Математические модели в лингвистике

Коллокации и их автоматическое определение.

Мати Пентус, Александр Пиперски, Алексей Сорокин

МГУ им. М. В. Ломоносова, межфакультетский курс,
осенний семестр 2017–2018 учебного года

Композициональность

- В языке потенциально неограниченное число предложений и тем более текстов.
- Однако носитель языка способен восстановить значение элементов более высокого уровня.
- Значение предложений сводится к значению словосочетаний, а значения словосочетания — к значению слов...
- Естественный язык обладает композициональностью:

"Значение сложного выражения есть функция значений его частей и синтаксических правил, соединяющих эти части."

Пример композициональности

- То есть определение должно единообразно модифицировать значение существительного, к которому оно относится:
  - *дом* $\rightarrow$ *красный дом*,
  - *сапог* $\rightarrow$ *красный сапог*,
  - *куртка* $\rightarrow$ *красная куртка*
- Однако не всегда смысл изменяется одинаковым образом:
  - *уголок* $\rightarrow$ *красный уголок*,
  - *армия* $\rightarrow$ *Красная Армия*,
  - *книга* $\rightarrow$ *Красная Книга*
- То есть композициональность часто нарушается.

Нарушения композициональности

- Основной пример: идиомы и фразеологизмы.
- Классификация по В. В. Виноградову:
  - Фразеологические сращения (смысл выражения не восстанавливается по смыслу его компонентов):
    “ни в зуб ногой”, *“собаку съесть”*.
  - Фразеологические единства (смысл частично восстанавливается по значению компонентов):
    “делать из мухи слона”, *“плыть по течению”*, *“звонить в колокола”*
  - Фразеологические сочетания (устойчивое сочетание слов, чей смысл восстановим из значения компонентов, одно из слов менее свободно чем другое):
    “потупить взгляд”, *“нанести урон”*, *“обрести спокойствие”*.

Нарушения композициональности

- Основной пример: идиомы и фразеологизмы.
- Классификация по А. Баранову и Д. Добровольскому:
  - Идиомы (“*шишка на ровном месте*”),
  - Коллокации (“*зло берёт*”),
  - Пословицы (“*цыплят по осени считают*”),
  - Грамматические фразеологизмы (“*во что бы то ни стало*”),
  - Синтаксические фразеологизмы (X *он и в Африке* X).
- Существуют и другие классификации и определения.

Варианты определения коллокаций

- “неслучайное сочетание двух и более лексических единиц, характерное как для языка в целом, так и определенного типа текстов”(Ягунова, Пивоварова, 2010).
- “семантическая фразема, такая что в состав её означаемого входит означаемое одной из лексем в качестве семантической доминанты и означаемое второй лексемы в качестве некоторого дополнительного компонента, причём вторая лексема выбирается в зависимости от первой” (Иорданская, Мельчук, 2007).
- “lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other” (Bartsch, 2004).

Вычислительный подход к коллокациям

- “Collocations of a given word are statements of the habitual and customary places of that word.” (Firth, 1957).
- “Collocations [are] recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages” (Smadja, 1993).
- “A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon.” (Evert, 2007)
- “I use collocation thus as a generic term whose specific meaning can be narrowed down according to the requirements of a particular research question or application.” (Evert, 2004)

Основные свойства коллокаций

Основные свойства коллокаций:

- Некомпозициональность — значение коллокации не сводится к значению элементов:

$$\textit{сыграть в ящик} \neq \textit{сыграть} + \textit{ящик}$$

- Нерегулярность — коллокации не порождаются в соответствии со стандартной моделью.
- Устойчивость — элементы коллокации нельзя заменить на синонимичные.

$$\textit{крепкий чай} \neq \textit{сильный чай}$$

- Частотность.

Применения коллокаций

Какие есть применения у коллокаций?

