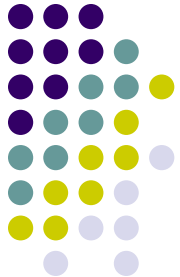
СОЧЕТАЕМОСТЬ СЛОВ: КОЛЛОКАЦИИ И ТЕРМИНЫ В КЛ

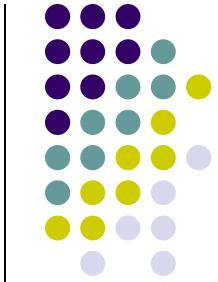
Большакова Елена Игоревна

СОДЕРЖАНИЕ



1. Словосочетания, виды сочетаемости
2. Грамматическая (синтаксическая) сочетаемость
   - валентность и модель управления слова
3. Лексическая сочетаемость в ЕЯ
   - фразеологичность и устойчивость словосочетаний
4. Коллокации и методы их извлечения из текстов
   - меры ассоциации (лексической связности)
   - коллокации в приложениях КЛ
5. Термины и их извлечение из текстов
   - лингв. и статистические признаки терминов
   - машинное обучение, оценка качества извлечения
6. Заключение

СОЧЕТАЕМОСТЬ СЛОВ



- Сочетаемость слов ЕЯ: способность слов соединяться друг с другом, образовывая единицы более высокого уровня
- В языке словосочетания (*phrases*) образуют подуровень для синтаксического уровня
- Сочетаемость учитывается при синтаксическом и семантическом анализе
- Определяется несколькими факторами:
  - Грамматическая (синтаксическая)

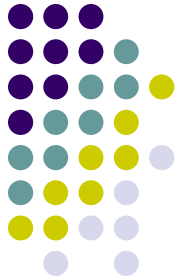
    сильный дождь – ~~сильный дождя~~
  - Семантическая *сильный дождь – ~~сильный шкаф~~*
  - Лексическая (лексико-семантическая)

    сильный дождь – ~~тяжелый дождь~~

ТИПЫ СОЧЕТАЕМОСТИ СЛОВ

- Грамматическая (синтаксическая)
  - зависит от принадлежности слов к частям речи
  - выражает синтаксическую связность слов
  - требует определенной грамматической формы
    A N: яркий свет, N N : песнь птицы, но bird song
- Семантическая предполагает согласованность семантических классов сочетающихся слов (отсутствие противоречий):
  глупый человек, поет птица
  но не: *глупый дом, поет апельсин*
- Лексическая проявляется в избирательности лексем, нестандартная сочетаемость единиц ЕЯ:
  оказать услугу/ внимание, но не *заботу/интерес*

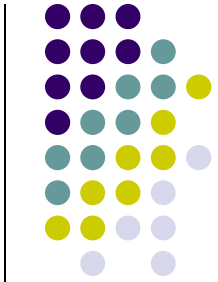
ГРАММАТИЧЕСКАЯ СОЧЕТАЕМОСТЬ: РАЗНОВИДНОСТИ



Три основных вида:

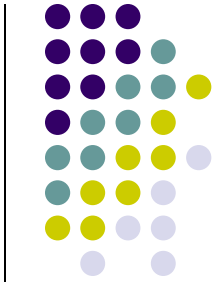
- Согласование *(agreement)* слов языка,
  например: *(нет) большого самолета*
  - грамматическое уподобление морфологических параметров (рода, падежа, числа) у зависимых слов (например, прилагательных); Широко представлено во флективных языках.
- Управление (падежом): *подарить сестре цветок*
  - управляет выбором падежа подчиненного слова
  - зависит от конкретной управляющей лексемы и соотносится с понятиями *валентности* слова и *модели управления*
- Примыкание: не фиксируется грамматически,
  например: *директор Иванов, чай с сахаром*

СИНТАКСИЧЕСКИЕ ТИПЫ СЛОВОСОЧЕТАНИЙ РЯ



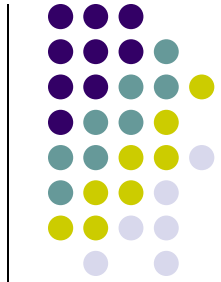
- По составу: двучленные, трехчленные и т.п.:
  яркий свет, передать книгу преподавателю
- По типу синтаксической связи, распространенные в РЯ :
  - Определяемое слово (существительное, глагол, наречие, прилагательное) → его определение (прилагательное, наречие): *актовый зал, вполне приемлемо*
  - Глагол → его дополнение (прямое или предложное) в виде существительного:
    уделить внимание, ворвались в здание
  - Сказуемое → подлежащее:
    спасатели обнаружили, отправлен груз
  - Существительное → его дополнение:
    хлеб с маслом, жертвы теракта
  - Прилагательное или причастие → его дополнение:
    занятый трудом

СИНТАКСИЧЕСКАЯ ВАЛЕНТНОСТЬ



- Валентность – способность слова присоединять другие единицы определенным синтаксическим способом (сопоставимо с понятием предиката в логике)
- *Слова-предикаты* описывают ситуации и действия:
  - глаголы и глагольные формы: *идти, приходящий*
  - существительные: имена действия (*преобразование*), качества (*красота*), параметрические (*высота*), категориальные (*тип*)
  - краткие прилагательные: *рад, должен*
  - предлоги: *к (морю)*
- Слова-предикаты имеют места для заполнения – *валентности*, например:
  - *Подарить*: *кто? (1) что? (2) кому? (3)*
  - *Видеть*: *кто? (1) кого/чего ? (2)*

