Построение словарных ресурсов

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание

- Словари и их составление
- Извлечение коллокаций
  - сочетаемость в ЕЯ
  - понятие коллокации в КЛ
  - меры ассоциации для извлечения коллокаций
  - коллокации в задачах КЛ
- Извлечение терминов
  - особенности терминов
  - методы выявления терминов и связей между ними
- Дистрибутивная семантика и технология Word2Vec

Поставьте правильное ударение

- Тефтели
- Апартаменты
- Ржаветь
- Творог
- Гренки
- Жалюзи
- Средам (день недели)
- Обеспечение
- Асимметрия
- Одновременно
- Граффити
- Аранжировать
- Повторим
- Ненецкий
- Однояйцевые
- Памятуя
- Обнял
- Издревле
- Знамение
- Завидно
- Банты
- Балашиха

Правильное ударение по Викисловарю

- Т**е**фт**е**ли
- Апарт**а**м**е**нты (двойное ударение?)
- Рж**а**в**е**ть
- Тв**о**р**о**г
- Гренк**и** (один грен**о**к)
- Жалюз**и**
- Сред**а**м (день недели)
- Обеспеч**е**н**е**ние
- Асимметр**и**я
- Одновр**е**м**е**нно (двойное ударение?)
- Граффи**т**и
- Аранж**и**ровать
- Повтор**и**м
- Н**е**нецкий
- Однояйцев**ы**е (слова нет)
- П**а**мятуя (слова нет)
- **О**бнял (двойное ударение?)
- Издр**е**вле
- Зн**а**мение
- Зав**и**дно
- Б**а**нты
- Балаш**и**ха

Литературная норма и речевая практика

Литературная норма – предписания словарей и грамматик, касающиеся правил употребления языковых средств

Речевая практика – представления носителей языка о норме

- Представления неодинаковы у разных носителей языка
- Представление зависит от личностных свойств и того, как и где индивид провел жизнь
- Любые слова имеют единообразное понимание лишь в обществе, где социальные характеристики индивидов чрезвычайно близки

Норма и содержание словарей

- В разных словарях указана речевая практика разных городов, народностей, социальных групп, профессий
- Составитель словаря нередко фиксирует в нем свою речевую практику
- Норма более или менее общепринята для ядра языка, но ее не существует для
  - малочастотной лексики
  - новых значений слов и моделей управления
  - распределения конкурирующих способов выражения

    шаурма, *шаверма*, *шаварма*

Словари и способы их составления

- В разных задачах КЛ требуются разные словари
- В словарь входят:
  - слова/словосочетания (термины, стоп-слова, дискурсные слова и выражения и пр.)
  - статистическая/лингвистическая информация (частота употребления, определения, связи между терминами и т.д.)
- Способы составления словарей:
  - ручной (работа с картотеками и информантами)
  - автоматический по коллекциям текстов
- Составление словарей вручную не лучший способ из-за быстрой изменчивости ЕЯ и субъективности автора

Автоматическое составление словарей

- Предобработка коллекции текстов
- Отбор кандидатов (слова/словосочетания, соответствующие шаблонам, N-граммы, коллокации)
- Присвоение кандидатам весов
- Ранжирование по весам и отбор наиболее подходящих слов и словосочетаний

✓ Проблема: корпус может неадекватно отражать употребление исследуемых единиц (низкая частота, неверное употребление и др.)

Словари словосочетаний

- Сейчас актуальны словари словосочетаний
- Сочетаемость слов ЕЯ – способность слов соединяться друг с другом, образовывая единицы более высокого уровня
- Есть свободные словосочетания: их смысл складывается из смысла входящих в них слов *снежная горка*, *дорога из желтого кирпича*
- Есть несвободные словосочетания: их смысл полностью не складывается из смысла слов *дать сдачи*, *острая борьба*
- Именно несвободные словосочетания включаются в словари
- В КЛ они называются *коллокациями*

Понятие коллокации

Нет однозначного понимания этого термина
в теоретических и прикладных работах

- И. Мельчук (лингвист): словосочетание, смысл которого частично выводится из его компонентов *затронуть интересы*, *острая борьба*, *вор в законе*
- J. R. Firth (лингвист), 1957: *«часто встречающиеся сочетания слов, чье появление рядом основывается на регулярном характере взаимного ожидания и задается не грамматическими, а чисто семантическими факторами»*
- Компьютерная лингвистика: комбинация двух и более слов, имеющих тенденцию к совместной встречаемости

