Лингвистика для NLP

Ася Ройтберг, asya.roytberg@bigdatateam.org
ML Instructor at BigData Team, http://bigdatateam.org/
https://www.facebook.com/bigdatateam/

16.12.2020

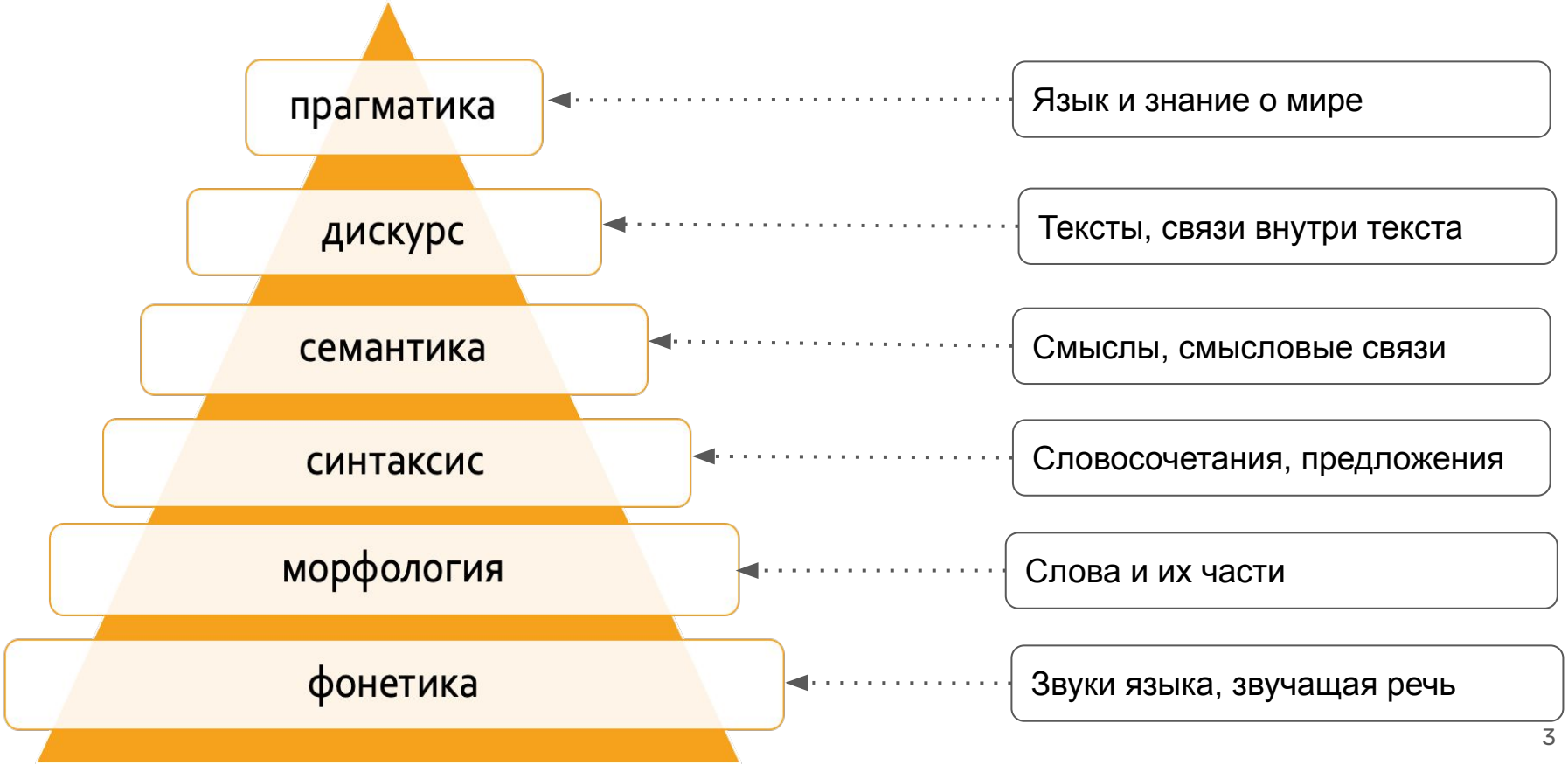
MEGAFON

Зачем нам лингвистика?