- Лексикология (как компьютерная, так и традиционная).
- Подготовка словарей (как дву-, так и одноязычных).
- Автоматический перевод:

A leopard cannot change its spots

$$\Updownarrow$$

Чёрного кобеля не отмоешь добела

Применения коллокаций

- Уточнение языковых моделей.
- Информационный поиск:

  Дума наложила вето . . . $\leftrightarrow$ *Дума отвергла . . .*
- Автоматическая разметка (морфологическая, синтаксическая и т.д):
  - Коллокации могут нарушать морфологические закономерности:

    задать перцу, *ни в зуб ногой*, . . .
  - Может использоваться нестандартный порядок слов:

    справедливости ради, *бил озноб*.
  - Морфология и синтаксис элементов коллокации известны:

    во что бы то ни стало, *кто в лес, кто по дрова*

Наш подход к коллокациям

Мы придерживаемся статистического подхода к коллокациям:

Определение коллокаций

Коллокация — это сочетание слов (возможно, синтаксически связанных), частотность которого существенно выше, чем была бы в предположении о независимости его компонент.

Математически,

$$c(w_1 w_2) \gg c(w_1)c(w_2),$$

где $c(\alpha)$ — относительная частота сегмента α в корпусе.
То есть надо вычислять *взаимную информацию* слов w_1 и w_2:

$$MI(w_1, w_2) = \log \frac{c(w_1 w_2)}{c(w_1)c(w_2)}$$

Этап 1: создание конкорданса

Создание конкорданса при помощи SketchEngine:

Query (в)-s **34,220,576** (27,291.0 per million) | using first 10,000,000 lines only (use random).

Page 1 of 500,000 Go Next | Last

doc#0	другого года, даже если виноград собран	в	пределах одного виноградника. Истинные
doc#0	загородными домами, с удовольствием устраивают	в	них винные погреба. Вино - напиток
doc#0	хранения оборудуют специальные винные погреба,	в	которые в идеале хорошо бы установит систему
doc#0	оборудуют специальные винные погреба, в которые	в	идеале хорошо бы установит систему климат
doc#0	очень красиво. Для хранения бутылок	в	винном погребе можно использовать деревянные
doc#1	тебе и уважуха)) Вот только минус подборки	в	том, что ты искал только по тем группам
doc#1	Nile. По моему скромному мнению уделывает	в	плане дизайна большую часть представителей
doc#1	анимированная змея и сменные бэкграунды	в	стиле древнего Египта. А это я только тыкнул
doc#2	являются нелинейными. Оттенки отражают различие	в	свойствах. Количество нелинейных слоев
doc#2	слоев можно выбирать произвольное, но авторы	в	своих дальнейших вычислениях ограничились
doc#2	которую можно определить интенсивность волны	в	заданном слое N . Эта функция находится
doc#2	нелинейного уравнения Шрёдингера. Подробности см.	в	тексте. Рисунок из обсуждаемой статьи в
doc#2	в тексте. Рисунок из обсуждаемой статьи	в	Phys. Rev. Lett. Итальянские физики-теоретики
doc#2	рассчитали параметры структуры, которую	в	дальнейшем можно использовать для создания
doc#2	электромагнитные или акустические волны	в	одном направлении и полностью их блокировать
doc#2	полностью их блокировать, когда они движутся	в	противоположную сторону. В отличие от предыдущих
doc#2	они движутся в противоположную сторону.	В	отличие от предыдущих теоретических моделей
doc#2	устройство, которое пропускает электрический ток	в	одном направлении и не позволяет ему течь
doc#2	одном направлении и не позволяет ему течь	в	противоположном. Наряду с транзистором
doc#2	света или звука, могли бы использоваться	в	тепло- и звукоизоляции, направленной передаче

