Метод выделения коллокаций с использованием степенного показателя в распределении Ципфа

Клышинский Э.С., Кочеткова Н.А., Национальный исследовательский университет «Высшая школа экономики
eklyshinsky@hse.ru, natalia_k_11@mail.ru,
Карпик О.В. ИПМ им. М.В. Келдыша РАН.
parlak@mail.ru

Аннотация

Для выделения из коллокаций текста мы предлагаем использовать степенной показатель распределения Ципфа. Для этого предлагается рассчитывать распределение Ципфа для фиксированного слова и его соседей. В статье проводится исследование получаемых результатов для таких пар как прилагательное+существительное, существительное+глагол и др. Предложенный метод сравнивается с результатами расчета меры MI.

1 Введение

Распределение Ципфа является одним из фундаментальных законов в области компьютерной лингвистики [Маслов и др., 2006]. На данный момент существует несколько теорий относительно природы этого закона [i Cancho et al., 2005], однако исследователи не пришли к единому мнению на данный счет.

В исследованиях было показано, что средний показатель распределения Ципфа для разных языков колеблется в районе 1 (см., например, работу [Piantadosi, 2014], в которой проведено исследование для 18 европейских языков). Также было показано, что закон Ципфа соблюдается не только для отдельных слов, но и для n-грамм слов и отдельных букв [Ha et al., 2003; Hachisuka et al., 2014].

Большинство исследований посвящено поведению отельных слов и групп слов в языке в целом. В работе [Кочеткова и др, 2016] мы уже рассмотрели распределение степенного показателя в распределении Ципфа для групп слов в специальном корпусе. В данной работе мы продолжаем наше исследование для разных синтаксических групп слов русского языка. В основе работы лежит предположение, что знания о форме распределения степенных показателей может быть использовано для извлечения коллокаций из текстов.

2 Статистическое распределение степенных показателей

В предыдущих работах было показано, что большинство тематически окрашенных коллокаций в тексте образуются с существительными в сочетании с прилагательными, другими существительными и реже с глаголами [Tutubalina, 2016]. В связи с этим в данной работе мы ограничимся только этими тремя видами коллокаций. Так как в работе [Кочеткова, 2016] было показано, что форма распределения сохраняется при переходе от униграмм в 4-граммам, то в данной статье мы будем рассматривать лишь биграммы.

Для расчетов использовалась следующая формула распределения Ципфа:

$$f(n) = A_1 * 2^{-an},$$

где a — степенной показатель распределения, а A_1 — абсолютная встречаемость самого частотного первого слова, входящего в состав пары.

2.1 Исходные данные

Для проведения экспериментов мы использованные размеченные коллекции со снятой омонимией. Омонимия снималась с использованием программы, описанной в [Рысаков, 2015]. При разметке использовалась языковая модель СинТагРус, то есть, например, притяжательные местоимения считались как прилагательные. В качестве текстовых коллекций использовались статьи специализированного журнала «САПР и графика» (7.2 млн токенов), подборка романов по тематике «мастера детектива» (34.5 млн токенов) и тексты, входящие в Библиотеку Мошкова (881 млн токенов). Вслед за [Tutubalina, 2016], мы извлекали из коллекций пары слов, подходящие под один из следующих шаблонов: прилагательное + существительное, существительное + существительное, существительное + глагол и

глагол+существительное. Все слова ставились в начальную форму.

Для каждой пары фиксировалось второе слово после чего отбрасывались те слова, с которыми встретилось менее 5 разных начальных форм в первой позиции, каждая из которых встретилась не менее 3 раз. Для оставшихся слов рассчитывался степенной показатель распределения Ципфа по первым словам при фиксированном втором. Таким образом, для одного шаблона получалось столько степенных показателей, сколько разных начальных форм встретилось на второй позиции, причем сами показатели рассчитывались для изменяющихся слов в первой позиции.

Приведем пример рассматриваемых распределений. Для корпуса «САПР и Графика» для шаблона «прилагательное+существительное» для зафиксированного существительного «ЕДИНИЦА» получалось следующее распределение частот для встретившихся с ним прилагательных: СБОРОЧНЫЙ, 621; СТРУКТУРНЫЙ, 19; МОНТАЖНЫЙ, 18; КАЖДЫЙ, 17. Степенной показатель для данного распределения равен 4.686.