- Читать SOTA статьи
- Улучшение моделей
- Извлечение “умных фичей”

BIGDATA TEAM

Уровни языка



Морфология

Что такое слово?

Токенизация, стемминг, лемматизация

BIGDATA TEAM

Одно слово или разные?

1

Написаны через пробел = разные слова?

Не присесть ли тебе на диван-кровать?

2

Формы одного слова или разные слова

прыгал
прыгающий
прыгнувший
попрыгунчик
прыгнул

BIGDATA TEAM

Разбиение на слова

- дефис

  счет-фактура, сильно-сильно, English-speaking, easy-to-remember, well-known vs specially disigned

- апостроф

  isn’t - cannot, 's,

- слитные предлоги в иврите (, לארץ , ץ בער מארץ)
- Word2Vec, TF\IDF, P.S. и т.д.

Токенизация

1

Не просто разбиение по пробелам

2

Сложные случаи – вопрос договоренностей

BIGDATA TEAM

Слово, словоформа

прыгал
прыгающий
прыгнувший
попрыгунчик
прыгнул

вода
воды
водой
воде

water
waters
water

В каком языке у слов больше словоформ?

английский vs. русский

склонение и спряжение

Стемминг

Какие есть проблемы?

BIGDATA TEAM

Стемминг: проблемы

1. Чередования букв: кошк~~а~~ - кош~~ек~~ круг - круж~~ок~~
2. Нерегулярные формы: ребен~~ок~~ – дет~~и~~ бы~~л~~ – ест~~ь~~
3. Короткие слова: е~~л~~, бы~~л~~
4. Не различает, где аффикс, а где часть слова:

   рук~~и~~ vs. пр~~и~~ ele~~ment~~ vs. measure~~ment~~

Стемминг: преимущества

СКОРОСТЬ

BIGDATA TEAM

Лемматизация: к чему сводим?

- Для существительных – номинатив (Что это?)
- Для остальных частей речи - словарная форма

Словарная форма – вопрос традиции.

Вот так выглядит “словарная форма” латинского глагола

facio, feci, factum, ere (делать)

BIGDATA TEAM

Лемматизация: проблемы

1

Омонимия

зАмок - замОк

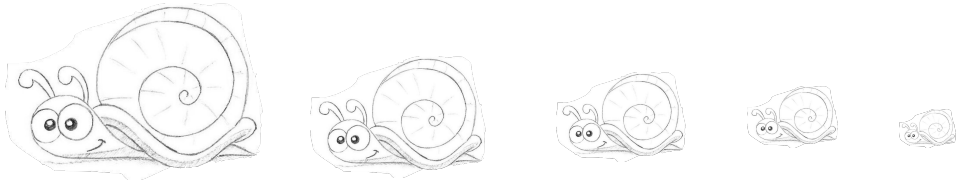
пятнистая рысь - бежал рысью

три – три | тереть

села - селО | сесть

2

Скорость



Части речи (POS)

- Список частей речи
- Частеречная омонимия
- Знаменательные\служебные части речи
- Нас не интересуют сложные случаи

BIGDATA TEAM

Части речи (POS)

UPOS – набор универсальных тегов.

XPOS – специфические наборы тегов для отдельных языков

- StanfordNLP
- Stanza

```
word: Barack    upos: PROPN    xpos: NNP
word: Obama     upos: PROPN    xpos: NNP
word: was       upos: AUX      xpos: VBD
word: born      upos: VERB     xpos: VBN
word: in        upos: ADP      xpos: IN
word: Hawaii    upos: PROPN    xpos: NNP
word: .         upos: PUNCT    xpos: .
```

BIGDATA TEAM

Universal Dependencies POS

https://universaldependencies.org

Open class words	Closed class words	Other
ADJ: adjective (прилагательное) ADV: adverb (наречие) INTJ: interjection (междометие) NOUN: noun (существительное) PROPN: proper noun (имя собственное) VERB: verb (глагол)	ADP: adposition (пред\послелоги) AUX: auxiliary (вспомогательный глагол) CCONJ: coordinating conjunction (сочинительный союз) DET: determiner (детерминатив) NUM: numeral (числительные) PART: particle (частица) SCONJ: subordinating conjunction (подчинительный союз)	PUNCT: punctuation (пунктуация) SYM: symbol (символы) X: other (другое)