Этап 2: выделение коллокаций

Выделение коллокаций при помощи SketchEngine:

user: Dr. Alexey Sorokin corpus: ruTenTen [2011, sample] Search

Concordance
Word List
Word Sketch
Thesaurus
Find X
Sketch-Diff
Corpus Info

< Concordance
Sample
Filter
Overlaps
1st hit in doc
Frequency
Node tags

Collocation candidates

Attribute: word In the range from: 0 to: 1

Minimum frequency in corpus: 5

Minimum frequency in given range: 5

Show functions: T-score, MI, MI3, log likelihood, min. sensitivity, logDice

Sort by: T-score, MI, MI3, log likelihood, min. sensitivity, logDice

Make Candidate List Save Options

Этап 3: список коллокаций

Получаемый список наиболее частых коллокатов:

Collocation candidates

Page 1 Go Next >

	Freq	MI
P | N Кефиристане	5	6.970
P | N Крайске	5	6.970
P | N Торвил	5	6.970
P | N Красносельске	6	6.970
P | N МЦИПе	5	6.970
P | N энерготическом	5	6.970
P | N Мурниеках	5	6.970
P | N Набадвипе	5	6.970
P | N Эспиньолях	26	6.970
P | N Падьюке	7	6.970
P | N Хостинг-Центре	5	6.970
P | N Стегеборге	5	6.970
P | N Уранополь	7	6.970
P | N буржнете	8	6.970
P | N квадрапространстве	10	6.970
P | N Печникове	8	6.970
P | N мутазилизме	5	6.970
P | N Клариды	5	6.970
P | N Аксапте	6	6.970

Анализ результатов

Пусть словоформа X встречается только внутри биграммы в X, тогда

$$MI(\text{в}, X) = \log \frac{p(\text{в } X)}{p(\text{в})p(X)} = -\log p(\text{в}) = 5.19$$

Получилось 6.79, потому что “Although it is labeled as standard "Mutual Information", Sketch Engine actually uses a slightly different calculation: «a scaled version of Dice»”. (A. Kilgariff)

Недостатки взаимной информации

- Пусть корпус содержит 1000000 слов.
- Минимальная ненулевая частота биграммы: 0.000001.
- Значит, для биграммы XY, где каждое слово имеет частоту 100, получим

$$MI(X, Y) = \log_2 \frac{0.000001}{0.0001 * 0.0001} = \log_2 100 = 6.64$$

- Метод взаимной информации предпочитает редкие слова!
- Для них любая случайная биграмма даёт большое значение информации.
- Вывод: надо рассматривать не абсолютную величину отличия, а его значимость.

Математическое отступление: проверка гипотез

- Мы будем проверять не то, насколько сильно отклоняется значение, а насколько невероятно такое отклонение.
- Будем считать, что слово w появляется с вероятностью, равной его относительной частоте:

$$p(u = w) = c(w) = \frac{n(w)}{N}, \text{ где } N \text{ — размер корпуса}$$

- Рассмотрим гипотезу H_0, что появление слов w_1 и w_2 независимо (это означает, что $w_1 w_2$ не образует коллокацию).
- Тогда биграмма $w_1 w_2$ распределена биномиально с вероятностью