2.2 Полученное распределение степенных показателей

Форма полученных распределений визуально похожа на распределение Пуассона. Мы не проводили статистическую проверку данной гипотезы, но будем использовать ее в дальнейшем в качестве истинной. На рис. 1 показано, что соответствие не является полным, однако, это лучшее совпадение из рассмотренных нами распределений. Красным и зеленым показаны разные значения коэффициента распределения Пуассона. В табл. 1 также приведены средние значения для каждой из коллекций.

Заметим, что полученные результаты согласуются с другими, полученными ранее [Piantadosi, 2014].

Табл. 1. Средние значения степенного показателя распределения Ципфа

	Библ. Мошкова	**Мастера детектива**	**САПР и Графика**
Сущ+прил	0/946	0.944	1.013
Сущ+Сущ	0/836	0.854	0.960
Сущ+гл		0.936	0.947
Гл+Сущ	0/929	0.949	0.860

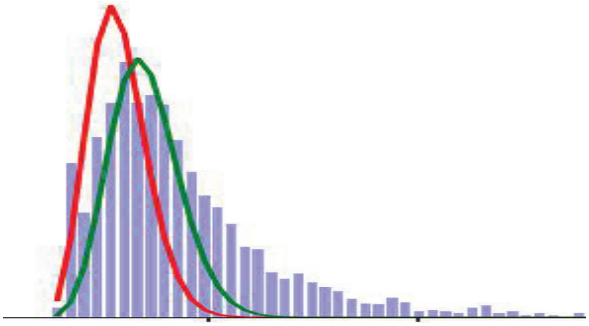

Рис. 1. Распределение степенных показателей распределения Ципфа (син.) и распределение Пуассона для сравнения (кр. и зел.).

3 Выделение коллокаций

Для выделения коллокаций мы использовали следующую гипотезу. Для распределения Пуассона значение дисперсии равно среднему значению. Слова, для которых показатель распределения Ципфа близок к среднему, не представляют для нас интерес. Однако если отступить от среднего на значение дисперсии, то это будет означать, что первое слово встречается в 3.5-4 раза чаще, чем второе слово, что является серьезным отклонением. В некоторых случаях это будет означать, что самое частотное сочетание встречается чаще, чем все остальные пары, образуемые данным словом.

Для проверки данной гипотезы мы отобрали все слова, для которых показатель Ципфа превышал 2. Далее, для них были взяты самые частотные пары слов. Ниже приведены списки пар с самым высоким показателем Ципфа и находящихся на нижней границе (удвоенное среднее). Для шаблона *прилагательное+существительное* были получены пары, представленные на табл. 2.

Заметим, что наличие местоимений в выдаче изрядно зашумляет ее. Однако если убрать местоимения, результаты будут достаточно чистыми. Так, например, при устранении местоимений первые десять пар для Библиотеки Мошкова будут следующие: *всячина всякий, замыкание короткий, сошка мелкий, разбирательство судебный, ловля рыбный, дровосек железный, желток яичный, хрыч старый, гвинея новый, убор головной*.

Табл. 2. Примеры пар для сущ+прил.

Существит.	Прилагат.	Коэф. Ципфа
Мастера детектива (среднее 0,94)		
Начало	Сам	6,35
Видимость	Весь	5,35
Будка	Телефонный	4,83
Крючек	Спусковой	4,45
Главное	Сам	4,44
Возмущение	Свой	1,90
Сервиз	Чайный	1,90
Вонь	Этот	1,90
Допрос	Перекрестный	1,90
Репетиция	Генеральный	1,89
Библиотека Мошкова (среднее 0,95)		
Превосходительство	Ваш	9,15
Сиятаельство	Ваш	8,78
Всячина	Всякий	7,65
Благородие	Ваш	7,61
Светлость	Ваш	7,11
Евразия	Весь	1,90
Цветик	Мой	1,90
Отпущение	Полный	1,90
Долг	Свой	1,90
Разница	Какой	1,90
САПР и Графика (среднее 1,01)		
Единица	Сборочный	4,69
Генерация	Автоматический	4,58
Обеспечение	Программный	4,46
Ссылка	Внешний	4,15
Выступение	Свой	3,99
Расположение	Взаимный	2,04
Развитие	Дальнейший	2,03
Индустрия	Строительный	2,03
Квалификация	Высокий	2,02
Версия	Новый	2,02