BIGDATA TEAM

Universal Dependencies POS

UD POS	He UD POS
SpaCy	nltk
Flair	
UDPipe	
Stanza	Stanza

Вспоминаем про HMM

Три задачи HMM

- Правдоподобие
- Декодирование
- Обучение

BIGDATA TEAM

Три задачи HMM: декодирование

Дано:

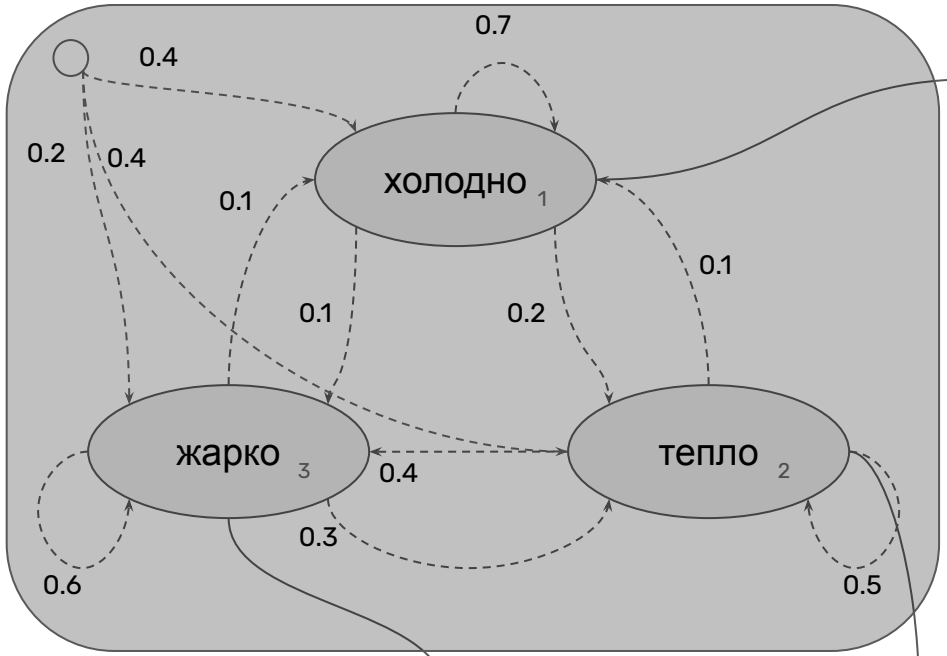
модель $\lambda = (\{a_{ij}\}, \{b_j(k)\}, \{\pi_i\})$;

наблюдаемая последовательность $o_1 o_2 \ldots o_T$

Необходимо найти наиболее вероятную последовательность скрытых состояний

$\mathit{argmax}\ \mathrm{P}[q_1 q_2 \ldots, q_T \mid \lambda, o_1 o_2 \ldots o_T]$

Три задачи HMM: декодирование



$b_1(0) = P[0|холодно] = 0.1$
$b_1(1) = P[1|холодно] = 0.3$
$b_1(2) = P[2|холодно] = 0.4$
$b_1(3) = P[3|холодно] = 0.2$

Какая последовательность погодных условий наиболее вероятна, при условии, что Петя съел:
1 июня - 2 мороженых
2 июня - 1 мороженое
3 июня - 0 мороженых?

Перебирать все варианты вычислительно долго!

