	log-likelihood	2		243
	MI.log_f	2		243
	MI3	2		243
	Min.sensivity	3		243
	t-score	7		82
служебный	**MI.log_f**	**4**	**6**	**91**
	log-likelihood	5		91
	MI3	5		91
	t-score	5		91
	logDice	16		91
	Min.sensivity	29		91
фонетический	**t-score**	**4**	**6**	**95**
	log-likelihood	9		95
	MI.log_f	9		95
	MI3	10		95
	logDice	14		95
	Min.sensivity	25		95
знаменательный	**MI.log_f**	**3**	**5**	**66**
	log-likelihood	8		66
	MI3	8		66
	t-score	8		66
	logDice	30		66
иной	**logDice**	**4**	**5**	**189**
	log-likelihood	4		189
	MI3	4		189
	Min.sensivity	5		189
	MI.log_f	7		189
каждый	**logDice**	**11**	**5**	**120**
	Min.sensivity	14		120
	log-likelihood	18		120
	MI3	21		120
	MI.log_f	30		120
количество	**logDice**	**20**	**5**	**84**
	log-likelihood	21		84
	t-score	21		84
	MI.log_f	24		84
	MI3	25		84

Из таблицы видно, что некоторые меры, не являясь математически эквивалентными, выдают «кандидатов» в терминологические сочетания с ключевой лексемой (в данном примере, *слово*) с одинаковым или близким рангом, например, *ключевой*, *значение*, *порядок*. Возможно, это как раз и есть значимые коллокации для данной терминосистемы.

Выявленные коллокации для других слов показывают ту же картину, а именно: наличие двух типов терминологических сочетаний и определенный «изоморфизм» в работе разных мер ассоциации.

5. Заключение

Вышеописанная методология с использованием инструментов статистических мер позволяет на материале научных текстов русского языка извлекать терминологические словосочетания и оценить количественно (до определенной степени) силу связи между элементами словосочетаний.

Одновременное использование нескольких мер ассоциации в системе Sketch Engine расширяет возможности статистического подхода, увеличивает достоверность того, что найденные коллокаты действительно тесно связаны с заданным ключевым словом, а в ряде случаев позволяет более гибко выбирать кандидатов в устойчивые терминологические сочетания.

Для более качественной работы системы требуется совершенствование средств морфологической разметки. Некоторые «шумовые» результаты в выдаче объясняются именно ошибками в морфологической разметке, в частности: 1) частям сложных существительных приписаны разные леммы; 2) отсутствуют или недостаточно эффективны процедуры снятия морфологической неоднозначности; 3) ошибки в приписывании правильных лемм из-за отсутствия в морфологическом словаре специальных терминов.

Представляется, что в процедуры выявления кандидатов в коллокаты следует ввести так называемые стоп-слова, в первую очередь, знаки препинания и предлоги.

Совместное использование нескольких мер позволяет ввести интегрированный показатель синтагматической связанности. Такой показатель может включать в себя суммарный (или средний) ранг по всем мерам и нормированную частоту совместной встречаемости.

Литература

[1] Захаров В.П. Тезаурус по корпусной лингвистике // Информационные технологии и письменное наследие. El'Manuscript-10. Материалы Международной научной конференции. — Уфа, 2010. — С. 95-98.

[2] Evert S. Computational Approaches to Collocations. URL: http://collocations.de

[3] Evert S. The statistics of word cooccurrences: Word pairs and collocations. Ph. D. thesis, University of Stuttgart, 20; Krenn B., Evert S. Can we do better than frequency? A case study on extracting PP-verb collocations. // Proceedings of the ACL Workshop on Collocations. 2001. P. 39-46; Křen M. Kolokační miry a čestina: srovnaní na datech ČNK. // Kolokace. Praha: Ústav Českého národního korpusu, 2006. P. 223–248; Pečina P. Lexical association measures: collocation extraction. Praha, 2009.

[4] Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. — Lorient: Universite de Bretagne-Sud, 2004. — P. 105–116.

[5] Pedersen T. Dependent Bigram Identication. //Proceedings Fifteenth National Conference on Artifical Intelligence. 1998. P. 1197.

[6] 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.

Extracting Terminological Phrases By Different Association Measures

V.P. Zakharov, M.V. Khokhlova

The paper presents the results of automatic term extraction from a special text corpus (a collection of papers on computational linguistics) by means of statistical measures (t-score, MI, MI^3, min. sensitivity, log-likelihood, logDice, and MI.log_f). The applied method includes statistical analysis that enables estimating the value of syntagmatic relations between lexemes. The results of experiments on the base of Sketch Engine system are shown.