ВАЛЕНТНОСТИ И МОДЕЛЬ УПРАВЛЕНИЯ

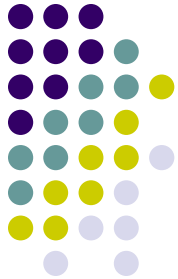


Валентность описывает сочетательную способность слов:
ряд заполняемых позиций-валентностей:

Рубить: *кто? (1) что? (2) чем? (3)*

- *Актант* – заполнитель валентности: слово, фраза, словосочетание (обозначает лицо/существо/предмет)
- Валентности отличаются по степени обязательности
- *Сирконстанты* – необязательные валентности многих слов-предикатов, обычно обстоятельства (время, место, образ действия): *когда? где? как? …*
- *Модель управления* слова-предиката – набор его валентностей + синтаксические ограничения на актанты, (в частности: падеж актантов),
  часто: учет взаимного расположения актантов
- Слово может иметь несколько моделей управления

МОДЕЛЬ УПРАВЛЕНИЯ: ПРИМЕР



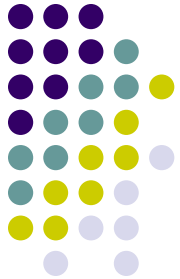
Наказывать: A *наказал* B D *за* C

Директор наказал Иванова рублем за неточный отчет

Модель управления (МУ) – таблица (*падежно-актантная рамка*):

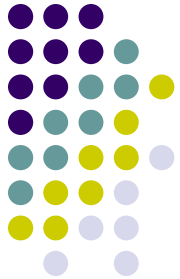
1 = A субъект	2 = B объект	3 = C причина	4 = D средство
Sим (группа существит., имен.падеж)	Sвин (группа существит., вин.падеж)	*за* Sвин (*за* + группа существ., вин.падеж)	Sтв (группа существит., твор.падеж)

СЕМАНТИЧЕСКИЕ ВАЛЕНТНОСТИ



- Валентность имеет
  синтаксический и семантический аспект,
  семантический первичен!
- Fillmore C. (60-е гг.) ввел т.н. *семантические падежи* (=валентности), и описал их инвентарь:
  - Агентив: *сестра*
  - Объектив: *портфель, коробку*
  - Датив: *брату*
  - Инструменталь: *рукой*
  - Локатив: *в комнате*
- Расширение набора сем.падежей (темпоратив, директив…)
- Семантика предложения описывается через связи слова-предиката (сказуемого) с его семантич. падежами
- Синтаксические валентности –
  синтаксическое оформление семантических
  ⇨ *семантико-синтаксическая* модель управления

ЛЕКСИЧЕСКАЯ СОЧЕТАЕМОСТЬ

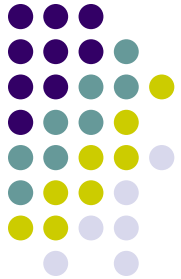


- Более точно: лексико-семантическая
- Не определяется правилами грамматики, зависит от конкретного ЕЯ:
  сильный дождь – ~~strong rain~~, heavy rain
  (пословный перевод не возможен)
- С ней связаны понятия *идиоматичности* и *фразеологичности* словосочетаний

- *Идиоматичность* словосочетания: привычное, традиционное сочетание слов в речи, звучащее естественно для носителей языка *пороть чушь*
- *Фразеологичность*: семантическая спаянность

  Фразеологизм – устойчивое словосочетание, часто употребляемое как целое: *анютины глазки*

ВИДЫ СЛОВОСОЧЕТАНИЙ ПО ФРАЗЕОЛОГИЧНОСТИ

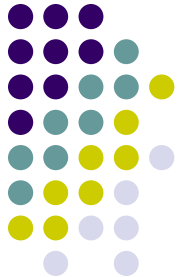


По степени семантической слитности /И.Мельчук/

- *Фразеологические единства* (= *полные фраземы*) – полная семант. спаянность, неотделимость компонент:
  сломя голову, сесть в галошу, черт знает что
- *Полусвободные сочетания* (= *коллокации*) – неполная семантическая слитность
  - смысл частично выводится из компонентов
    острая борьба, вор в законе, роза ветров
  - один компонент сочетается лишь с определенными словами: *затронуть интересы, бросить взгляд*
    to land a job, to stand a comparison with…
- *Свободные* (композиционные) *сочетания* – их смысл складывается из смысла компонент *идти в лес/школу/сад/* допускают вставки: *большой дом, большой красивый дом*

Насколько много словосочетаний этих трех видов ?