$b_3(0) = P[0|жарко] = 0$
$b_3(1) = P[1|жарко] = 0.1$
$b_3(2) = P[2|жарко] = 0.3$
$b_3(3) = P[3|жарко] = 0.6$

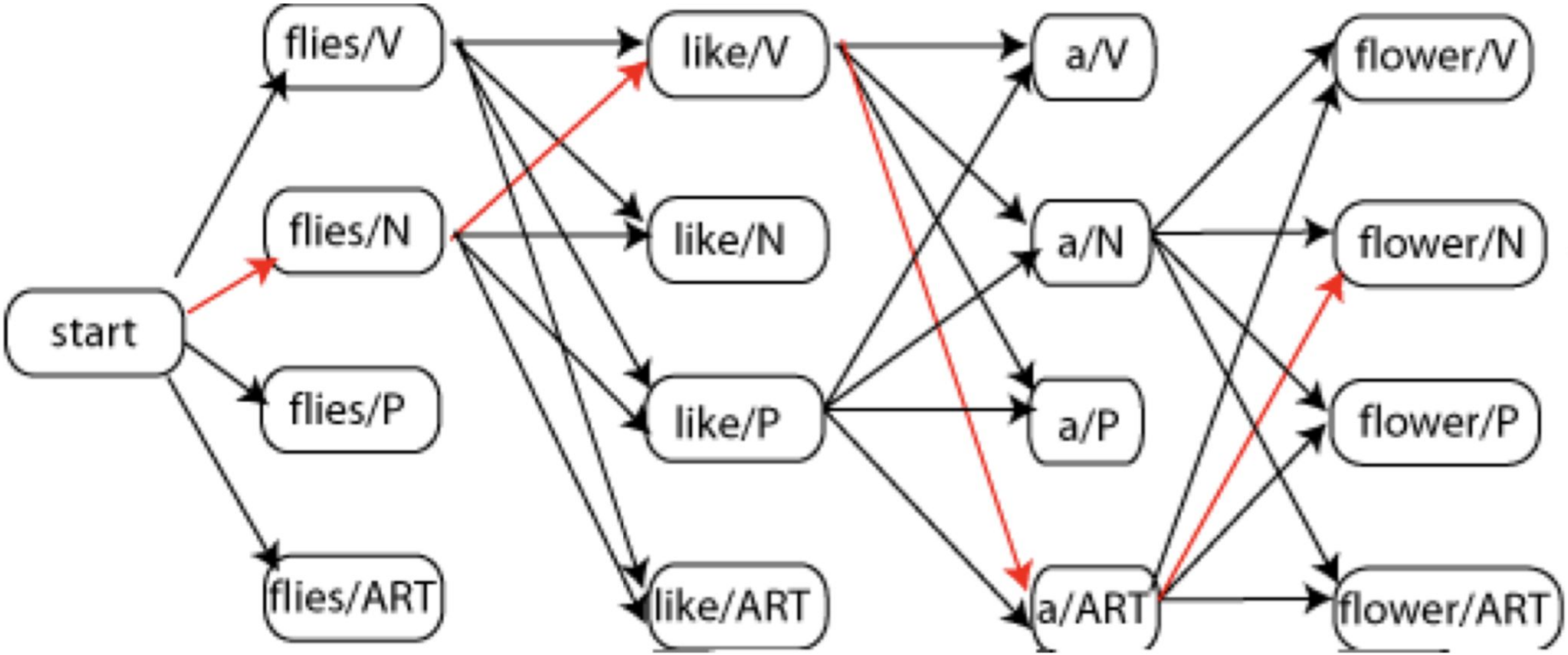
$b_2(0) = P[0|тепло] = 0.1$
$b_2(1) = P[1|тепло] = 0.3$
$b_2(2) = P[2|тепло] = 0.4$
$b_2(3) = P[3|тепло] = 0.2$

POS-tagging и HMM

Что в случае задачи разметки частей речи скрытые состояния, а что наблюдаемые?

