Устойчивые словосочетания, термины и их извлечение из текстов

В прошлый раз

- Синтаксический анализ
- Полный синтаксический анализ
  - Сложно
  - Нужно много знаний, в т.ч. и знаний о мире
- Полный vs. частичный синтаксический анализ
- Частичный синтаксический анализ
  - Извлечение устойчивых словосочетаний, терминов
  - Автоматическое присваивание ключевых слов, тегов
  - Извлечение синтаксических контекстов и др.

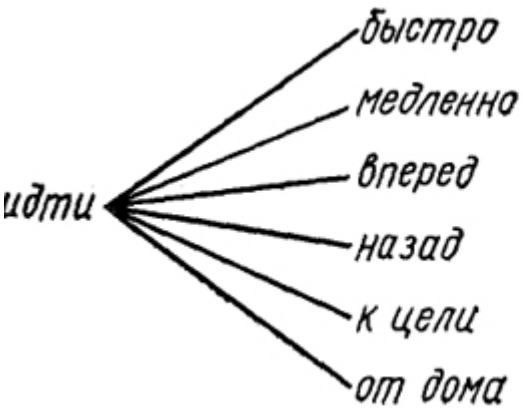
Словосочетания vs. N-граммы

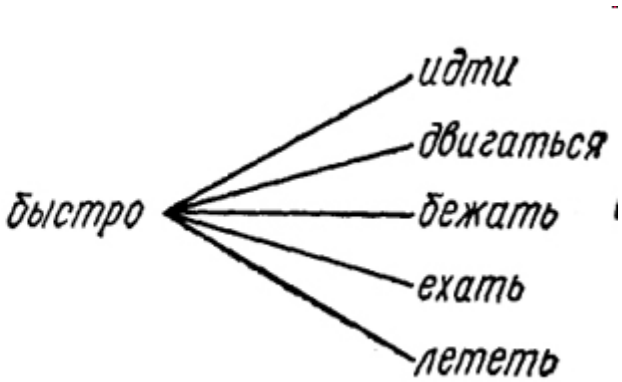
Словосочетания по главному слову

- Именные группы (зеленый сад)
- Предложные группы (в доме)
- Глагольные группы (участвовать в игре)
- Группа прилагательного (приятный на вид)
- Причастный оборот (соответствующий образцу)
- Деепричастный оборот (имея в кармане)

Свободные словосочетания

- Свойства
  - Высокая степень комбинаторности,
  - Значение словосочетания строится из значений слов компонентов
  - Подчинение синтаксическим правилам, могут разбиваться другими словами





Факторы устойчивости/неустойчивости

- Факторы
  - семантическая отделимость компонентов
  - выводимость значения фразы из значений компонентов
  - возможность замены каждого слова на синонимы
  - подчиненность синтаксическим правилам (в т.ч. возможность вставки других слов)

- Для устойчивых словосочетаний не выполняются какие-то из этих факторов

Фразеологическое сращение (идиома)

- Семантически неделимый оборот, значение которого совершенно не выводимо из значений составляющих его компонентов
- Одно или оба слова не употребляются вне данной фразы
- Синтаксически не разделимы – жесткое расположение в тексте, т.е. ведут себя практически как слова с пробелами
- Обычно перечисляются в фразеологических словарях – их относительно немного
  - *содом и гоморра, бить баклуши*, *спустя рукава*
  - *от мала до велика, ничтоже сумняшеся*

Фразеологические единства

- устойчивый оборот, в котором сохраняются признаки семантической раздельности компонентов.
- каждое слово имеет свое значение, но в совокупности они приобретают переносный смысл.
- слова плохо поддаются замене на синонимы
- допустимы вставки слов
  - *гранит науки, зайти в тупик*, *бить ключом*, *плыть по течению*, *держать камень за пазухой*, *водить за нос*

Фразеологические сочетания (коллокации)

- устойчивый оборот, в состав которого входят слова как со свободным значением, так и с фразеологически связанным, несвободным (употребляемым лишь в данном сочетании).
- целостное значение следует из значений составляющих их отдельных слов.
- состав допускает ограниченную синонимическую подстановку - один из членов фразеологического сочетания оказывается постоянным

- *сгорать от любви, ненависти, стыда, нетерпения*

Еще примеры коллокаций

- Одно из слов встречается только (или почти) только в данном выражении
  - *насупить брови, прищурить глаза, грецкий орех, окладистая борода*

- Одно из слов в узко специальном значении
  - *глухая стена, мертвая петля, бросить взгляд, внести ясность*

Почему важно извлечение устойчивых словосочетаний

- Извлечение и поиск информации
  - Слово, попавшее в оборот, меняет свое значение
- Синтаксический анализ
  - Могут быть нарушены правила грамматики
- Понятийный анализ предметной области
  - многословные термины
- и др.