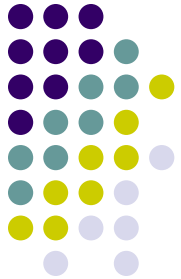
ФРАЗЕОЛОГИЧНОСТЬ: ЛИНГВИСТИЧЕСКИЕ КРИТЕРИИ



- *Фразеологические сочетания* (= *идиомы*, фразеологизмы) объединяют фраземы и коллокации:
  - самостоятельные, неделимые, устойчивые лексические (смысловые) единицы
  - нестандартная сочетаемость, идиоматичность
- Лингвистические критерии фразеологичности:
  - композиционность (свободность) – выводится ли смысл словосочетания из его компонент?
  - синтаксическая гибкость: возможна ли замена одного из слов сочетания синонимом?
    возможна ли вставка внутрь него другого слова?
  - стандартизованность, воспроизводимость в речи, ***устойчивость***

Устойчивость – *лексическая связанность* компонент

ФРАЗЕОЛОГИЧНОСТЬ: ПРИМЕРЫ

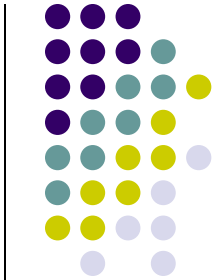


Охарактеризуйте фразеологичность сочетаний:

- *Железная логика*
- *Намотать на ус*
- *Выходить на свободу*
- *В последнюю очередь*
- *Прививать любовь*
- *Держи карман шире*
- *Глубокая реформа*
- *Гусь лапчатый*
- *Получать медаль*
- *Плакали денежки*
- *Личная заинтересованность*
- *Second language learning*

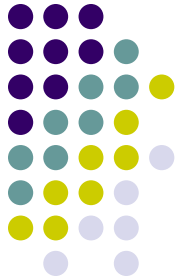
ПОНЯТИЕ КОЛЛОКАЦИИ

Multiword expression (*MWE*)



- В широком смысле *коллокация* – это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости (т.е. устойчивое словосочетание):
  острая борьба, отдавать приказ
- И.Мельчук (лингвист): полусвободное словосочетание
- Возможные свойства:
  - включает ли служебные слова ? *bag of, of the ?*
  - семантически/грамматически допустима?
  - разрывна ли в тексте? *плакали твои денежки ?*
  - устойчива в языке/тексте?
- Наиболее частая трактовка: несколько (2-5) знаменательных слов, синтаксически связанных, возможно разрывных, но устойчиво встречающихся в текстах (в речи) *горячая дискуссия, решить проблему*

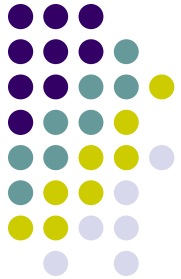
ЗАДАЧА ИЗВЛЕЧЕНИЯ КОЛЛОКАЦИЙ



- Традиционная тема КЛ: более 30 лет, актуальнсть:
- Из-за своих свойств коллокации при анализе текста должны обрабатываться как единое целое, а значит, собираться и храниться в словарях *можно ли без словарей ?*
- Коллокации включают:
  - имена собственные, названия и наименования: *Нижний Тагил, Михаил Таль, Высшая школа экономики*
  - устойчивые обороты (клише): *в первую очередь*
  - производные служебные слова: *за счет, в течение*
  - многословные *термины:* *кратный интеграл, оружие массового поражения, переходный глагол*
- Как коллокации распознавать в тексте и извлекать?
  При каких условиях добавлять в словарь?

Устойчивость в КЛ понимается статистически

МЕТОДЫ ИЗВЛЕЧЕНИЯ КОЛЛОКАЦИЙ



При автоматическом извлечении коллокаций из коллекций/корпусов текстов должны быть учтены:

- Лингвистические критерии охватывают *синтаксические (грамматические) образцы* (т.е. типы) сочетаний:

  $A \leftarrow N$ *полевая форма*

  $V \rightarrow N$ *заметить разницу*

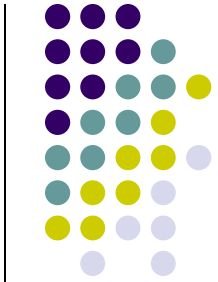
  $N \rightarrow Prep \rightarrow N$ *хлеб с маслом*

  $V \rightarrow и \rightarrow V$ *грабить и убивать* и др.

  (*N – Noun, V – Verb, A – Adjective, Prep – Preposition*)

- Статистические критерии оценивают устойчивость через частоту совместной встречаемости слов
  - простейший критерий: подсчет частоты сочетания
  - более сложные критерии – *меры ассоциации*

СТАТИСТИЧЕСКИЕ КРИТЕРИИ: МЕРЫ АССОЦИАЦИИ



Гипотеза: если употребление слова a

не зависит от употребления слова b, то

$$P(ab) = P(a) * P(b)$$

Меры ассоциации (меры *лексической связанности*)

- проверяют эту гипотезу
- учитывают не только частоту сочетания слов, и частоту входящих в него слов, но также и размер текстовой коллекции/корпуса
- упорядочивают (ранжируют) извлеченные коллокации

Чаще всего применяются для извлечения двусловных неразрывных коллокаций (биграмм)

МЕРА *MI*

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

- Значением меры может быть любое число, зависит от N: чем больше корпус, тем выше в среднем значения меры
- Мера оценивает степень (не)зависимости появления двух слов в корпусе друг от друга
- Если *MI > 1*, то словосочетание статистически значимо (слова употребляются вместе чаще, чем по отдельности)
- Из теории вероятностей:

  I – взаимная информация (*mutual information*),

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

  P – вероятности слов и их сочетаний

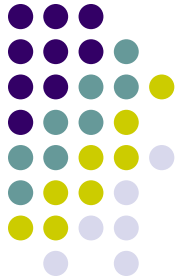
  (если слова независимы, мера = 0, если связаны, то > 0)