Признаки коллокаций

- Возможные характеристики коллокации (зависят от рассматриваемой задачи):
  - содержит несколько (2-5) слов
  - включает знаменательные (реже – служебные) слова
  - осмысленна и ее смысл не выводится из входящих в нее слов
  - устойчива – воспроизводима в речи в виде готовой единицы
  - частотна
- При автоматической обработке устойчивость понимается статистически

Коллокации: примеры

- Имена собственные, названия и наименования:
  Нижний Новгород, *Анхела Меркель*, *Высшая школа экономики*
- Устойчивые обороты (клише):
  решить проблему, *в первую очередь*
- Производные служебные слова:
  за счет, *в течение*, *несмотря на*
- Многословные термины:
  оружие массового поражения, *вытесняющая многозадачность*

Методы извлечения коллокаций

- Лингвистические критерии – синтаксические образцы коллокаций

  A ← N *полевая форма*
  V → N *заметить разницу*
  N→ Prep → N *хлеб с маслом* и др.

- Статистические критерии – частота совместной встречаемости слов (для оценки устойчивости)
  - простейший критерий: подсчет частоты коллокации
  - более сложные критерии: *меры ассоциации*

Меры ассоциации

Гипотеза: если употребление слова a не зависит от употребления слова b, то

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

Меры ассоциации (связанности) проверяют эту гипотезу

- Чаще всего применяются для извлечения двусловных неразрывных коллокаций
- Упорядочивают (ранжируют) извлеченные коллокации
- Учитывают не только частоту сочетания, но и частоту входящих в него слов и размер коллекции/корпуса

Меры ассоциации: используемые обозначения

- Применяемые меры (*association measures*):
  - *mutual Information: MI* и MI_3
  - *t-score* (*t-тест*, *t-критерий Стьюдента*)
  - *log-likelihood* и т.п.
- Используемые обозначения:

  N – размер корпуса в словах или словоформах

  $f(a)$ – частота встречаемости слова a

  $f(b)$ – частота встречаемости слова b

  $f(a,b)$ – частота (*frequency*) совместной встречаемости слов a и b

Мера MI

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

- Оценивает степень зависимости появления двух слов в корпусе друг от друга
- Значением может являться любое число; если $MI > 1$, то словосочетание статистически значимо
- Зависит от N: чем больше корпус, тем выше в среднем получаемые по нему значения меры
- MI можно обобщить для любого числа слов в словосочетании

Модификация MI: MI_3

Мера *MI* завышает значимость редких словосочетаний ➔ выявляются опечатки

- требует *порог отсечения* по частоте снизу, подбираемый экспериментально; иногда подбирают и порог сверху
- существуют модификации *MI*, которые пытаются устранить данный недостаток

$$MI_3 = \log_2 \frac{f^3(a,b) * N}{f(a) * f(b)}$$

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

Мера t-score

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

- Показывает, насколько неслучайна взаимная встречаемость двух слов в корпусе
- Принимает любые значения
- Зависит от N
- Не требует порога отсечения снизу по частоте
- Завышает значимость сочетаний с частотными словами ➔ извлекаются сложные предлоги, числа

✓ для исключения требуется заранее составлять списки стоп-слов

t-score: пример

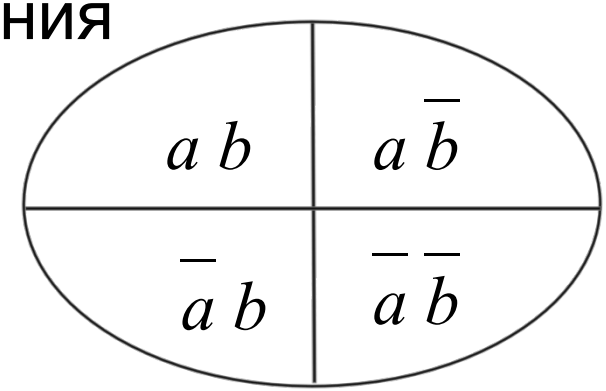
На основе данных из НКРЯ
N = 229 968 798

	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

Мера log-likelihood

$$log\text{-}likelihood = 2 \sum_{A,B} f(A,B) * \log_2 \frac{f(A,B) * N}{f(a) * f(b)}$$

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



A – это a или $\overline{a}$, B – это b или $\overline{b}$

Особенности мер ассоциации

- *Ранги* (порядковые номера) извлеченных коллокаций для разных мер могут не совпадать
- Результаты извлечения при использовании мер для словоформ и для лемм (нормализованных слов) не совпадают
- Высокоранговые коллокации часто входят в словари устойчивых словосочетаний (коллокаций)
- Общая проблема – разрывные коллокации
- Меры зависят от N ➔ результаты зависят от объема и типа корпуса

✓ для текстов разных жанров – разные меры?