Автоматическое извлечение устойчивых словосочетаний

- Синтаксически правильная группа
  - Чаще всего
- Факторы для учета
  - Частотность словосочетания
  - Статистические меры взаимной ассоциации – разные виды учета
  - Заменяемость компонентов на синонимы
    - Forest fire – wood fire
  - Сравнение схожести контекстов употребления словосочетания и отдельных слов
  - Извлечение разрывных словосочетаний – фактор отклонения от среднего расстояния между элементами словосочетания

Извлечение коллокаций: лингвистические критерии

- Лингвистические критерии – *синтаксические образцы* сочетаний по их типам, например:
- $A \leftarrow N$ *полевая форма*
- $V \rightarrow N$ *заметить разницу*
- $N \rightarrow Prep \rightarrow N$ *хлеб с маслом*
- $V \rightarrow и \rightarrow V$ *грабить и убивать* и др.

Статистические критерии: ассоциативные меры

- Меры ассоциации (совместной встречаемости) учитывают не только частоту сочетания, но и частоту входящих в него слов
- Извлечение двухсловных неразрывных коллокаций
- Применяемые меры (*association measures*):
  - *Mutual Information: MI* и *MI3*
  - *t-score*
  - *Log-likelihood*
  - *Dice* (для n-словных сочетаний)
  - ……

- Упорядочивают (ранжируют) извлеченные коллокации.

Мера Mutual Information (MI)

- ***N*** – размер корпуса в словах или словоформах;
- ***f*** – *frequency,* частота совместной встречаемости пары слов ***a, b*** или абсолютная частота отдельного слова ***a*** или ***b*** соответственно;
- Из теории вероятностей: ***I*** – взаимная информация, ***P*** – вероятности слов и их сочетаний (если слова независимы, мера равна 0, если связаны, то больше 0), т.о., ***MI*** оценивает степень независимости появления двух слов в корпусе.
- ***MI*** > 1, то словосочетание статистически значимо

$$MI = \log_2 \frac{f(a,b) \times N}{f(a) \times f(b)}$$

$$I(a,b) = \log_2 \frac{P(a,b)}{P(a) \times P(b)}$$

Модификации *MI*

- *MI* >1 означает обычное, что слова употребляются вместе чаще, чем по отдельности
- *MI* можно обобщить для любого числа слов в сочетании.
- Возможны модификации, усиливающие влияние отдельных компонент формулы.
- Например, кубическая взаимная информация:

$$MI_3(a,b) = \log\left(\frac{N \cdot f^3(a,b)}{f(a) \cdot f(b)}\right)$$

Мера *t-score*

$$t - score = \frac{f(a,b) - \frac{f(a) \times f(b)}{N}}{\sqrt{f(a,b)}}$$

где

N – размер корпуса в словах или словоформах;

f – *frequency*, частота совместной встречаемости пары слов ***a, b*** или абсолютная частота отдельного слова ***a*** или ***b*** соответственно.

Мера показывает, насколько неслучайной является сила ассоциации (связанности) компонент коллокации.

Мера *Log-likelihood*

$$\log - likelihood = 2\sum f(a,b) \times \log_2 \frac{f(a,b) \times N}{f(a) \times f(b)}$$

где

N – размер корпуса в словах или словоформах;

f – *frequency*, частота совместной встречаемости пары слов ***a, b*** или абсолютная частота отдельного слова ***a*** или ***b*** соответственно.

Мера выражает отношение функций правдоподобия, соответствующим двум гипотезам – о случайной и неслучайной природе двухсловного сочетания.

Меры ассоциации: пример

/Данные М.В. Хохловой/

Collocation	***MI-score***	***LL-score***	***T-score***
искренне говоря	2,94/***4.92***	4,49/***6.11***	2,74/***2.16***
точно говоря	2,64/***5.29***	21,09/***55.31***	2,21/***6.24***
просто говоря	2,19/***5.60***	79,38/***209.98***	2,02/***11.75***
откровенно говоря	6,12/***9.67***	230,24/***299.54***	2,09/***10.19***
честно говоря	7,08/***10.98***	1064,06/***1690.55***	1,96/***22.33***
объективно говоря	4,24/***6.82***	4,37/***11.22***	4,16/***2.43***
образно говоря	3,00/***10.80***	102,07/***145.01***	2,32/***6.63***
строго говоря	4,55/***8.34***	184,16/***351.80***	2,08/***12.05***

Первое числовое значение дано для леммы,
второе значение *(курсивом)* - для формы деепричастия.