BIGDATA TEAM

POS-tagging и HMM



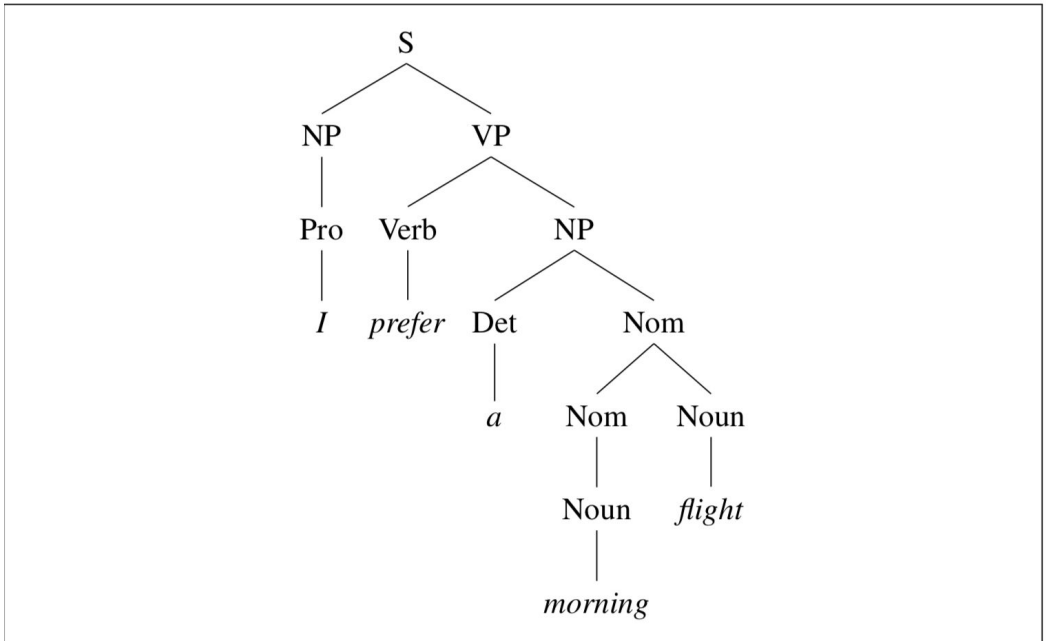
Морфология: summary

- Морфология занимается словами и их частями
- Не всегда ясно, где разные слова, а где одно слово
- У слов есть части речи
- Можно сделать POS таггер на основе HMM
- UD - популярная мультиязычная POS разметка
- UD - не единственный способ размечать POS

Составляющие и зависимости

Грамматика составляющих

Constituency Grammar



S- sentence

NP - Nominal Phrase

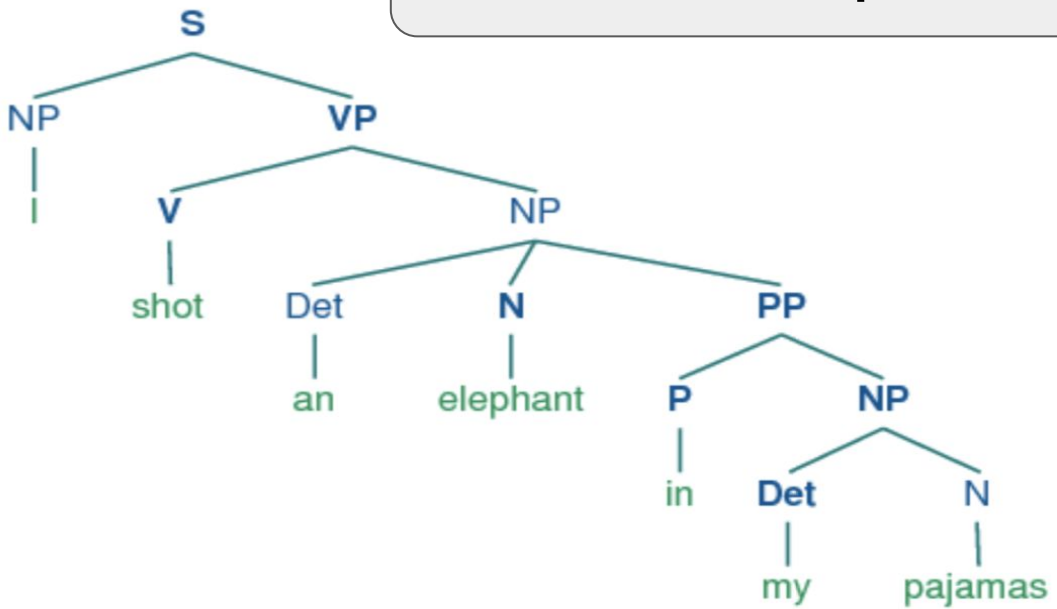
VP - Verbal Phrase

PP - Prepositional Phrase

Грамматика составляющих

Неоднозначность (Ambiguity)

I shot an elephant in my pajamas.

