Интеллектуальные информационные системы

Тема 5

Выделение устойчивых словосочетаний

Павел Исаакович Браславский
pb@imach.uran.ru
весенний семестр 2006

План

- Что такое «устойчивое словосочетание»?
- Зачем их выделять?
- Методы
  - Частота + морфологические шаблоны
  - t-тест
  - статистика χ^2
  - отношение правдоподобия

«Природа» collocations

- Ограниченная (избирательная) сочетаемость слов:
  - фразеологизмы
  - идиомы
  - имена собственные и торговые марки

Collocations – примеры

- играть роль, иметь значение, оказывать влияние, производить впечатление
- средства массовой …, оружие массового …, высшее учебное …
- глубокий старец ↔ поверхностный/мелкий юноша
- крепкий чай ↔ сильный чай
- Кока-Кола, «Русский Стандарт», Microsoft Windows
- Нижний Новгород, Сергиев Посад, Комсомольск-на-Амуре, Михаил Горбачев, Борис Ельцин

Устойчивые СС – зачем?

- Корректная токенизация
- Составление словарей (лексикография)
- Машинный перевод, обучение иностранным языкам (крепкий чай – strong tee ↔ крепкий сон – sound sleep)
- Извлечение терминологии

Частота

Прямой подсчет частоты встречаемости пар (троек):

В литературе описаны несколько подходов к автоматическому выделению устойчивых словосочетаний.

→

в литературе; литературе описаны; описаны несколько; несколько подходов; подходов к; к автоматическому; автоматическому выделению; выделению устойчивых; устойчивых словосочетаний

“мусор” из-за высокой частоты служебных слов

Частота + шаблоны

- Учет морфологии:
  - A N: *турецкий гамбит, первая производная, вопиющая ложь*
  - N N_G: *роза ветров, ошибка резидента, баланс интересов*
  - N Pr N: *вор в законе, торба на круче, хлеб с маслом, песня о буревестнике*

Проверка статистических гипотез

- Статистическая модель
- H_0: Слова «встретились» случайно
- $P(w^1w^2)=P(w^1)P(w^2)$
- Учет не только частоты пар, но и частоты встречаемости отдельных слов (составляющих пару)
- Не совсем корректно для языка, но позволяет получить результаты на практике

Содержание базируется на Manning&Schuetze, 1999

t-тест

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

$\bar{x}$ – эмпирическое среднее

μ – теоретическое среднее

s^2 – эмпирическая дисперсия

N – размер эмпирической выборки

Пример

P(new) = 15828/14307668
P(companies) = 4675/14307668
H_0: *P (new companies) = P(new)P(companies)* ≈3,615·10^{-7}
Схема Бернулли: $s^2 = p(1-p) \approx p$
$\bar{x}$ = 8/14307668
$t \approx 0{,}999932$
Критическое значение для уровня значимости α=0,005 – 2,576 ➔ не можем отвергнуть нулевую гипотезу

χ^2 Пирсона

- Применяется к таблицам 2x2
- Не предполагает нормальности

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

Пример

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$\chi^2 \approx 1{,}55$

Критическое значение для уровня значимости $\alpha=0{,}05$ (степень свободы = 1 для таблиц 2x2)

$\chi^2 = 3{,}841$ → не можем отклонить гипотезу

Отношение правдоподобия

- H_1: $P(w^2|w^1)=p=P(w^2|\neg w^1)$
- H_2: $P(w^2|w^1)=p_1 \neq p_2=P(w^2|\neg w^1)$ $(p_1 >> p_2)$

$$p = \frac{c_2}{N}; \quad p_1 = \frac{c_{12}}{c_1}; \quad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

Биномиальное распределение

$$b(m, n, p) = C_m^n p^m (1-p)^{n-m}$$

Отношение правдоподобия – 2

$L(H_1) = b(c_{12}, c_1, p)b(c_2-c_{12}, N-c_1, p)$

$L(H_2) = b(c_{12}, c_1, p_1)b(c_2-c_{12}, N-c_1, p_2)$

$\log \lambda = L(H_1)/ L(H_2)$

$-2 \log \lambda$ в асимптотике распределено как χ^2

Итог

- Статистические методы позволяют учесть встречаемость отдельных слов
- Тонкости связаны с применимостью методов для разных объемов данных и диапазонов вероятностей (χ^2 лучше чем t-test для больших p, где нормальность нарушается; отношение правдоподобия лучше аппроксимируется χ^2 чем таблицы 2x2 для малых объемов)
- Чаще используется не для принятия/ отклонения гипотез, а для ранжирования словосочетаний-кандидатов

Пример – извлечение терминов

- Две книги из разных предметных областей
  - Олифер Н.А., Олифер В.Г. Сетевые операционные системы. СПб.: Питер, 2005.
  - Щедровицкий Г.П. Философия. Наука. Методология. М.: ШКП, 1989.
- Шаблоны:
  [Прил. + Сущ.], [Прич. + Сущ.],
  [Сущ. + Сущ., Род.п.], [Сущ. + Сущ., Твор.п.],
  [Сущ. + '-' + Сущ.]
- 4 метода

Top-10 (COC)

freq, t-тест	**LR**	**χ^2**
• операционная система •файловая система •адресное пространство •ввод-вывод •оперативная память •рабочая станция •системный вызов •база данных •право доступа •программное обеспечение	• операционная система •файловая система •адресное пространство •ввод-вывод •рабочая станция •оперативная память •база данных •системный вызов •критическая секция •программное обеспечение	• Карнеги Меллон •ввод-вывод •накладные расходы •грамматический разбор •оранжевая книга •доска объявлений •адресное пространство •рабочая станция •Денис Ритчи •критическая секция

Тоз-10 (ГПЩ)

freq	t-тест	LR	χ^2
• процесс мышления •процесс мысли •знаковая форма •суть дела •научное мышление •картина мира •математическое отношение •научный предмет •методологическая работа •целый ряд	• процесс мышление •процесс мысли •знаковая форма •суть дело •картина мира •математическое отношение •научное мышление •научный предмет •методологическая работа •целый ряд	• процесс мышление •суть дело •знаковая форма •сия пора •картина мира •математическое отношение •целый ряд •процесс мысли •онтологическая картина •единая картина	• филиал ВНИИТЭ •Миклухо-Маклай •родимое пятно •Павлик Морозов •категорический императив •экологическая ниша •древние греки •бочка портвейна •конная армия •уральский филиал

Различия методов (по Top-100)

COC

	freq	t-тест	χ^2	LR
freq	1	0,93	0,25	0,73
t-тест	0,93	1	0,26	0,77
χ^2	0,25	0,26	1	0,39
LR	0,73	0,77	0,39	1

ГПЩ

	freq	t-тест	χ^2	LR
freq	1	0,94	0,17	0,71
t-тест	0,94	1	0,19	0,75
χ^2	0,17	0,19	1	0,26
LR	0,71	0,75	0,26	1