Особенности мер ассоциации

- *Ранги* (порядковые номера)
  извлеченных коллокаций для разных мер не совпадают.
- Разные результаты извлечения при использовании мер для словоформ и для лемм (нормализованный текст).
- Зависимость результатов от объема и типа корпуса (например, 6 млн. или же 200 тыс. слов)

  для текстов разных жанров – разные меры?
- Высокоранговые коллокации часто входят в словари устойчивых словосочетаний (коллокаций).
- Общая проблема: разрывные коллокации.

Сравнение мер ассоциации

- ***Mutual Information**:*
  - Завышает значимость редких словосочетаний, делая возможным их выявление, но при этом выявляются и случайные сочетания (опечатки);
  - Требует порог отсечения по частоте снизу, подбираемый экспериментально;
  - Обычно подбирается и пороговое значение сверху.
- ***t-score***:
  - Не требует порога отсечения снизу по частоте;
  - Завышает значимость сочетаний с высокочастотными словами, и в результате извлекаются: сложные предлоги, предложные группы, обстоятельства, числа (для исключения требуется заранее составлять списки стоп-слов).

Термины и их автоматическое извлечение из текстов

Определение термина

- Слово (или сочетание слов), являющееся точным обозначением определенного понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п.
  (*БТС РЯ, Лингвистический словарь*)

- Понятие – мысль, отражающая в обобщенной форме предметы и явления действительности посредством фиксации их свойств и отношений; последние (свойства и отношения) выступают в понятии как общие и специфические признаки, соотнесенные с классами предметов и явлений **(***Лингвистический словарь***)**

Словосочетания из математической области

- Минимальное количество
- Дифференциал высшего порядка
- Формула площади
- Неравенство Бесселя
- Метод трапеций
- Уравнение окружности
- Разностный метод
- Формальный параметр
- Погрешность решения
- Идея метода
- Ненулевое решение
- Способ построения
- Свойство функции

Где здесь термины и нетермины?

Терминоведение

- наука, изучающая специальную лексику с точки зрения её типологии, происхождения, формы, содержания (значения) и функционирования,
- а также использования, упорядочения и создания.

Раньше считалось:
Свойства идеального термина

- **Sager, J.C.: A Practical Course in Terminology Processing**
- the term must relate directly to the concept. It must express the concept clearly,
- there should be no synonyms where absolute, relative or apparent,
- the contents of terms should be precise and not overlap in meaning with other terms,
- the meaning of the term should be independent of context
- Такие свойства хороши при составлении терминологических словарей, но…

Раньше считалось: Термины и определения

- Наличие определения является для многих исследователей обязательным, конституирующим признаком термина.
- На практике критерием разграничения термина и нетермина часто служит наличие
  или, соответственно, отсутствие дефиниции.
- Некоторые исследователи полагают, что:
  - Термин – это слово или словосочетание, имеющее дефиницию или требующее

Теория терминологии

- Основное назначение теории:
- создание описания понятийной системы предметной области, создав терминологическую систему однозначно понимаемых терминов
- отражение этой системы в терминологических словарях (поэтому так важна дефиниция – зачем иначе включать в словарь)

- но нас интересует другой аспект рассмотрения: какие выражения содержатся в текстах предметной области и какие из них мы будем считать терминами

Текущий взгляд: свойства терминов

- Термин – обозначение понятия предметной области
- Термины не всегда точны
- Термины не могут избавиться от своей языковой формы
  - имеют синонимы
  - могут быть многозначны
- Границы между общей лексикой и терминологией не так жестки

Термины: синонимы и варианты

- ***Кредитование***
  - *кредит, кредитная услуга, кредитное обслуживание, кредитная операция, выделение кредита, выдача кредита, выделение кредитных средств, предоставление кредита*