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

Коллокации в задачах КЛ

- Машинный перевод:
  как правило, коллокации не переводятся дословно
- Обучение иностранным языкам:
  владение языка определяется, в частности, знанием нестандартной сочетаемости
- Исправление лексических ошибок, обусловленных опечатками или неверным употреблением близких по смыслу слов *сачок цен*, *неутомимый голод*, *деловой* вместо *деловитый*, *массивный* вместо *массовый*
- Извлечение терминологических словосочетаний

Термины

- *Термины* – слова и словосочетания, называющие понятия предметной области
- Смысл большинства терминов не выводится из смысла их частей ➔ термины – это коллокации (?)
- Свойства термина:
  - наличие определения
  - связь слово/словосочетание ⇔ обозначаемое понятие обеспечивается соглашениями между учеными (определением термина)
  - тенденция к однозначности в пределах своей ПО: не должно быть омонимов, полисемии
  - отсутствие экспрессии, стилистическая нейтральность

Термины: языковая форма

- Имена существительные и именные словосочетания (преимущественно) *метафаза*, *пенсионное обеспечение*, *период упреждения*, *абберация оптической систем*
- Глаголы (реже, но не реже чем вообще в языке): *компилировать*, *опреснять*
- Прилагательные (обычно – часть термина): *земноводные*, *слепой*, *инфантильный*
- Наречия (редко): *piano*, *электростатически*
- Фразы (очень редко): *отдать швартовы*, *раздернуть снасть*

Термины и общая лексика

- Типы лексических единиц ЕЯ:
  - только нетермины: *поварешка*, *сугроб*
  - только термины: *γ-лучи*, *митоз*, *дифракция*
  - как термин, так и нетермин: *корень*, *соль*
- Граница между единицами общей лексики и терминами расплывчата (есть промежуточный слой)
- Границы между терминологиями разных предметных областей также расплывчата *функция*, *дорожная карта*, *ядро*

Примеры терминов/ нетерминов

Из математической области:

свойства функции
простая итерация
наклонная асимптота
точка функции
математический аппарат
неизвестный параметр
единственное решение
сумма углов
постоянный коэффициент

- Скорей всего не все есть в словарях. Можно ли утверждать, что они не термины?
- Соотносится ли слово/словосочетание с понятием, может определить только эксперт-терминолог

Задача извлечения терминов

- Необходимо извлечь термины из научно-технических текстов (НТТ)
- Сложность:
  - в НТТ перемешаны термины и нетермины
  - термины неидеальны
- Лексический состав научных текстов:
  - ✓ термины
  - ✓ словосочетания с терминами
  - ✓ общенаучная лексика: *проблема*, *определение*
  - ✓ общая лексика (в том числе служебные слова)
- Не существует формального критерия термина ➔ нужны вычислительные критерии

Неидеальность терминов

Реальные термины:

- **Не имеют [точного] определения**. Это зависит в том числе от степени зрелости ПО
- **Многозначны**. Обычный тип многозначности в тексте – перенос слова на смежное явление. Для разрешения требуется контекст

  тригонометрические функции:
  косинус=функция косинус

  отношение сторон в прямоугольном треугольнике:
  косинус=косинус угла
- **Вариативны**. Бывают варианты:
  - ✓ записанные в словаре – синонимы (дублеты)
  - ✓ возникающие в тексте
    дисковый контроллер – контроллер диска

Критерии извлечения терминов (1)

Какие свойства терминов можно использовать для извлечения потенциальных терминов?

- Термин – это коллокация
  нулевое окончание, *бюджетные средства*
- Термины можно описать синтаксическими образцами
  A N *логический вывод*, *темная материя*
  N N *период упреждения*, *шина адреса*
- Новые термины часто определяются в тексте:
  Под прерыванием понимается сигнал…

Для извлечения терминов можно использовать их статистические и лингвистические свойства

Критерии извлечения терминов (2)

- Статистические критерии:
  - частота встречаемости
  - специальные метрики (C-Value) и меры ассоциации для коллокаций
- Лингвистические критерии:
  - грамматические (образцы терминов)
  - лексические (списки стоп-слов – обычно не входят в термины: *каждый*, *другой*, *плохой*)
  - контекстные (окружение термина в тексте)
- Обычно: комбинация критериев
- Результат – упорядоченный список *кандидатов в термины*