$$p_0 = p(u_1 = w_1)p(u_2 = w_2) = c(w_1)c(w_2)$$

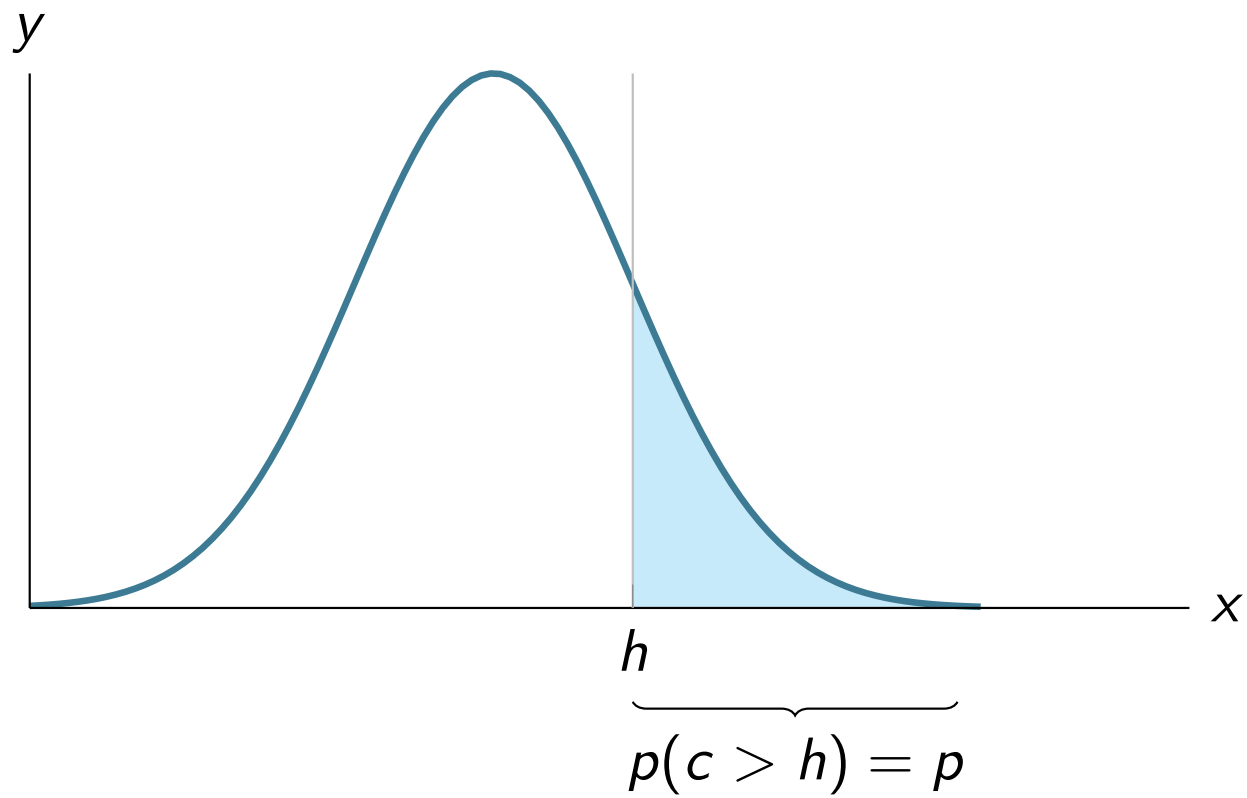
- Гипотеза отсутствия коллокации: $H_0 : p(w_1 w_2) = p_0$.
- Гипотеза наличия связи (присутствия коллокации): $H_1 : p(w_1 w_2) > p_0$.

Уровни значимости

- Пусть корпус содержит N биграмм, а биграмма $w_1 w_2$ встречается n раз. Нам требуется подобрать порог частоты h так, чтобы вероятность ошибочного принятия альтернативной гипотезы H_1 (“ложной тревоги”) была не выше некоторого значения p.
- То есть при принятии гипотезы мы должны ошибаться с вероятностью не больше p.
- Предположим, что H_0 верна, тогда в качестве порогового значения для относительной частоты $c(w_1 w_2)$ нужно взять такое h, что
$$p(c > h | H_0) \leqslant p.$$
- Тогда H_1 принимается если $c(w_1 w_2) \geqslant h$ или, что то же самое,
$$p(c > c(w_1 w_2) | H_0) \leqslant p.$$

Уровни значимости

- H_0 отвергается (принимается альтернативная гипотеза о наличии коллокации), если c попадает в проекцию заштрихованной области.



Уровни значимости

- Найдём вероятность того, что $p(c \geqslant c(w_1 w_2))$ при условии H_0, то есть при условии, что вероятность данной биграммы равна $p_0 = c(w_1)c(w_2)$.
- Она задаётся формулой $\sum_{m=n}^{N} C_N^m p_0^n (1-p_0)^{N-n}$, где m — число вхождений биграммы $w_1 w_2$ в корпус.
- Если $\sum_{m=n}^{N} C_N^m p_0^m (1-p_0)^{N-m} \leqslant p$, то гипотезу H_0 можно отвергать с риском ошибиться не более, чем p.
- То есть в качестве показателя наличия коллокации можно брать величину

$$1 - \sum_{m=n}^{N} C_N^m p_0^m (1-p_0)^{N-m}$$

- Недостаток: невозможно считать.