- ***Линейный оператор***
  - *линейное отображение, линейное преобразование*

Многозначность терминов

- Тригонометрические функции
  - Косинус=функция косинус
  - Синус=функция синус
- Отношения сторон в прямоугольном треугольнике
  - Косинус=косинус угла
  - Синус= синус угла
- Тип многозначности – метонимия (перенос слова на смежное явление):
  - смазка (процесс)
  - смазка (вещество)
- Среди терминов, конечно, меньше многозначных слов и выражений

Принципы для распознавания терминов

Экспертная практика извлечения терминов зависит от типа создаваемых ресурсов

- Терминологические ресурсы и особенности включения в них терминов
  - Словари для людей
  - Информационно-поисковые тезаурусы
  - Терминологические ресурсы для автоматической обработки текстов
- Модели извлечения терминов
  - Признаки для извлечения терминов
  - Комбинирование признаков

Терминологические словари для людей

- Иногда есть проблемы с границами ПО
  - Государственный финансовый контроль
  - Онтологическое моделирование
- Основной принцип отбора терминов – необходимость дефиниции
  - Терминологизация (или уточнение) известного выражения
  - Неизвестное слово (выражение)
- Типовой размер: от нескольких сотен до 2-3 тысяч

Информационно-поисковые тезаурусы

- Информационно-поисковый тезаурус – нормативный словарь терминов предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области
- Национальные и международные стандарты
- Используются в ряде международных организаций и парламентский организаций
  - Европейский парламент – EUROVOC
    ООН – UNBIS Thesaurus
- Повышение интереса в последнее время
  - За счет роста интереса к предметно-ориентированному и корпоративному поиску

Включение терминов в ИПТ на основе многословных выражений (Амер. стандарт Z39.19)

- Признанность в литературе
- Расщепление термина увеличивает многозначность: *plant food*
- Смысл выражения зависит от порядка слов: *информационная наука - научная информация*
- Одно из слов-компонент находится вне сферы тезауруса или слишком общее: *first aid*
- Отношения дескриптора не следуют из его структуры:
  - *Искусственные почки, статус беженца, traffic lights*

Шемакин: Научно-технический тезаурус

- Значение одного из терминов изменилось бы в результате комбинации
  - *посадочные площадки*;
- термин-словосочетание обозначает некоторую физическую целостность или специфическое вещество
  - *цифровые вычислительные машины, перекись водорода*;
- термин-словосочетание имеет один или несколько синонимов на уровне словосочетания
  - *полупроводниковые триоды - транзисторы*;
- термин-словосочетание употребляется только в единственном или множественном числе
  - *автоматический перевод, английский язык, строительные материалы*

Научно-технический тезаурус-2

- для термина словосочетания существует общепринятая аббревиатура, составленная из первых букв компонентов словосочетания
  - *электронно-цифровые вычислительные машины - ЭЦВМ*;
- для некоторых элементов термина-словосочетания мала вероятность использования вне данного словосочетания
  - *обзор веерным лучом, этажерочные микромодули*
- один из элементов термина-словосочетания снимает неоднозначность другого:
  - *автоматы: автоматы дозирования, автоматы курса*;
- словосочетания являются единственным способом уменьшения информационного шума,
  - *преобразователи последовательного кода в параллельный*
  - *преобразователи параллельного кода в последовательный*

Выводы по включению терминов в ИПТ

- Стандартный размер – несколько тысяч терминов
- Много разных принципов отбора (включая многозначность составных частей, синонимы, и др.)
- Устойчивость совместного употребления (как в коллокациях) не самый важный фактор
- (!) Отнесенность вводимого термина к определенному типу сущностей (связь с понятием ПО) сохраняется

Методы извлечения терминов

Традиционное извлечение терминов

- Отдельные слова и синтаксически правильные группы существительного
- Могут быть лексические ограничения
  - словарь слов, которые редко встречаются в терминах:
    - Оценочная лексика и др.
- Статистический критерий
  - Для терминов-словосочетаний – частотность, взаимная информация (MI), C-value (учет более длинных словосочетания) и др.
  - Для отдельных слов tf.idf, модифицированные виды C-value и др.