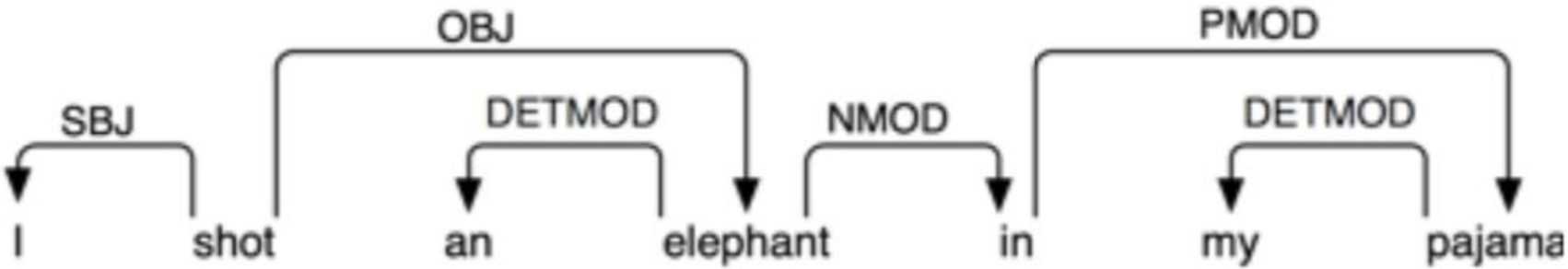


Who wears the pajamas?