Табл. 3. Примеры для сущ+гл

Существит.	Прилагат.	Коэф. Ципфа
Мастера детектива (среднее 0,95)		
Прощение	Просить	6,90
Сознание	Терять	5,03
Значение	Иметь	4,37
Воля	Давать	4,27
Плечо	Пожимать	4,23
Эмили	Сказать	1,96
Профессионал	Быть	1,94
Сообщник	Быть	1,93
Контрабанда	Заниматься	1,92
Маршалл	Сказать	1,92
Библиотека Мошкова (среднее 0,93)		
Участие	Принимать	7,29
Прощение	Просить	5,34
Реверанс	Делать	5,33
Мах	Давать	5,32
Обыкновение	Иметь	5,24
Хаджи	Сказать	1,87
Уловка	Быть	1,87
Хладнокровие	Сохранять	1,86
Пронин	Сказать	1,86
Авантюра	Быть	1,86
САПР и Графика (среднее 0,86)		
Вывод	Делать	3,97
Внимание	Обращать	3,37
Эффективность	Повышать	3,06
Срок	Сокращать	2,68
ГОСТ	Соответствовать	2,67
Задача	Решать	1,82
Поставка	Осуществлять	1,80
Решение	Принимать	1,79
Часть	Являтья	1,76
Точность	Повышать	1,76

В Табл. 3 также приведены примеры для пар *существительное+глагол* (Табл. 3).

Мы провели сравнение выражений, выделенных по степенному показателю распределения Ципфа, с результатами, получаемыми при помощи меры MI=p(a,b)/p(a)*p(b). Из первой сотни коллокаций с самым большим степенным показателем было найдено лишь 46, причем для показателя MI они входили в лучшем случае во вторую или третью сотню.

Приведем в качестве примера десять пар *прилагательное+существительное* с самыми большими значениями показателя MI для коллекции «САПР и Графика»:
ГАДКИЙ УТЕНОК, 10.28,
РЫБООБРАБАТЫВАЮЩИЙ ТРАУЛЕР,10.06,
ДЕЛАНЫЙ ВОРОНОЙ, 10.06,
НЕОПАЛИМЫЙ КУПИНА, 9.87,
ПЛАВЛЕНЫЙ КВАРЦ, 9.87,
ПОБЕДОНОСНЫЙ ШЕСТВИЕ, 9.77,
БЕРИНГОВ ПРОЛИВ, 9.72,
ПОБЕДНЫЙ РЕЛЯЦИЯ, 9.72,
МЕКСИКАНСКИЙ ЗАЛИВ, 9.65,
ПОКОРНЫЙ СЛУГА, 9.59

Аналогично для коллекции “Мастера детектива”:
СОСЛАГАТЕЛЬНЫЙ НАКЛОНЕНИЕ, 11.30,
ПЕРИФЕРИЧЕСКИЙ НЕВРИТ, 10.71,
ПРЕФРОНТАЛЬНЫЙ ЛОБОТОМИЯ, 10.61,
ДИАЛЕКТИЧЕСКИЙ МАТЕРИАЛИЗМ,10.27
ГОТСКИЙ АЛЬМАНАХ, 10.12
МАКОВЫЙ РОСИНКА, 10.05
ТОКАРНЫЙ СТАНОЧЕК, 9.92
ОРЛЕАНСКИЙ ПЛОВ, 9.73
САУДОВСКИЙ АРАВИЯ, 9.68
ЯЧМЕННЫЙ ОТВАР, 9.67

Подобные коллокации сложно назвать характерными для выбранной предметной области.

Однако предложенный метод игнорирует слова, для которых можно указать более одного часто употребляемого слова. Так, например, пары *Северная Корея* и *Южная Корея* встречаются в текстах с примерно одинаковой частотой, тогда как третье слово, следующее за ними по частоте будет встречаться значительно реже. В подобных ситуациях степенной показатель распределения Ципфа падает ниже выбранного порога в 2. В связи с этим мы исследовали не только значение степенного показателя, но и значение отношения между частотой встречаемости первого и второго слова, а если отношение их частот было близко к единице, то второго и третьего. График на рис. 2 показывает взаимосвязь между степенным показателем распределения Ципфа и значением описанного выше отношения. Данные, показанные на рисунке, были получены для коллекции «Мастера детектива» для сочетаний прилагательных и существительных. Среднее отношение составило 2,35, дисперсия 1,77. Ось X показывает значения соотношения частот в логарифмической шкале (значения соотношения достигали нескольких сотен), ось Y показывает значения степенного коэффициента.