Основные шаги процесса извлечения терминов

Извлечение кандидатов из текстовой коллекции

↓

Переупорядочивание списка выбранных кандидатов для получения большего числа терминов в начале списка

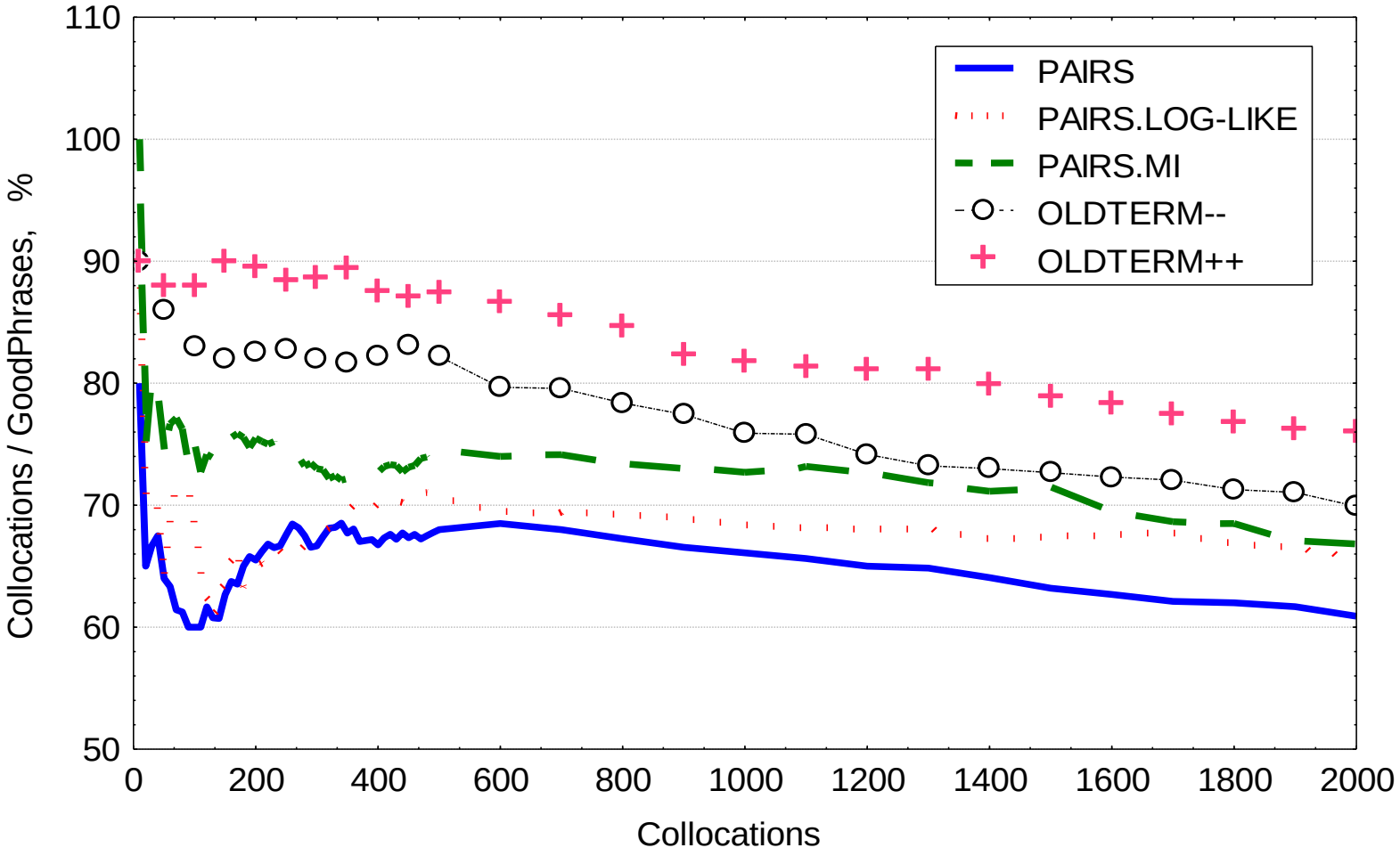
↓

Комбинирование признаков (применение методов машинного обучения) для улучшения итогового результата

Проблемы извлечения терминов

- На первых 100-200 словосочетаниях упорядоченного списка многие методы работают очень хорошо
- Далее качество снижается
- Эксперту
  - Нужно много просматривать
  - Много промежуточных случаев
- Но термины встречаются далеко от начала списка
- Нужны содержательные критерии

Оценка качества сборки двухсловных словосочетаний



Автоматическое извлечение терминов: факторы

- Частотные характеристики в коллекции предметной области
- Использование контрастных коллекций (общелитературных коллекций)
- Контекстные меры – сравнение частотности термина-кандидата с частотностями окружения
- Неравномерная распределенность по коллекции
  - В отдельном тексте
  - В кластере близких по смыслу текстов
- Лингвистические признаки:
  - написание с большой буквы, употребление как подлежащего
- Термины-словосочетания: ассоциативные меры