МЕРЫ *MI* и MI_3

- Мера MI завышает значимость редких словосочетаний, делая возможным их выявление, но при этом выявляются и случайные сочетания
- MI требует подбора порога отсечения снизу, он подбирается экспериментально (иногда: порог сверху)
- MI можно обобщить для любого числа слов в сочетании
- Возможны модификации, усиливающие влияние отдельных компонент формулы, например: кубическая взаимная информация:

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

НОРМАЛИЗОВАННАЯ ВЗАИМНАЯ ИНФОРМАЦИЯ



Поскольку у Меры MI нет верхней границы, были предприняты попытки по её нормализации

– в частности, для улучшения работы с низкочастотными словами

$$Normalized\,MI(ab) = \frac{\log \dfrac{N \cdot f(ab)}{f(a)\,f(b)}}{-\log \dfrac{f(ab)}{N}} = \frac{\log \dfrac{p(ab)}{p(a)\,p(b)}}{-\log\ p(ab)} = \frac{MI(ab)}{-\log\ p(ab)}$$

MI и MI_3: ПРИМЕР

На основе данных из НКРЯ

N = 229 968 798

	f(a)	*f(b)*	*f(a,b)*	*MI*	MI_3	Ранг
Красивый лес	42915	60367	23	1,03	10,08	3
Красивая дорога	42915	113910	21	-0,02	8,77	4
Железная дорога	29226	113910	9846	9,41	35,94	1
Компьютерная лингвистика	3578	585	5	9,10	13,74	2

MEPA *t-score*

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

- Показывает, насколько неслучайна взаимная встречаемость двух слов в корпусе
- Принимает любые значения
- Зависит от N
- Не требует подбора порога отсечения снизу
- Завышает значимость сочетаний с высоко частотными словами, и в результате извлекаются сложные предлоги, предложные группы, числа
(для их исключения требуются словари стоп-слов)

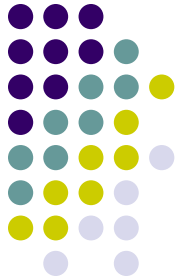
t-score: ПРИМЕР

На основе данных из НКРЯ

N = 229 968 798

	t-score	Ранг	Ранг (MI)
Красивый лес	2,45	2	3
Красивая дорога	-0,06	4	4
Железная дорога	99,08	1	1
Компьютерная лингвистика	2,23	3	2

МЕРЫ *Log-likelihood, Ch-Square*



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

- Выражает (через функцию правдоподобия) отношение гипотез о случайной и неслучайной природе сочетания
- Принимает любые значения, зависит от N
- Дает результат схожий с *t-score*

Эта мера и статистический критерий Хи-квадрат применяются реже других мер

$$Chi - Square(ab) = \frac{(f(ab) - \frac{f(a)f(b)}{N})^2}{f(a)f(b)}$$

МЕРЫ АССОЦИАЦИИ: СРАВНИТЕЛЬНЫЙ ПРИМЕР

/Данные М.В. Хохловой/

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

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***

МЕРА *DICE*

$$Dice = \frac{2 * f(a,b)}{f(a) + f(b)}$$

- Показывает, какую долю от количества словосочетаний с a и с b составляет ab
- Принимает значения от 0 до 1
- Не зависит от N

На основе данных из НКРЯ:

	Dice	Ранг	Ранг (MI)	Ранг (t-score)
Красивый лес	0,00045	3	3	2
Красивая дорога	0,00027	4	4	4
Железная дорога	0,01376	1	1	1
Компьютерная лингвистика	0,00240	2	2	3

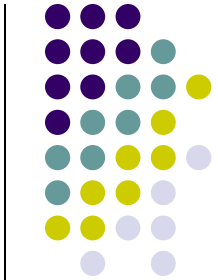
ОСОБЕННОСТИ МЕР АССОЦИАЦИЙ

- *Ранги* (порядковые номера) распознанных коллокаций для разных мер часто не совпадают,
  в целом результаты применения различных мер могут существенно различаться
- Не совпадают и результаты при использовании мер для словоформ и для лемм (нормализованных слов)
- Результаты зависят от объема и типа корпуса (например, 6 млн. или же 200 тыс. слов)
  Для текстов разных жанров – разные меры?
- Мера MI_3, возможно, дает наилучшие усредненные результаты.
- Общая проблема – разрывные коллокации
- Тем не менее: все высокоранговые коллокации обычно входят в словари устойчивых словосочетаний

КОЛЛОКАЦИИ В ПРИЛОЖЕНИЯХ КЛ

- Составление и обновление словарей:
  - фразеологизмов
  - устойчивых словосочетаний, в том числе: полусвободных сочетаний, имен, названий фирм и др.
- Обучение родному и иностранному языкам
- Машинный перевод
- Генерация и литературная правка текстов
- Автоматизированное исправление лексических ошибок: случайных и стилистических

БАЗЫ СЛОВОСОЧЕТАНИЙ РЯ



Система ***КроссЛексика*** – словарь сочетаемости для обучения языку и помощи при написании текстов

- База: 1.75 млн.свободных и полусвободных словосочетаний различных тематик
- Словосочетания отбирались вручную (с 1990 г.)
- Основные источники базы: словари по разным тематикам, тексты Интернета, статистика Яндекса

CoSyCo – база синтаксически связанных словосочетаний, для лингвист. исследований и применения в задачах АОТ