Статистический подход к извлечению коллокаций

- Общая методология: подобрать некоторую статистику z с функцией распределения F, монотонно зависящую от числа вхождений биграммы $w_1 w_2$.
- Тогда порогом наличия коллокации является либо само значение этой статистики z_0, либо $F^{-1}(z_0)$:

  $z \geqslant z_0$ — принимаем гипотезу о наличии коллокации,
  $z < z_0$ отвергаем гипотезу

- Лучше всего брать легко вычислимую статистику с известной функцией распределения (например, нормальным или χ^2).

Нормальное приближение

- Однако распределение нормированной суммы большого числа независимых бернуллиевых величин очень напоминает нормальное. Нельзя ли этим воспользоваться?
- Если предположить, что $p(w_1 w_2) = p_0$, то величина $z(w_1 w_2) = \frac{c(w_1 w_2) - p_0 N}{\sqrt{p_0 N}}$ стремится по распределению к стандартной нормальной величине с плотностью

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

- Проблема в том, что при малых p_0 сходимость очень медленная, а знаменатель очень мал, поэтому пользоваться этим приближением нельзя.

Нормальное приближение

- В работе [Church, 1991] была предложена величина

$$t(w_1 w_2) = \frac{c(w_1 w_2) - p_0 N}{\sqrt{c(w_1 w_2)}}$$

- Данная статистика называется t-мерой для коллокаций

Извлечение коллокаций с помощью t-меры

Получаемый список наиболее частых коллокатов:

P | N том	182,902	407.129
P | N этом	149,908	361.273
P | N течение	79,760	274.793
P | N качестве	77,418	270.522
P | N России	82,347	261.724
P | N результате	64,477	246.876
P | N рамках	62,878	244.362
P | N соответствии	62,570	243.973
P | N случае	61,178	229.087
P | N конце	51,142	218.857
P | N области	56,473	214.022
P | N Москве	48,054	212.279
P | N виде	47,541	209.237
P | N связи	46,249	202.241
P | N целом	42,123	199.421
P | N настоящее	42,080	198.786
P | N частности	41,590	198.225
P | N котором	42,715	197.801
P | N этой	46,253	190.531
P | N зависимости	37,361	187.205
P | N основном	36,151	184.666
P | N ходе	35,978	184.058
P | N мире	36,065	181.299
P | N нем	34,109	176.181

Недостаток t-меры

- t-мера зависит лишь от $c(w_1 w_2)$ и произведения $e(w_1 w_2) = c(w_1)c(w_2)$.
- Воображаемый пример (Evert, 2007):

$$\begin{aligned} C(\textit{Iliad}) &= 10, & C(\textit{the}) &= 100000, \\ C(\textit{must}) &= 1000, & C(\textit{also}) &= 1000, \\ C(\textit{the Iliad}) &= 10, & C(\textit{must also}) &= 10 \\ E(\textit{the Iliad}) &= 10^6, & E(\textit{must also}) &= 10^6 \end{aligned}$$

- То есть с точки зрения t-меры данные биграммы в одинаковой степени являются коллокациями.
- Различие: *must* и *also* встречаются со многими другими словами, *Iliad* — только с *the*.
- Следовательно, одной “положительной” информации недостаточно.
- Выход: нужно рассматривать всю таблицу сопряжённости биграм w_1 и w_2.