Частотные признаки слов, вычисляемые только на базовой коллекции

Частотность – кол-во словоупотреблений Tf

Документная частотность – кол-во документов (df)

Tf.Idf $$TF_t(w) \times \log \frac{|D_t|}{DF_t(w)}$$

TF-RIDF (Пуассоновский процесс) $$TF_t(w) \times (\log \frac{|D_t|}{DF_t(w)} + \log(1 - e^{-\frac{TF_t(w)}{|D_t|}}))$$

Domain Consensus (подобно энтропии) $$-\sum_{d \in D}(freq(w, d_k) \times \log(freq(w, d_k))$$

Модификации TF-IDF, использующие тестовую и контрастную коллекции

- *Contrastive Weight*
  - *Идея*: слова из общей лексики распределены одинаково в обеих коллекциях

$$\log(TF_t) * \log\left(\frac{|W_t| + |W_r|}{TF_t + TF_r}\right)$$

- где TF_t и TF_r - частотности слова в тестовой и контрастной коллекциях, $|W_t|$ и $|W_r|$ - число слов в тестовой и контрастной коллекциях
- *KF-IDF*
  - Отражает новизну слова в тестовой коллекции
  - $DF * \log\left(\frac{2}{|D|_w} + 1\right)$ $|D|_w = \begin{cases} 1, \text{ если слова нет в контрастной} \\ 2, \text{ если слово есть в контрастной} \end{cases}$

46

Признаки, использующие статистическую и контекстную информацию: MC-/MNC-value

- *Основная идея*: объединение частотности слов-кандидатов и информации о контекстных словах
- *MC-value*
  - Ищет термины — части объемлющих словосочетаний
  - $TF - \frac{\sum_{p \in P} TF(p)}{|P|}$, где P — множество словосочетаний, содержащих данное слово
- *MNC-value*
  - Добавляет контекстную информацию в MC-value

$$0.8 * MC - value + 0.2 * \sum_{c \in C} freq(c)$$

  - где $\sum_{c \in C} freq(c)$ - контекстный фактор

Лингвистические признаки

- Берем только прилагательные и сущ-ные
- Подмножества:
  - Существительные, встречающиеся в тексте в именительном падеже
  - Слова-кандидаты, начинающиеся с заглавной буквы
  - Слова с заглавной буквы, не стоящими первыми в предложении текста
- На этих подмножествах можно вычислять различные статистические характеристики

Выводы по признакам извлечения терминов

- Важно, что вышеперечисленные признаки – это лишь некоторые примеры из предложенного в различных статьях
- Каждая группа признаков – отражает какие-то специфичные свойства терминов
- Для всестороннего учета – комбинация принципов

Мера для оценки качества упорядочения

- Средняя точность AvP (адаптирована из информационного поиска) :
  - Пусть в списке k - терминов
  - Точность $PrecTerm_i$ вычисляется в момент поступления очередного правильного термина