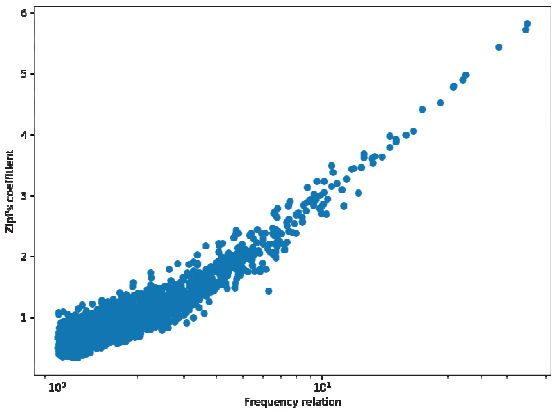

Рис. 2. Зависимость степенного показателя коэффициента Ципфа (ось Y) от соотношения частот первых слов (ось X в логарифмическом масштабе)

График показывает, что при отношении частот выше 2 поведение функции меняется. Но мы выделим на рис. 2 три зоны: соотношение частот меньше 4 (среднее плюс дисперсия), соотношение частот больше 4, но степенной показатель меньше 2 и степенной показатель выше 2.

Мы исследовали сочетания, обладающие соотношением частот выше 4. Для коллекции «Мастера детектива» в первой десятке оказались такие сочетания существительных и прилагательных как:
САМ НАЧАЛО, 91.8
ВЕСЬ ВИДИМОСТЬ, 44.3
БУДКА ТЕЛЕФОННЫЙ, 33
САМ ГЛАВНОЕ, 25.3
КРЮЧЕК СПУСКОВОЙ, 25.2
САМ СЕБЯ, 24.5
ЧАСТНЫЙ ДЕТЕКТИВ, 24.3
СУДЕБНЫЙ РАЗБИРАТЕЛЬСТВО, 23,2
ЗУБНОЙ ЩЕТКА, 22.5
ДВЕРНОЙ ПРОЕМ, 21.5

Для коллекции «САПР и Графика» в первой десятке оказались следующие сочетания:
СБОРОЧНЫЙ ЕДИНИЦА, 32.7

АВТОМАТИЧЕСКИЙ ГЕНЕРАЦИЯ, 27.6
ПРОГРАММНЫЙ ОБЕСПЕЧЕНИЕ 24.6
ВЕСЬ ЭТОТ, 21,2
ВЕСЬ МИР, 19.4
ВНЕШНИЙ ССЫЛКА, 19.2
СВОЙ ВЫСТУПЛЕНИЕ, 18.5
РАБОЧИЙ МЕСТО, 17.9
ВЫСОКИЙ КАЧЕСТВО, 15.4
ОПЕРАТИВНЫЙ ПАМЯТЬ, 15.3

Если рассчитать отношение частот между вторым и третьим вариантом, то в первую десятку для коллекции «Мастера детектива»
ПОРА ТОТ СЕЙ
ГИЛЬЗА СТРЕЛЯНЫЙ ПУСТОЙ
СИДЕНЬЕ ЗАДНИЙ ПЕРЕДНИЙ
ПОМАДА ГУБНОЙ ГУБНЫЙ
КАРТА СВОЙ ВЕСЬ
ОБЪЯТИЕ МОЙ СВОЙ
КОЛЯСКА ДЕТСКИЙ ИНВАЛИДНЫЙ
СОБЕСЕДНИК СВОЙ МОЙ
СПИНКА ВЫСОКИЙ ПРЯМОЙ
СКВАЖИНА ЗАМОЧНЫЙ НЕФТЯНОЙ

Для коллекции «САПР и Графика»:
ПОРА СЕЙ ТОТ
ЭКРАН ВИДОВОЙ СЕНСОРНЫЙ
МАКЕТ ЭЛЕКТРОННЫЙ ЦИФРОВОЙ
ДОРОГА АВТОМОБИЛЬНЫЙ ЖЕЛЕЗНЫЙ
КРИЗИС ФИНАНСОВЫЙ ЭКОНОМИЧЕСКИЙ
УЧАСТИЕ АКТИВНЫЙ НЕПОСРЕДСТВЕННЫЙ
ПОСТАВКА СТАНДАРТНЫЙ БАЗОВЫЙ
СТАНЦИЯ РАБОЧИЙ ГРАФИЧЕСКИЙ
РЯД ЦЕЛЫЙ МОДЕЛЬНЫЙ
ВВОД РУЧНОЙ ПОВТОРНЫЙ