Таблицы сопряжённости

Таблица сопряжённости показывает совместное распределение вхождений w_1 и w_2.

	w_2	$\neg w_2$	
w_1	n_{11}	n_{12}	L_1
$\neg w_1$	n_{21}	n_{22}	L_2
	R_1	R_2	N

$$
\begin{aligned}
n_{11} &= c(w_1, w_2) \\
n_{12} &= c(w_1, \cdot) - c(w_1, w_2) = \sum_{w \in T_2, w \neq w_2} c(w_1, w) \\
n_{21} &= c(\cdot, w_2) - c(w_1, w_2) = \sum_{w \in T_1, w \neq w_1} c(w, w_2)
\end{aligned}
$$

$$
\begin{aligned}
n_{22} &= c(\cdot, \cdot) - c(w_1, \cdot) - c(\cdot, w_2) + c(w_1, w_2) \\
&= \sum_{\substack{u \in T_1, v \in T_2 \\ u \neq w_1, v \neq w_2}} c(u, v)
\end{aligned}
$$

Здесь T_1, T_2 — множества слов, откуда берутся w_1, w_2.

Меры коллокативности для таблиц сопряжённости

Пусть $e_{11} = p_1 p_2,\ e_{12} = p_1 \overline{p}_2,\ e_{21} = \overline{p}_1 p_2,\ e_{22} = \overline{p}_1 \overline{p}_2$ — ожидаемые значения элементов таблицы.

$$MI_a(w_1, w_2) = \sum_{i,j} c_{ij} \log_2 \frac{c_{ij}}{e_{ij}}$$ — средняя взаимная информация

$$\chi^2(w_1, w_2) = \sum_{i,j} \frac{(c_{ij} - e_{ij})^2}{e_{ij}}$$ — статистика χ^2

$$OR(w_1, w_2) \approx \frac{c_{11}}{c_{21}} / \frac{c_{12}}{c_{22}} = \frac{(c_{11} + 0.5)(c_{22} + 0.5)}{(c_{12} + 0.5)(c_{21} + 0.5)}$$

$$D_1(w_1, w_2) = \frac{n_{11}}{L_1} \qquad D_2(w_1, w_2) = \frac{n_{11}}{R_1}$$

$$Dice(w_1, w_2) = \frac{2 D_1 D_2}{D_1 + D_2} = \frac{2 n_{11}}{L_1 + R_1}$$

И ещё примерно 80 вариантов (Pecina, 2005)

Таблица мер коллокативности

1. Mean component offset	$\frac{1}{n}\sum_{i=1}^{n} d_i$
2. Variance component offset	$\frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$
3. Joint probability	$P(xy)$
4. Conditional probability	$P(y\|x)$
5. Reverse conditional prob.	$P(x\|y)$
*6. Pointwise mutual inform.	$\log \frac{P(xy)}{P(x*)P(*y)}$
7. Mutual dependency (MD)	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
8. Log frequency biased MD	$\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
9. Normalized expectation	$\frac{2f(xy)}{f(x*)+f(*y)}$
*10. Mutual expectation	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
11. Salience	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
12. Pearson's χ^2 test	$\sum_{ij} \frac{(f_{ij}-f_{ij}')^2}{f_{ij}'}$
13. Fisher's exact test	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
14. t test	$\frac{f(xy)-f'(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
15. z score	$\frac{f(xy)-f'(xy)}{\sqrt{f'(xy)(1-(f'(xy)/N))}}$
16. Poisson significance measure	$\frac{f(xy)-f'(xy)\log f(xy)+\log f'(xy)!}{\log N}$
17. Log likelihood ratio	$-2\sum_{ij} f_{ij} \log \frac{f_{ij}}{f_{ij}'}$
18. Squared log likelihood ratio	$-2\sum_{ij} \frac{\log f_{ij}^2}{f_{ij}'}$

Сравнение различных мер

- *MI* — чувствительна к случайным совпадениям, находит тематические коллокации.
- t-мера — смещена в сторону частотных пар, находит устойчивые сочетания, но не имеет строгого обоснования.
- *Dice*-мера — находит симметричные устойчивые сочетания (w_1 и w_2 встречаются только вместе).
- Отношение вероятностей не отличает положительную и отрицательную ассоциацию.
- Выбор оптимальной меры зависит от задачи (например, нужны ли симметричные или несимметричные коллокации).