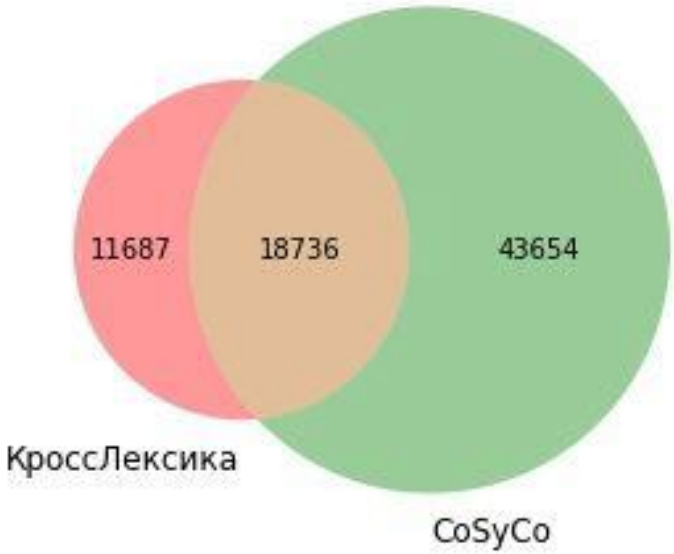
- собрана автоматически (поверхн.синтаксический анализ)
- База данных: только словосочетаний определительного вида $A + N$ ≈ 20 млн. единиц
- Основные источники базы: тексты новостей, коллекция Либрусек, научные статьи, ядро русской Википедии

СРАВНЕНИЕ БАЗ КРОССЛЕКСИКИ И *COSYCO*

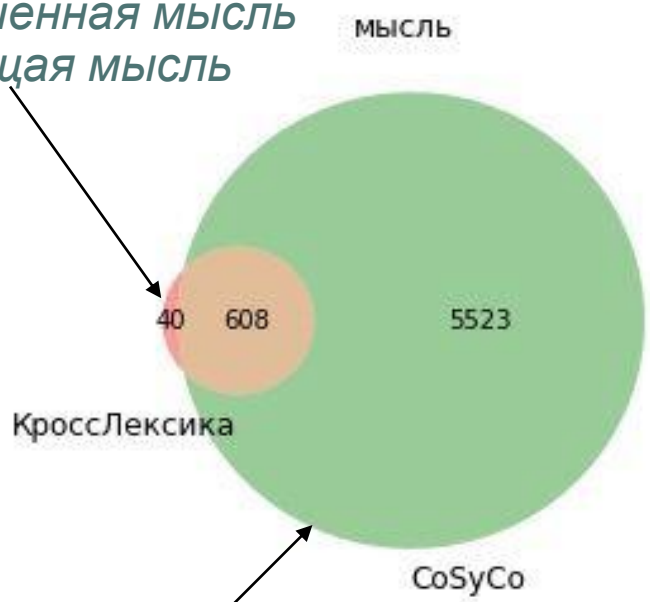
Словосочетания вида A + N для слова *мысль*

Устойчивые (коллокации)?

Число существительных, входящих в словосочетания вида A + N

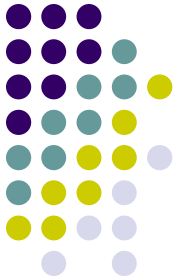


Коллокации:
- *захватывающая мысль*
- *внушенная мысль*
- *ноющая мысль*



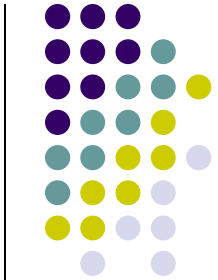
Коллокации:
- *посторонняя мысль*
- *непрошеная мысль*
- *грязная мысль*

ПРИЛОЖЕНИЯ: ИСПРАВЛЕНИЕ ЛЕКСИЧЕСКИХ ОШИБОК



- Один из видов лексических ошибок:
  одно слово заменяется похожим, но с другим смыслом:
  Появилась спасательная мысль, массивная миграция, оказывать оральную поддержку…
- Причины:
  - случайная ошибка (*малапропизм*):
    трафик ввода, сачок цен, неутомимый голод
  - ошибочное употребление близких по смыслу слов:
    паронимические ошибки: деловой вместо *деловитый*
    массивный вместо *массовый*
- Сложность автоматического выявления:
  одно слово текста заменяется другим, причем в той же синтаксической роли и с теми же морфологическими признаками (число, род, падеж, лицо)

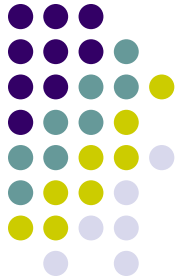
МЕТОД ОБНАРУЖЕНИЯ И ИСПРАВЛЕНИЯ ОШИБОК



Предполагается: ошибки разрушают *лексическую связность* пары слов, но сохраняют их синтаксическую связанность

- Просмотр всех пар знаменательных слов предложения, проверка их на синтаксическую и лексическую связность
- Если лексическая связность нарушается (например, значение меры ***MI*** ниже установленного порога), то сигнализируется ошибка
- Отбор кандидатов на исправление из словарей *паронимов*:
  - *буквенных* (*каска*: *качка*, *кашка*, *краска* и др.) или
  - *морфемных* (слова с одинаковым корнем: *человечный*, *человеческий*, *очеловеченный*)
- Проверка кандидатов на *лексическую связность,* ранжирование и предъявление пользователю