Метод на основе C-Value

1. Из текста с помощью шаблонов извлекается множество слов и словосочетаний, имеющих определенную грамматическую структуру
2. Множество слов и словосочетаний сокращается на основе списка стоп-слов
3. Слова и словосочетания упорядочиваются согласно значению метрики C-Value

Особенность: C-Value поощряет отбор словосочетаний большей длины, не входящих в состав других словосочетаний

Метрика C-Value

$$\text{C-Value(a)} = \begin{cases} \log_2 |a| * fr(a) & (1) \\ \log_2 |a| * (fr(a) - \frac{1}{P(T_a)} * \sum_{b \in T_a} fr(b)) & (2) \end{cases}$$

(1) если a не вложено в другие словосочетания
(2) если a вложено в другие словосочетания

a – кандидат в термины,
$|a|$ – длина словосочетания (количество слов)
$fr(a)$ – частота a
T_a – множество словосочетаний, содержащих a
$P(T_a)$ – количество словосочетаний, содержащих a

Метрика C-Value: пример

Пусть есть термины *число*, *представление числа*, *число большой разрядности*

fr(*представление числа*) = 1

fr(*число большой разрядности*) = 1

fr(*число*) = 3

число является кандидатом

C-value(*представление числа*) =

C-value(*число большой разрядности*) = 1

C-value(*число*) = 0

число не является кандидатом

Особенности списка терминов-кандидатов

- Методы извлечения работают хорошо для первых 100-200 словосочетаний упорядоченного списка
- Далее число терминов среди выявленных сочетаний уменьшается, но они там есть

➔ нужна работа экспертов по отбору терминов

➔ нужно другое ранжирование (машинное обучение)

Фрагмент списка кандидатов, упорядоченного по частоте, с 403-ого элемента и далее:

точка координатной плоскости
основание трапеции
поверхность тела
теория сплайнов
вариант разложения
интервал убывания

Извлечение терминов: машинное обучение

- Задача: упорядочить список терминов-кандидатов так, чтобы максимальное число реальных терминов оказалось в начале списка
- Для этого необходимо **скомбинировать** различные лингвистические и статистические признаки
- Для поиска наилучшей комбинации признаков – машинное обучение

Признаки терминов для машинного обучения

- Вычисляются на основе
  - базовой, основной (*target*) коллекции текстов
  - контрастной (*reference*) коллекции
- Лингвистические признаки (булевские признаки для однословных терминов):
  - *неоднозначность*: имеет ли слово более одного варианта нормализации
  - *новизна:* отсутствует ли слово в морфословаре
  - *специфичность*: присутствует ли слово в контрастной коллекции
- Контекстные признаки
  - вложенность в объемлющие словосочетания
  - разнообразие контекстов

Статистические признаки для машинного обучения

- Частота употребления
- Документная частотность
- Мера tf-idf
- Относительная частотность (странность) – поощряет слова, встречающиеся чаще в базовой коллекции, чем в контрастной
- Значение мер ассоциации

Вклад признаков в извлечение терминов может зависеть от особенностей предметной области

Оценка качества упорядочивания терминов

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

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

D – множество из k терминов-кандидатов

$D_q \subseteq D$ – подмножество действительно терминов

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

- *AvP* 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..

Извлечение связей терминов

Виды связей (примеры):

- ✧ род-вид — *мебель – мягкая мебель*
- ✧ часть-целое — *клетка – ядро*
- ✧ синонимы — *траектория – путь*
- ✧ причина-следствие — *сила – ускорение*

Методы распознавания:

- Распознавание контекстов употребления терминов на базе шаблонов:
  such T1 as T2 – such crimes as money laundering
- Оценка совместной встречаемости терминов в одном документе
- Методы дистрибутивной семантики

Дистрибутивная семантика

- Основана на изучении окружения (*дистрибуции*, *распределения*) единиц в тексте
- Слова имеют сходство в значении, если они употребляются в схожих контекстах: *кофе, чай, сок*
- Метод дистрибутивного анализа (сходство в значении по контексту слова): по коллекции текстов строятся классы близких по семантике слов
- Можно выявить слова с общими *семами* (элементами смысла):
  - Синонимы
  - Антонимы
  - Слова одного семантического класса

Технология Word2Vec

- Статистическая обработка больших текстовых массивов на любом языке
- Вход: корпус/коллекция текстов и параметры
  - модель и алгоритм обучения
  - размер контекста вокруг слова (от 3 до 10 слов)
  - размерность результирующих векторов
- Выход – набор (пространство) векторов чисел
  - размерность пространства – обычно несколько сотен
  - каждому уникальному слову – свой вектор
- Смысл имеют только расстояния между векторами, а не сами вектора
- Векторы слов, имеющие общие контексты в корпусе, близки в построенном пространстве (косинусная мера)