Варианты определения коллокаций

- Мы считали статистику “по словам”, используя счётчики для словоформ.
- Можно считать “по леммам”, используя счётчики для лексем.
- Например, для биграммы “придать значение” это позволит отследить все формы глагола:
  - “придаёт значение”
  - “придала значение”
  - “придадим значение”
  - и даже “не придаёт значения”.

Сравнение коллокаций по словам и леммам.

Коллокации глагола *придать* “по словам”.

Collocation candidates

Page 1 Go Next >

		Frequency	T-score
P | N	ему	715	26.522
P | N	этому	496	22.174
P | N	ей	449	21.072
P | N	им	433	20.552
P | N	вам	314	17.424
P | N	вашему	296	17.185
P | N	значения	213	14.540
P | N	новый	193	13.749
P | N	мне	166	12.221
P | N	Вашему	137	11.686
P | N	уверенности	136	11.649
P | N	сил	127	11.166
P | N	бы	141	10.590
P | N	вашей	107	10.258
P | N	вашим	91	9.505
P | N	интерьеру	89	9.431
P | N	дополнительный	86	9.252
P | N	Вам	85	8.915

Сравнение коллокаций по словам и леммам.

Коллокации глагола *придать* “по леммам”.

Collocation candidates

Page 1 Go Next >

		Frequency	T-score
P | N	ваш	710	26.260
P | N	он	788	25.240
P | N	этот	600	22.330
P | N	она	459	20.202
P | N	они	463	18.974
P | N	вы	401	18.282
P | N	значение	252	15.691
P | N	новый	277	15.605
P | N	сила	208	14.009
P | N	особый	181	13.260
P | N	уверенность	163	12.724
P | N	дополнительный	167	12.703
P | N	бы	141	10.588
P | N	интерьер	89	9.348
P | N	форма	83	8.494
P | N	я	170	8.484
P | N	свой	128	8.364
P | N	кожа	73	8.286

Варианты определения коллокаций

- Мы рассматривали только непосредственные биграммы, чьи элементы примыкают друг к другу.
- Однако это не всегда оправдано:
  - придать большое значение,
  - придать существенное значение,
  - не придать никакого значения.
- для некоторых коллокаций вероятней “разрывная” биграмма:
  - “взять X назад”, “по Y счёту”
- Можно рассматривать не просто биграммы, а биграммы, находящиеся внутри окна фиксированной ширины (чаще всего берут $w = 5$).
- Тогда формулы для взаимной информации и t-меры можно использовать без изменения.

Сравнение различных коллокаций.

Коллокации глагола *придать* “по словам”, $d = 1$.

Collocation candidates

Page 1 Go Next >

	Frequency	T-score
P | N ему	715	26.522
P | N этому	496	22.174
P | N ей	449	21.072
P | N им	433	20.552
P | N вам	314	17.424
P | N вашему	296	17.185
P | N значения	213	14.540
P | N новый	193	13.749
P | N мне	166	12.221
P | N Вашему	137	11.686
P | N уверенности	136	11.649
P | N сил	127	11.166
P | N бы	141	10.590
P | N вашей	107	10.258
P | N вашим	91	9.505
P | N интерьеру	89	9.431
P | N дополнительный	86	9.252
P | N Вам	85	8.915

Сравнение различных коллокаций.

Коллокации глагола *придать* “по словам”, $d = 2$.