COMPUTATIONAL TERMINOLOGY

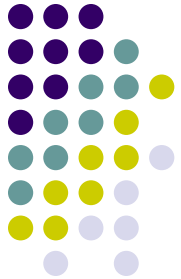


Вычислительная терминология

- Центральная задача –
  автоматическое извлечение терминов и их связей
- *Терминологические слова и словосочетания*:
  называют понятия специальной области знаний
  (науки, техники, искусства, общественной жизни):

  регистр адреса, число с плавающей точкой,
  пенсионное обеспечение, земноводные, piano
- Термины, за исключением базовых, имеют
  определения, основанные на других терминах,
  образуя тем самым терминологическую систему
- Важно: в совокупности термины представляют знания
  о предметной области
- Вычислительные критерии их извлечения? Свойства?

ТЕРМИНЫ И КОЛЛОКАЦИИ



- Многие терминологические словосочетания
  – некомпозиционны (их смысл не выводится из смысла компонент): *корень уравнения*
  – устойчивы, т.е. коллокации: *бюджетные средства*
- Грамматическая структура терминов: чаще всего – именные словосочетания, реже – глаголы, наречия и др.

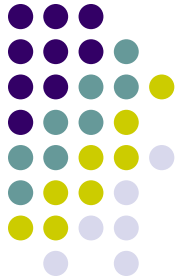
  Можно описать *грамматическими образцами:*

  N + N – *период упреждения*

  A + A + N – *абстрактная семантическая сеть*

  A + N + N – *спектральный коэффициент излучения* и т.п.
- Извлечение терминов их текстов: методы сходны с извлечением коллокаций, но есть отличия
- Сложность: в текстах перемешаны терминология и общая лексика, например, общенаучная лексика в научных текстах: *проблема, определение, другой*

ПРИМЕРЫ ТЕРМИН *Vs.* НЕТЕРМИН



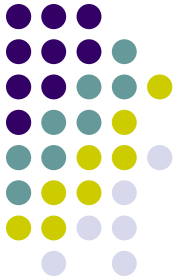
Из математической области:

- *дифференциал высшего порядка*
- *неизвестный параметр*
- *свойства функции*
- *наклонная асимптота*
- *разностный метод*
- *неравенство Бесселя*
- *погрешность решения*
- *единственное решение*
- *постоянный коэффициент*
- *сумма углов*
- *простая итерация*
- *идея метода*
- *уравнение окружности*
- *формальный параметр*
- *способ построения*

Какие из них термины?

- В словарях и тезаурусах есть нее все термины, постоянно возникают новые
- Сложность отбора терминов: Термин соотносится с некоторым понятием специальной области знаний – это обычно выявляет эксперт-терминолог

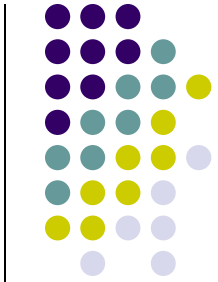
КРИТЕРИИ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ



Признаки терминов: каждая группа отражает специфичные их свойства

- Лингвистические свойства (ограничения) – отбор кандидатов в термины
- Статистические признаки – ранжирование кандидатов
  - Вычисляются по *базовому* тексту/коллекции ПО или с привлечением *контрастной коллекции* (например, общелитературных текстов)
  - Могут учитывать контекст (окружение в тексте):
    – вложенность в объемлющие словосочетания: *мера терминологичности C-Value*
    – разнообразие (частотность) контекстов
  - Учет лексической связанности: меры ассоциации,
- Обычно: комбинация критериев, в итоге – упорядоченный список *кандидатов в термины*

ИЗВЛЕЧЕНИЕ ТЕРМИНОВ: ЛИНГВИСТИЧЕСКИЕ ПРИЗНАКИ



- Грамматические (синтаксические) образцы терминов, например: *A + N + N*

  спектральный коэффициент излучения

- Графематич. признаки: написание с большой буквы
- Лексические особенности: для исключения нетерминов – списки *стоп-слов*, включающие имена, фамилии, географические названия, слова общей лексики и оценочные слова: *каждый, шаг, плохой* и т.д.
- Контексты употребления терминов:
  - Новые термины часто определяются в тексте:

    Под прерыванием понимается сигнал…

  Эту информацию можно записать в виде *лингвистических шаблонов*

МЕРА ТЕРМИНОЛОГИЧНОСТИ *C-VALUE*

Ранжирует термины, поощряя отбор словосочетаний большей длины, которые не входят в состав других:

$$\text{C - Value(a)} = \begin{cases} \log_2 |a| * freq(a) \quad , \quad \textit{если не вложен} \\ \log_2 |a| - (freq(a) - \frac{1}{P(T_a)} * \sum_{b \in T_a} freq(b)) \end{cases}$$

a – кандидат в термины,

$|a|$ – длина словосочетания (количество слов)