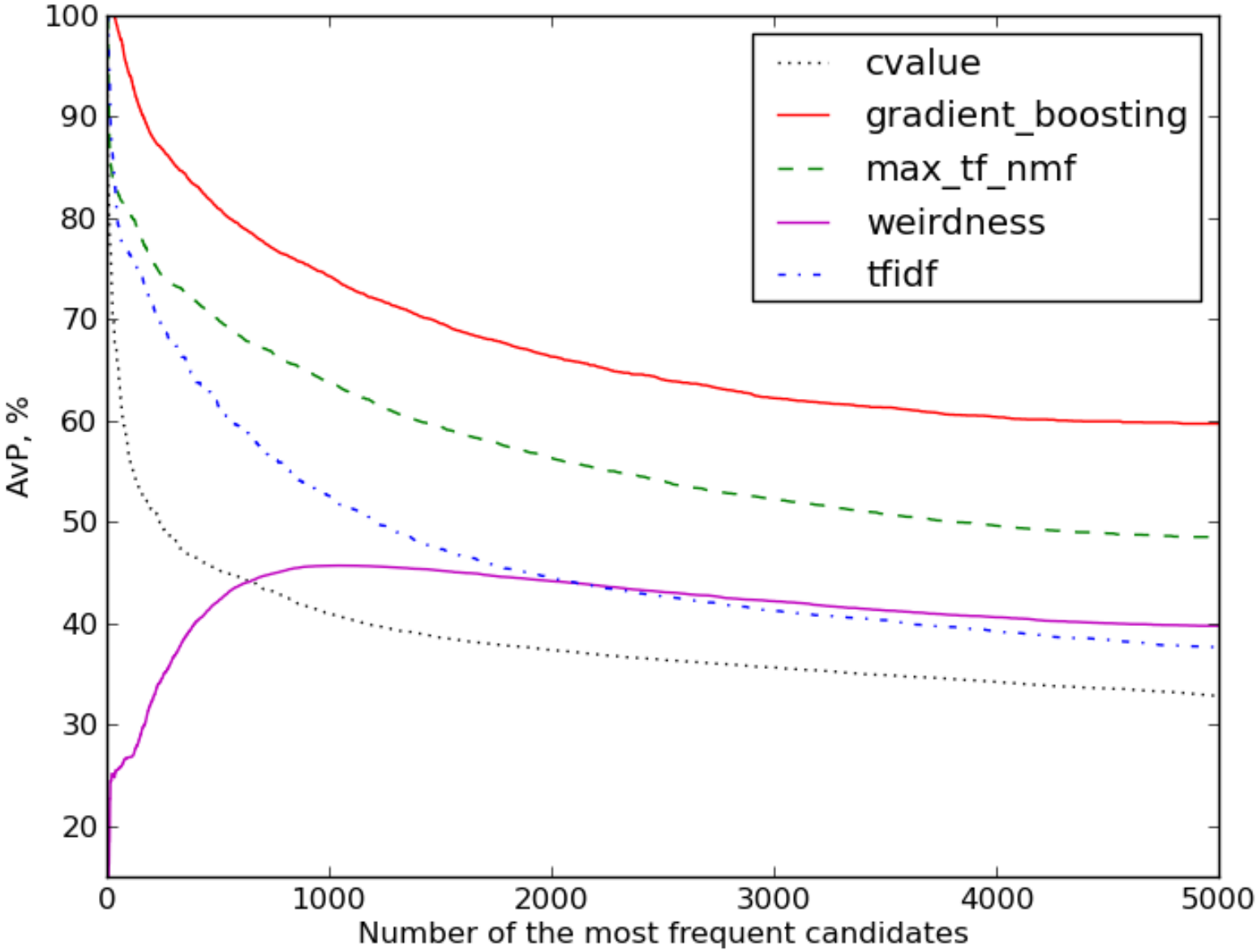
$$AvP = \frac{1}{k} \cdot \sum PrecTerm_i$$

- Example:
- T, N, T AvP = (1/2) (1+2/3) =5/6 = 0.888..
- N, T, T AvP = (1/2) (1/2+2/3) =7/12 = 0.68..

Извлечение слов-терминов предметная область: банки

- Базовая коллекция 10422 документа
  - Журнальные статьи из финансовых журналов
  - Более 15.5 млн. слов
- Контрастная коллекция
  - 1 млн. новостных документов
  - статистика о встречаемости в документах новостей
- Эксперименты с 5000 наиболее частотных слов
- Как оценивать качество: Банковский тезаурус, сделанный по проекту с ЦБ РФ
- Используется более 60 признаков