Collocation candidates

Page 1 Go Next >

	Frequency	T-score
P | N ему	723	26.672
P | N значения	564	23.715
P | N этому	550	23.360
P | N ей	457	21.261
P | N им	440	20.722
P | N уверенности	350	18.700
P | N вашему	344	18.529
P | N импульс	313	17.687
P | N вам	316	17.482
P | N сил	306	17.426
P | N форму	256	15.951
P | N новый	243	15.461
P | N силы	161	12.579
P | N Вашему	158	12.552
P | N мне	171	12.423
P | N интерьеру	150	12.245
P | N особый	144	11.980
P | N вид	143	11.816

Синтаксические коллокации

- Часто найденные коллокации не оказываются синтаксически связанными:
- Некоторые коллокации, найденные на материале НКРЯ:

1.	потому что
2.	может быть
3.	у меня
...	...
8.	том что
11.	ничего не
13.	я не

- Вывод: надо рассматривать только пары слов, связанные синтаксической зависимостью.
- Что произойдёт с вероятностной моделью?

Уточнение вероятностной модели

- Наблюдение: почти для любой биграммы $p(w_1 w_2) \gg p(w_1)p(w_2)$.
- Почему это так?
- Элементы биграммы не могут относиться к произвольным синтаксическим и морфологическим категориям.
- Чем может быть x в паре “придать x”:
  - Существительным в винительном падеже (“придать смысл”),
  - Прилагательным в винительном падеже (“придать неожиданный смысл”),
  - Наречием (“придать довольно неожиданный смысл”),
  - Существительным или прилагательным в дательном падеже: “придать фразе довольно двусмысленное звучание”

Уточнение вероятностной модели

- При вычислении ожидаемой вероятности мы предполагали, что все слова допустимы после w_1.
- Но это не так (даже в отсутствие коллокативности).
- Например, за предлогом не может идти глагол.
- Соответственно, в выражении “придать x” вариант $x =$ “значение” “конкурирует” в основном с существительными и прилагательными в винительном падеже.
- Если мы рассматриваем только биграммы, соединённые синтаксической связью, то оно конкурирует только с прямыми дополнениями при глаголе “придать”.
- Как это учесть в модели?

Уточнение вероятностной модели

- Нужно считать только вхождения w_1, w_2, находящиеся в биграммах заданного типа (например, “предлог + существительное”).
- Формально, пусть c_1, c_2 — категории в биграмме $w_1 w_2$.
- Тогда формула взаимной информации примет вид:

$$\begin{aligned} MI(w_1 w_2|c_1 c_2) &= \log c(w_1 w_2|c_1 c_2) - \log c(w_1 \_|c_1 c_2) \\ &\quad - \log c(\_ w_2|c_1 c_2), \text{ где} \\ c(w_1 w_2|c_1 c_2) &= \#(w_1 w_2|c(w_1) = c_1,\ c(w_2) = c_2), \\ c(w_1 \_|c_1 c_2) &= \sum_{u_2} \#(w_1 u_2|c(w_1) = c_1,\ c(u_2) = c_2), \\ c(\_ w_2|c_1 c_2) &= \sum_{u_1} \#(u_1 w_2|c(u_1) = c_1,\ c(w_2) = c_2), \end{aligned}$$

- Аналогично, в формулах для t-меры и в таблицах сопряжённости в качестве суммарной длины надо брать

$$c(\_\_|c_1 c_2) = \sum_{u_1, u_2} \#(u_1 u_2|c(u_1) = c_1,\ c(u_2) = c_2)$$

Таблицы сопряжённости: напоминание

	w_2	$\neg w_2$	
w_1	n_{11}	n_{12}	L_1
$\neg w_1$	n_{21}	n_{22}	L_2
	R_1	R_2	N

$$n_{11} = c(w_1, w_2) \qquad n_{12} = \sum_{\substack{w \in T_2, \\ w \neq w_2}} c(w_1, w)$$

$$n_{21} = \sum_{\substack{w \in T_1, \\ w \neq w_1}} c(w, w_2) \qquad n_{22} = \sum_{\substack{u \in T_1, v \in T_2 \\ u \neq w_1, v \neq w_2}} c(u, v)$$

Здесь T_1, T_2 — множества слов, откуда берутся w_1, w_2.
Например, T_1 — переходные глаголы, T_2 — прямые дополнения.