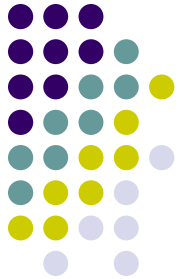
$freq(a)$ – частотность a

T_a – множество словосочетаний, которые содержат a

$P(T_a)$ – количество словосочетаний, содержащих a

электрический слой – *двойной электрический слой*

СТАТИСТИЧЕСКИЕ ПРИЗНАКИ: БАЗОВАЯ КОЛЛЕКЦИЯ ТЕКСТОВ

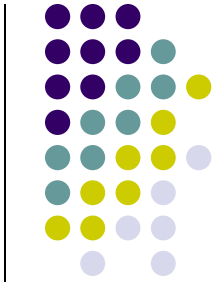


Базовая /исходная /целевая коллекция текстов

- *TF* (*term frequency*) – абсол. частота употребления слов и словосочетаний – иногда неплохой критерий
- *DF* (*document frequency*) – документная частотность
- Мера *TF-IDF* (из информационного поиска) оценивает тематическую содержательность кандидата в термины

  предлоги, артикли получают *TF-IDF* = 0, поскольку они есть во всех документах коллекции
- Меры ассоциации, но не всегда работают хорошо
- Вклад признаков в извлечение терминов может зависеть от особенностей предметной области

СТАТИСТИЧЕСКИЕ КРИТЕРИИ: КОНТРАСТНАЯ КОЛЛЕКЦИЯ



- Мера *Weirdness* – соотношение относит. частоты кандидата в основной и контрастной текстовой коллекции

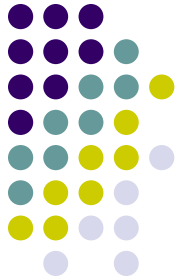
$$Weirdness(w) = \frac{TF_t(w)}{|W_t|} \Big/ \frac{TF_r(w)}{|W_r|}$$

где TF_t и TF_r – частотности слова в основной и контр. коллекциях, $|W_t|$ и $|W_r|$ – число слов в этих коллекциях

- Модификации *TF-IDF*:
  - *Contrastive Weight:* $\log(TF_t) * \log\left(\frac{|W_t| + |W_r|}{TF_t + TF_r}\right)$
  - *KF-IDF:* отражает новизну слова в тестовой коллекции

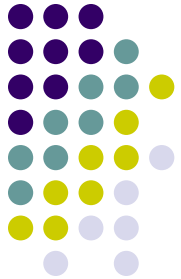
$$DF * \log\left(\frac{2}{|D|_w} + 1\right) \quad |D|_w = \begin{cases} 1, \text{ если слова нет в контрастной} \\ 2, \text{ если слово есть в контрастной} \end{cases}$$

ТЕХНОЛОГИЯ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ КОЛЛЕКЦИЙ ТЕКСТОВ



- Формирование текстовой коллекции для данной предметной области (мега- и гигабайты текстов)
- Процедуры автоматического извлечения слов и словосочетаний по выбранным признакам/критериям
- Ранжирование итогового набора извлеченных терминов-кандидатов
- Работа экспертов по проверке и отбору терминов: чем дальше от начала ранжированного списка извлеченных кандидатов, тем меньше реальных терминов:
  рациональное выражение, поверхность тела, точка координатной плоскости, основание трапеции, теория сплайнов …

МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧЕ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ

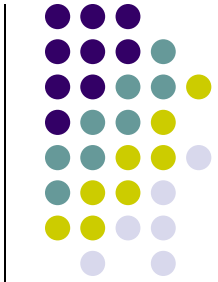


- Извлечение терминов – многофакторный процесс, учитывающий различные признаки терминов
- Поиск наилучшей комбинации признаков – на базе методов машинного обучения
  - использование для обучения размеченных терминологических ресурсов (например, тезаурусов)
  - учет большого числа признаков
  - например, метод логистической регрессии:

$$\sigma(a_1x_1 + a_2x_2 + \ldots + a_nx_n), \qquad \sigma(z) = \frac{1}{1+\exp(-z)}$$

- Задача: упорядочить список извлекаемых терминов-кандидатов так, чтобы максимальное число реальных терминов оказалось в начале списка
- Обученная модель – классификатор: Термин/Нетермин

ОЦЕНКА КАЧЕСТВА РАНЖИРОВАНИЯ ТЕРМИНОВ



- *Средняя точность AvP* (*Average Precision)* – адаптирована из информационного поиска):

$$AvP(D) = \frac{1}{\left|D_q\right|} \sum_{1 \le k \le |D|} \left(r_k \times \left(\frac{1}{k} \sum_{1 \le i \le k} r_i \right) \right)$$

