КОЛЛОКАЦИИ

Саша Ершова asershova@edu.hse.ru

August 7, 2017

ЛОШ-17, Компьютерная лингвистика

Что такое коллокации?

Коллокации — это устойчивые словосочетания.

ПРИМЕРЫ КОЛЛОКАЦИЙ

- бить баклуши
- ставить условия
- высокая температура
- лучший друг

Подходы к выделению коллокаций

1. Ориентация на значение (семантический подход)
2. Ориентация на статистику (статистический подход)

Семантический подход

СЕМАНТИЧЕСКИЙ ПОДХОД

Идея семантического подхода в том, что коллокации обладают некоторой семантической целостностью.

Что из этого следует?

ПРИЗНАКИ КОЛЛОКАЦИИ

1. Значение некомпозиционально (его нельзя вывести из значений отдельных слов в коллокации).

   сухое вино $\neq$ вино, которое высохло

2. Нельзя (или почти нельзя) подставить (квази)синоним, когипоним[1] и т.п. вместо одного из коллокантов.

   бить баклуши vs. *избивать баклуши

3. Части имеют «фиксированную» позицию.

   вокруг да около vs. *около да вокруг

4. Зависят от области употребления (например, терминологические словосочетания)

   суд постановил удовлетворить ходатайство

[1] Когипонимы — это, например, ”шахматы” и ”шашки”.

СЕМАНТИЧЕСКИЙ ПОДХОД

При ориентации на значение мы считаем пары типа ”месяц-год” не коллокацией, а просто устойчивой встречаемостью.

Статистический подход

Статистический подход

Идея статистического подхода в том, что мы считаем коллокациями N-граммы, которые встречаются в текстах вместе **чаще, чем случайно**.

Статистический подход

Идея статистического подхода в том, что мы считаем коллокациями N-граммы, которые встречаются в текстах вместе **чаще, чем случайно**.

Такой подход позволяет нам считать коллокациями гораздо большее количество словосочетаний.

что мы ищем?

1. Смотрим на сочетания лексем или словоформ?
2. Могут ли быть другие слова между членами коллокации?
3. А знаки препинания?

Коллокации vs. коллигации

Сочетание лексем — коллокация.

Сочетание словоформ — коллигация.

ОКНО ПОИСКА

Допустим, есть коллокация **принимать лекарство**.

В тексте она может встречаться в вариантах:

- **принимать лекарство**
- **принимать** горькое **лекарство**
- **принимать** назначенное врачом **лекарство**
- **принимать** горькое назначенное врачом **лекарство**
- и т.д.

Значит, между коллокатами могут быть другие слова.

Окно поиска

Окно поиска: в пределах скольки слов ищем коллокацию?

d — расстояние между словами

- **принимать лекарство** (d=0)
- **принимать** горькое **лекарство** (d=1)
- **принимать** назначенное врачом **лекарство** (d=2)
- **принимать** горькое назначенное врачом **лекарство** (d=3)
- **лекарство**, которое тебе надо **принимать** (d=-3)

ОКНО ПОИСКА

Выбираем размер окна:

- ±1: фразеологизмы, составные лексические единицы

  бить баклуши, НИУ ВШЭ, железная дорога
- ±5: устойчивые конструкции

  принимать лекарство, бросать взгляды
- ±предложение: лексемы, относящиеся к одному семантическому полю

  врач — больница — медсестра

Спасибо за внимание!