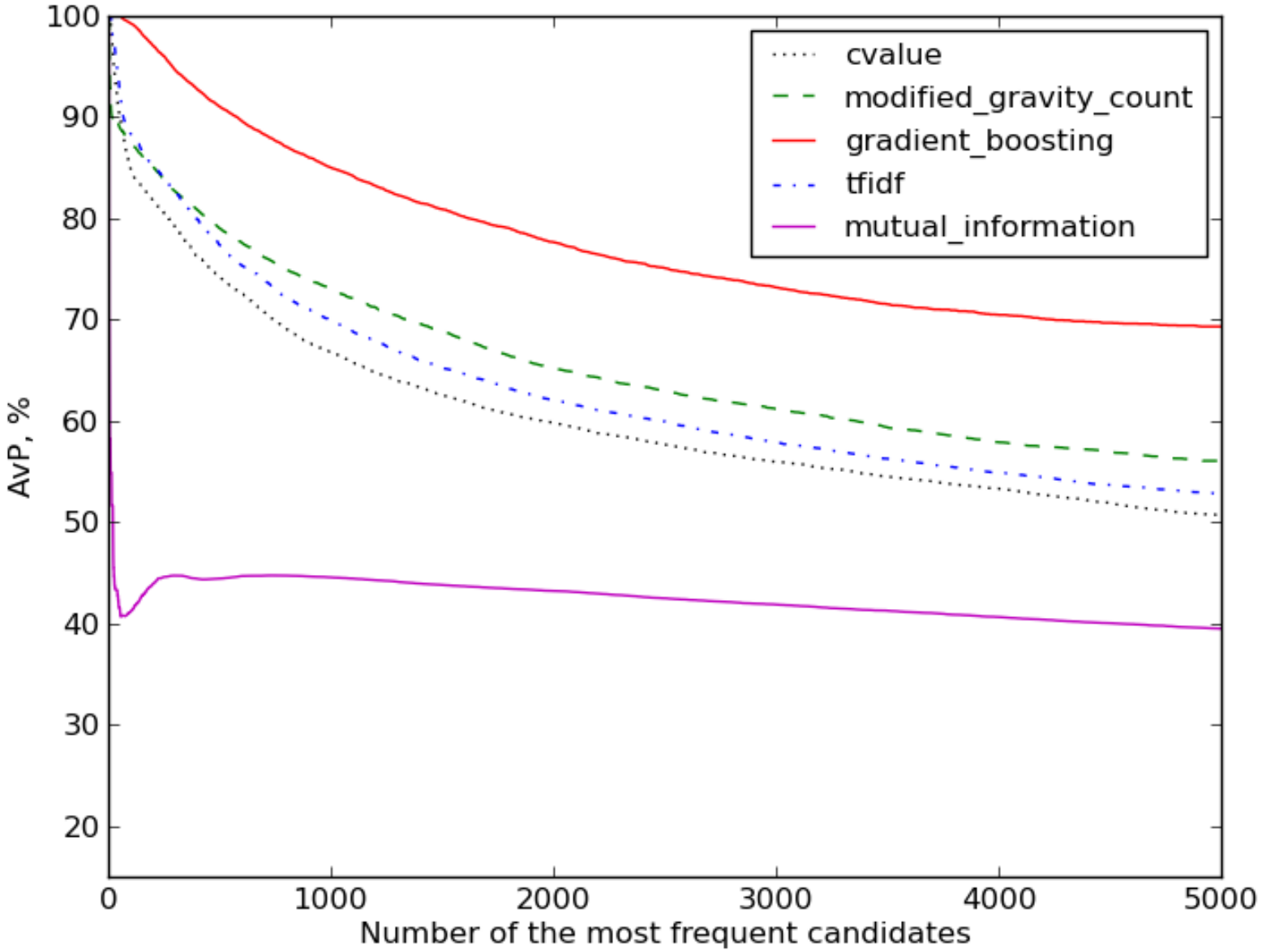
Качество извлечения слов-терминов (gradient boosting – другой метод комбинирования)



Лучшие 10 слов по комбинации признаков

- *Банковский*
- *Банк*
- Год
- *РФ*
- *Кредитный*
- *Налоговый*
- *Кредит*
- *Пенсионный*
- Средство
- Клиент

Качество извлечения двухсловных терминов



Как же решить, является ли выражение термином?

- Предметная область
  - Термин – должен принадлежать предметной области или иметь специализированное значение в этой области
- Термин выделяет значимую сущность или процесс в предметной области, со своими специфическими характеристиками
  - Если у выражения есть определение в заданной предметной области – то это термин (достаточное условие для термина)
- Синонимы или варианты к термину – нужно запоминать

Заключение

- Извлечение терминов – по своей природе – многофакторный процесс, т.е. для качественного извлечения необходимо комбинировать различные признаки, включая тип терминологического ресурса
- Необходимо исследовать устойчивость модели извлечения терминов для разных областей
- В современных технологиях извлечения знаний из текстов часто используются комбинированные модели

Заключение-2

- Вклад признаков в извлечение терминов может зависеть и от особенностей предметной области
  - Широкая или узкая предметная область
  - Насколько близка к общеупотребительному языку
- И от особенностей ресурса
  - Например, есть эксперименты, что ассоциативные меры не слишком важны при извлечении многословных терминов для тезаурусов

Задание

- Взять текст большого закона
- Например, здесь:
- http://base.consultant.ru/cons/cgi/online.cgi?req=home&utm_csource=online&utm_cmedium=button
- из текста закона извлечь конструкции
- прилагательное-существительное
- Упорядочить двумя способами
  - по частотности
  - по мере взаимной информации
- Оценить, при каком методе терминов больше
- среди первых 10, 20 словосочетаний.