D – множество из k терминов-кандидатов

$D_q \subseteq D$ – подмножество действительных терминов

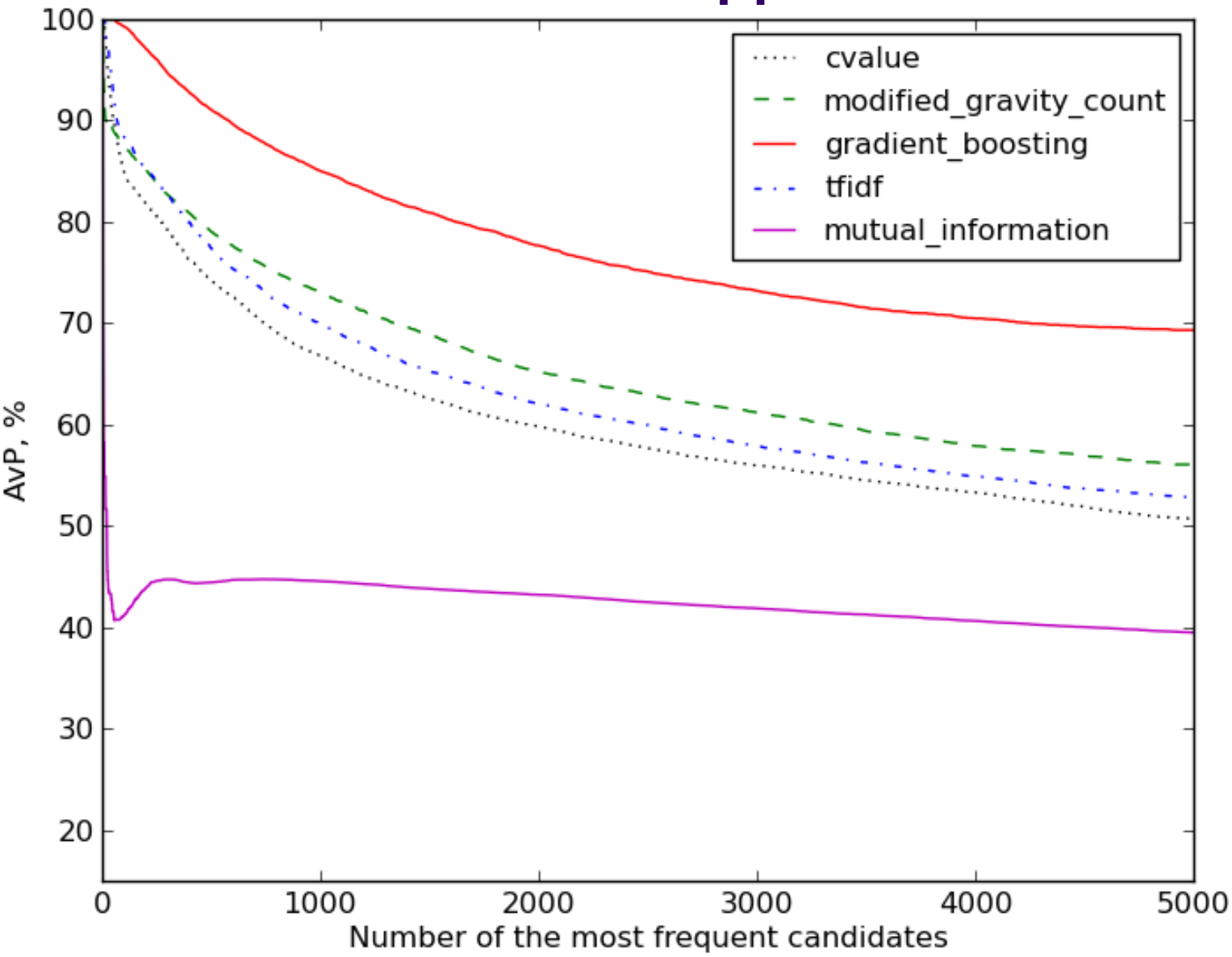
$r_i = 1$, если i-ый кандидат – термин, и $r_i = 0$ иначе

- Мера тем выше, чем больше терминов в начале списка
- Пример:

T, N, T AvP = (1/2) (1+2/3) = 5/6 = 0.888..

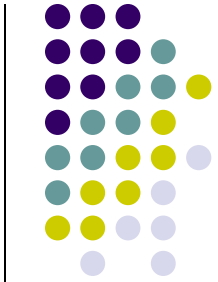
N, T, T AvP = (1/2) (1/2+2/3) = 7/12 = 0.68..

СРАВНЕНИЕ МЕТОДОВ ИЗВЛЕЧЕНИЯ



Банковская коллекция: 104 тыс. док., 15.5 млн.слов 60 признаков , метод *Gradient boosting* находит оптим. комбинацию признаков для предсказания терминов

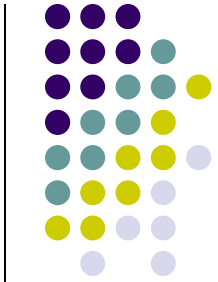
ИЗВЛЕЧЕНИЕ ТЕРМИНОВ: ПРИЛОЖЕНИЯ



Практика зависит от типа создаваемых ресурсов

- Построение по коллекциям текстов ПО терминологических ресурсов:
  - Словарь терминов (несколько сотен /тысяч терминов)
  - Информационно-поисковый Тезаурус (несколько десятков тысяч терминов и их смысловые связи)
    – стандарты: рекомендации по включению кандидатов
- Анализ и обработка отдельного текста (*Single Document Processing*):
  - Автоматическое *индексирование* текстов (выявление ключевых слов)
  - Построение *глоссариев* и *предметных указателей*
  - Быстрая навигация по объемному документу

ЗАКЛЮЧЕНИЕ



- Словосочетания – подуровень синтаксиса, сочетаемость определяется грамматическими, семантическими и лексическими факторами.
- Нестандартная лексическая сочетаемость выражается в КЛ понятием *коллокации*, для их извлечения из текстов применяются *меры ассоциации* (лексической связности)
- Для извлечения терминологических словосочетаний обычно используются комбинации лингвистических и статистических признаков, их вклад зависит от особенностей текстов предметной области
- Нет наилучшего метода для извлечения коллокаций или терминов, для конкретных приложений нужен подбор комбинации их признаков.

СПАСИБО ЗА ВНИМАНИЕ