Лаборатория естественного языка и обработки текста,
Центр Компьютерных Исследований (CIC),
Национальный Политехнический Институт (IPN),
г. Мехико, Мексика

Автоматический Поиск и Классификация Однословных Терминов в Корпусе Предметной Области с использованием Логарифмической Меры Сходства с Неспециализированным Корпусом

Гельбух А.Ф. (www.gelbukh.com),
Сидоров Г.О. (www.cic.ipn.m/~sidorov),
Лавин-Вийа Э., Чанона-Эрнандес Л.
(Представляет: Большакова Е.И.)

Цели

• Извлечение из текстов однословных терминов (или слов, которые являются частью многословных терминов).

• Предварительная оценка результатов.

• Автоматическая классификация терминов по «сходству».

Схема обработки

Входные данные = два корпуса

- специализированный (с терминами)
- неспециализированный (для сравнения).

Схема обработки:

- предобработка и подготовка данных,
- поиск терминов,
- вычисление меры их сходства и
- объединение терминов в классы

Предобработка и подготовка данных

- Лемматизация
- Фильтрация служебных слов
- Векторная модель данных
- Обработка обоих корпусов

Поиск терминов (1)

- Идея: сравнение взвешенных частот слов в двух корпусах, и если какое-либо слово гораздо чаще присутствует в корпусе предметной области, то это вероятный термин.

$$G = 2 * \left(\left(fr_{domain} * \log\left(\frac{fr_{domain}}{frExpected_{domain}} \right) \right) + \left(fr_{general} * \log\left(\frac{fr_{general}}{frExpected_{general}} \right) \right) \right)$$

Поиск терминов (2): ожидаемые частоты

$$R\_fr = \frac{fr_{domain} + fr_{general}}{size_{domain} + size_{general}}$$

где $size_{domain}$ и $size_{general}$ размеры соответствующих корпусов, вычисленные в количестве слов.

$$frExpected_{domain} = size_{domain} * R\_fr$$

$$frExpected_{general} = size_{general} * R\_fr$$

Поиск терминов (3)

- Доп. шаг: к какому корпусу относится предполагаемый термин? Важно только из корпуса преметной области.
- Если относительная частота в корпусе предметной области больше, чем в в неспециализированном корпусе.

Поиск терминов (4): Пример

Слово	fr_{domain}	$fr_{general}$	$frExpected_{domain}$	$frExpected_{general}$	G
socket	1	0	0.010286744	0.989713252	9.153798
sofisticado (сложный)	5	169	1.789893508	172.2101135	3.912798
soft	1	12	0.13372767	12.86627197	2.351035
software	430	831	12.97158432	1248.028442	2334.961
sofware[1]	2	2	0.041146975	3.958853006	12.8037
sol (солнце)	2	933	9.618105888	925.381897	-9.016687
solamente (только)	20	1714	17.83721352	1716.162842	0.254846

Эксперимент

• Специализированный корпус:
Статьи по информатике из Википедии
(26 страниц, содержащих 44,495 слов).

• Неспециализированный корпус:
выпуски газеты Excélsior (Мексика) конца 90х годов, всего 1,365,991 слов.

• Эмпирический порог likelihood для выделения терминов (=270).

Эксперимент: результаты

Термин	Log-likelihood
Dato (данные)	1506
Computador (компьютер)	863
Circuito (плата)	467
Memoria (память)	384
Señal (сигнал)	372
Secuencia (последовательность)	353
Computación (вычисление)	351
Información (информация)	346
Dispositivo (устройство)	342
Algoritmo (алгоритм)	341
Electrónico (электронный)	322
Base (основа)	319

Эксперимент: классификация

алгоритм, for, *реализация*, массив, реализовать, дерево (поиска)
аналоговый, напряжение, бинарный
as, if, int, integer (число), псевдокод, return (возврат), vtemp, *схема*, *описание*, Тьюринг, end (конец)
B2B, *бизнес*, хостинг, клиент, сервер, Интернет, ~~то~~, электронный, состоять
биология, биоинформатика, ДНК, выравнивание, ClustalW, фаг, ген, геном, геномы, геном, геномика, генный, гомология, *человек*, микрочип, *моделирование*, нуклеотиды, *прогнозирование*, белок, белок-белок, Sanger последовательность, эволюционный, *последовательность*, *биологический*, вычислительный, протокол, *отбор*, *анализ*, *технологический*, *структура*, *взаимодействие*, дополнить, *монтаж*, *инструмент*, ~~часто~~, ~~использовать~~, ~~рубить~~, программное обеспечение, визуализировать, *количественная оценка*, *модель*, автоматизировать, поиск
компонента, транзистор, трубка, функционировать, связь, устройство, ~~и т.д.~~, *технология*, цифровые, микропроцессоры, *скорость*, *логика*, случаться, динамик
компьютер, *наука*, *стабильный*, *научный*, вычисление, *дисциплина*, *математика*, ~~как правило~~, *теория*, вычислительные, *инженерный*, исследовать, искусственный, *математический*, информатика, параллельный, программирование

Эксперимент: оценка

Точность выделения терминов:

• Всего терминов 270, из них 31 глагол (подчеркнуты), то есть остается 239 терминов. Из этих терминов 19 явно не являются терминами данной области (зачеркнуты). точность = 92,5 %.

• Если мы добавим к словам, которые будем считать неправильно определенными, 48 общенаучных терминов (курсив), то получим точность в (239-(19+48))/239 = 72%

Будущие направления исследований

• Определить возможность автоматического определения порога при отборе терминов.

• Попробовать разные параметры алгоритма классификации *k-means*. Оценить возможность применения алгоритма, позволяющего определять количество классов автоматически.

• Вместо меры сходства по косинусу угла попробовать другие меры сходства при классификации.

• Сравнить разные логарфимические меры сходства при поиске терминов.

• Выполнить сравнение с одним или несколькими корпусами разных предметных областей, чтобы отфильтровать общенаучные термины.

Спасибо за внимание!

www.gelbukh.com

www.g-sidorov.org