Word2Vec: примеры

Упрощенно: близость контекстов – близость слов
Примеры работы: вывод по мере близости

avito

awito 0.693721
авито 0.675299
fvito 0.661414
авита 0.659454
irr 0.642429
овито 0.606189
avimo 0.598056

mail

rambler 0.777771
meil 0.765292
inbox 0.745602
maill 0.741604
yandex 0.696301
maii 0.675455
myrambler 0.674704

Word2Vec: модели

Нейронная двухслойная сеть прямого распространения

- CBOW (*Continuous Bag-of-Words*)
  - ✓ предсказывает слова при заданном контексте
  - ✓ контекст – «мешок слов»
  - ✓ работает быстро, лучше для больших корпусов (т.к. частота слов выше)
- SkipGrams (*Continuous Skip-n-Grams*)
  - ✓ предсказывает контекст при заданном слове
  - ✓ контекст – N-граммы
  - ✓ работает медленнее, но лучше для редких слов и небольших корпусов (<100 млн)

CBOW и SkipGrams для слова *кофе*

Взаимозаменяемые слова:

коффе 0.734483
чая 0.690234
чай 0.688656
капучино 0.666638
кофн 0.636362
какао 0.619801
эспрессо 0.599390
кофя 0.595211
цикорий 0.594247
кофэ 0.593993
копучино 0.587324
шоколад 0.585655
капучинно 0.580286

Характеризующие слова:

зернах 0.757635
растворимый 0.709936
чая 0.709579
коффе 0.704036
mellanrost 0.694822
сублемированный 0.694553
молотый 0.690066
кофейные 0.680409
чай 0.679867
декофеинизированный 0.67856
капучино 0.677856
monoarabica 0.676757

Word2Vec: применение

- Исправление опечаток и неверной транслитерации слов, например, для *пщщпду*
  пщщщпду - 0.723,… гугл - 0.649, поопду - 0.647…
- Построение классов семантически близких слов для разных приложений КЛ
- Определение аналогии между словами (семантических отношений)
  Найти такое слово, которое относится к Германии также, как Париж относится ко Франции:
  Мюнхен, Берлин, Дюссельдорф, Гамбург, …
- Расширение запросов к поисковой системе (учет статистики реальных ошибок)
  - генерация контекстно-близких слов
  - оценка важности слов в запросе

Термины: приложения КЛ

- Обработка отдельно взятого текста
  - автоматическое *индексирование*
  - построение *глоссариев* и *предметных указателей*
  - выявление ключевых слов документа
  - автоматизация терминологического редактирования
  - быстрая навигация по объемному тексту
- Обработка коллекций текстов – более развитое направление

Извлечение терминов и их связей из коллекций текстов

- Извлечение знаний (определений, терминов и отношений между ними) из текстов новой ПО
- Разработка терминологических ресурсов: словарей, тезаурусов, онтологий
- Предметно-ориентированный информационный поиск, например, расширение запроса на основе извлеченных синонимов:
  договор поручительства – договор поручения
  решение проблемы – решение задачи

Особенности извлечения ключевых слов

- *Ключевое слово* – слово или словосочетание, способное в совокупности с другими ключевыми словами представлять смысл текста
- Сложности выявления:
  - могут не быть терминами
  - могут не входить в текст
- Идеи выявления:
  - ✓ учет лингвистических свойств (структура, контексты, словари)
  - ✓ учет статистических свойств (tf-idf)
  - ✓ учет структуры текста (выбор слов из заголовков и первых предложений)
  - ✓ опора на семантический граф текста

Заключение (1)

- Коллокации – устойчивые в статистическом смысле словосочетания
- Их извлечение опирается на комбинацию статистических и лингвистических критериев и частичный синтаксический анализ
- Нет идеальных критериев извлечения, необходимо ориентироваться на конкретные приложения

Заключение (2)

- Терминологические словосочетания – тоже коллокации
- Результат работы методов извлечения терминов – список кандидатов
- «Статистика» хорошо работает для начала упорядоченного списка
- «Лингвистика» не универсальна, т.к. нет строгих правил именования понятий и выражения отношений
- При использовании машинного обучения большинство терминов оказывается в начале списка
- При создании словарных ресурсов дополнительно можно использовать аналогичные для обучения
- Способы оценки – полнота, точность, *AvP*, F-мера

Спасибо за внимание!