Наконец, в последнюю десятку перед соотношением частот 4 вошли следующие сочетания. Для коллекции «Мастера детектива»:
ЗАМОЧНЫЙ СКВАЖИНА
ЭТОТ ВОНЬ
СВОЙ ПОЖИТКИ
ЛЕВЫЙ ПЕДАЛЬ
ЭТОТ ИНЦИДЕНТ
ЦЕЛЫЙ СЕРИЯ
ЭТОТ ШАНТАЖ
ИНЖЕНЕРНЫЙ АТЛАС
ТОМАТНЫЙ СОУС
СВОЙ ПАСТЬ
Для коллекции «САПР и Графика»:
ГЛАВНЫЙ ИНЖЕНЕР
УЧЕБНЫЙ ЦЕНТР
ШИРОКИЙ АССОРТИМЕНТ
ЭТОТ ТЕРМИН
УНИВЕРСАЛЬНЫЙ МАРКЕР
КОМАНДНЫЙ СТРОКА
ТОРГОВЫЙ МАРКА
ТУРБИННЫЙ ЛОПАТКА
КАКОЙ ТОТ
УДАРНЫЙ ВОЛНА

Но в некоторых случаях появлялись артефакты. Так, в коллекции беллетристики выделилось сочетание «РОЗОВАЯ ЧУМА», так как при общем количестве упоминаний меньше 200, в одном из произведений данное словосочетание встретилось 166 раз, так как являлось важной частью повествования.

Отметим еще одно свойство предложенного метода: в общем случае данная мера не является симметричной. Пусть ведется расчет для фиксированного слова A и пусть для него выделилось наиболее частотное слово B. В этом случае наиболее частотным словом для слова B не обязательно окажется слово A. Так, например, для слова *будка* наиболее частым является слово *телефонный*, но для слова *телефонный* наиболее частым является слово *звонок*. Но, например, фраза *слуга покорный* ассоциативна в обе стороны, то есть каждое из двух слов наиболее тесно связано именно с другим словом. Подобное свойство ближе к ассоциативному ряду, имеющемуся в голове у человека.

Заметим также, что предварительные эксперименты показали метод работает не только с парами слов, но и с сочетаниями большего размера. Для этого необходимо рассматривать распределение не для одного слова, а для нескольких при одном зафиксированном.

4 Заключение

Предложенный метод успешно выделять коллокации из текстов большого объема. Это же свойство является минусом метода — для его работы требуются относительно большие коллекции, которые не всегда доступны. При этом следует отметить, что, например, коллекция статей монотематического журнала за несколько лет не является однозначной редкостью, то есть метод применим для работы с коллекциями реального размера. В дальнейшем мы планируем более подробно изучить вопрос выделения коллокаций из трех и более слов.

Список литературы

Кочеткова Н. А., Клышинский Э. С., Ермаков П. Д. 2016. *Подчиняются ли составные конструкции закону Ципфа?* Системный администратр, №11, 89-95.

Маслов В. П., Маслова Т. В. 2006. *О законе Ципфа и ранговых распределениях в лингвистике и семиотике*, Матем. заметки, том 80, выпуск 5, 718–732.

Рысаков С. В. 2015. *Методы борьбы с омонимией*, Системный администратор, №10.

i Cancho R. F., Riordan O., Bollobás B.. 2005. *The consequences of Zipf's law for syntax and symbolic reference*. Proceedings of Royal Society, B, 272, 561–565. doi:10.1098/rspb.2004.2957

Piantadoi S. T. 2014. *Zipf's word frequency law in natural language: A critical review and future directions.* Psychon Bullet Review, 21, 1112-1130.

Tutubalina, E. V., Braslavski, P. I. 2016. *Multiple features for multiword extraction: A learning-to-rank approach*. In Proc. of the Dialog-2016 Conf: *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, 782-791.

Hachisuka T., Kaplanyan A.S., Dachsbacher C. 2014. *Extension of Zipf's Law to Word and Character*

Ha L. Q., Sicilia-Garcia E. I., Ming J., Smith F. J. *N-grams for English and Chinese*. Computational Linguistics and Chinese Language Processing. Vol. 8, No. 1, February 2003 , pp. 77-102.