BIGDATA TEAM

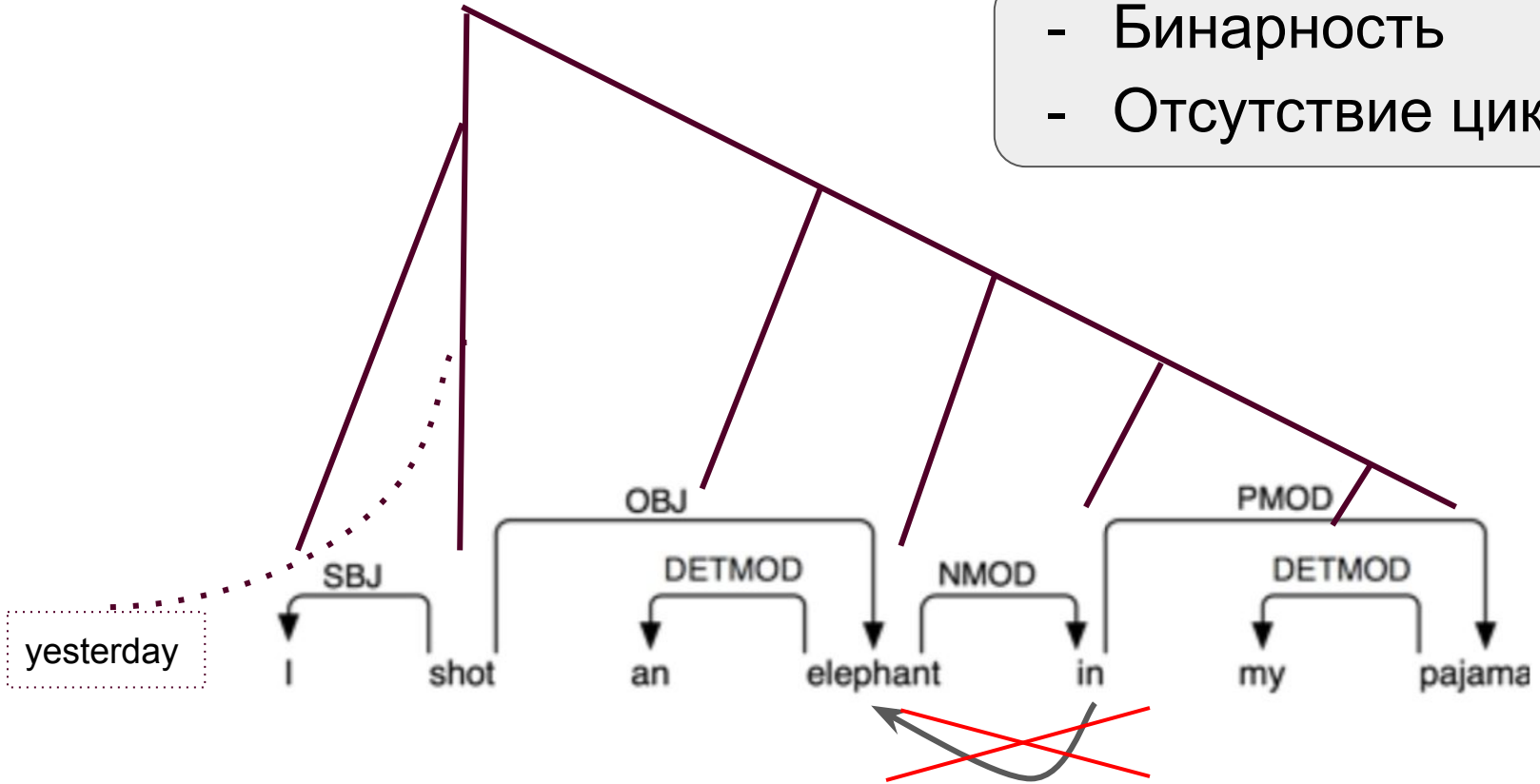
Грамматика зависимостей

Dependency Grammar



Грамматика зависимостей

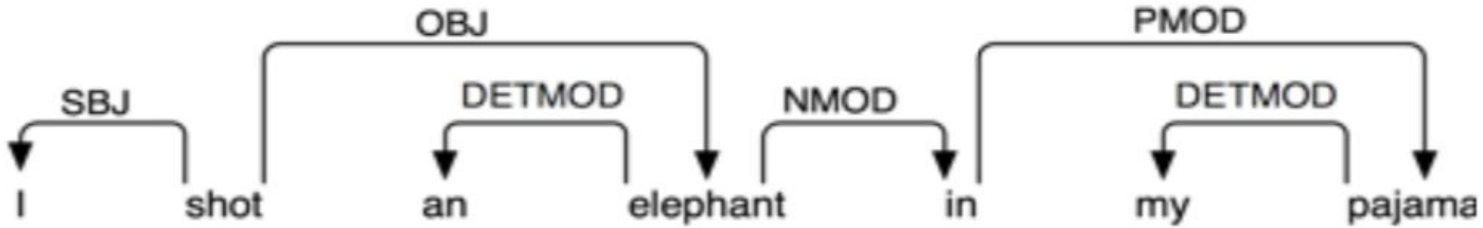
- Бинарность
- Отсутствие циклов



Universal Dependencies

List of UD relationships

Subject – Predicate – Object - центр большинства предложений



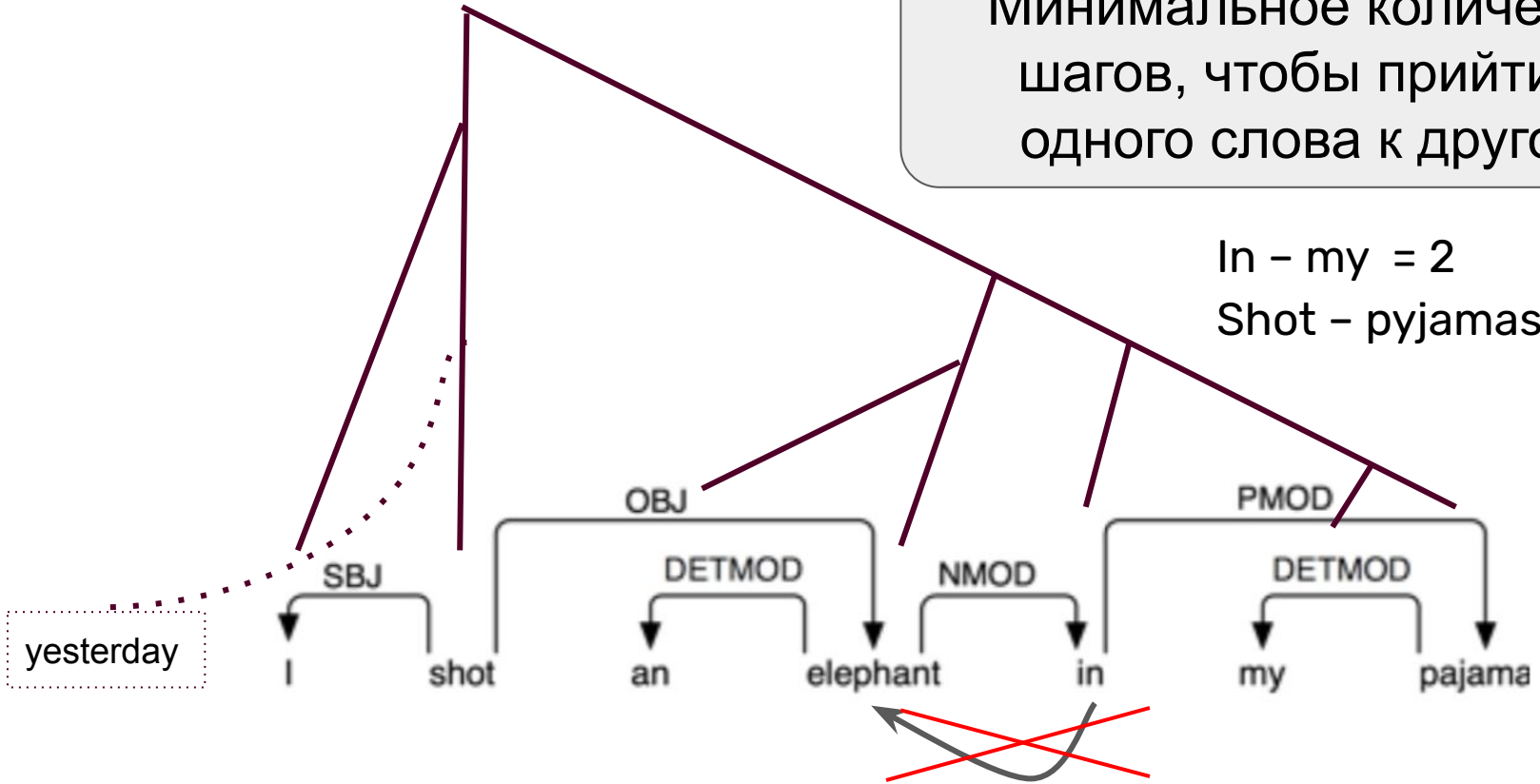
BIGDATA TEAM

Синтаксическое расстояние

Минимальное количество шагов, чтобы прийти от одного слова к другому

In – my = 2

Shot – pyjamas = ?



Наименьший общий предок (LSA)

Расстояние между вершинами a и b =
расстояние от a до LCA + расстояние от b до LCA

```
Процедура LCA(u, v):
    h1 := depth(u)          // depth(x) = глубина вершины x
    h2 := depth(v)

    while h1 ≠ h2:
       if h1 > h2:
          u := parent(u)
          h1 := h1 - 1
       else:
          v := parent(v)
          h2 := h2 - 1

    while u ≠ v:
       u := parent(u)       // parent(x) = непосредственный предок вершины x
       v := parent(v)

    return u
```

Синтаксис: summary

- У предложений есть очень четкая структура
- Структура предложения различна в разных языках
- Есть 2 основных формализма :
  a. Грамматика составляющих
  b. Грамматика зависимостей
- Понятие синтаксического расстояния
- Синтаксическое расстояние не прямо зависит от линейного расстояния

Слова, смыслы и связи

Дистрибутивная семантика

- Лексическая единица описывается вектором
- Измерения (или компоненты) – другие слова лексикона
- Значения этих компонентов – частота совместной встречаемости “нашей” единицы с другими словами в данном корпусе

Дистрибутивная семантика

Проблема: многозначность

Гуляли по стрелка Васильевского острова.

Повесили указатели в виде стрелок.

Соедини главное и зависимое слова стрелкой.

BIGDATA TEAM

“Атомы смысла”

Концепция

См. напр.
- Модель Смысл ↔Текст
- Теория семантических примитивов

Есть некоторый ограниченный набор смыслов и отношений между этими смыслами, из которых можно собрать значение любого слова.

Очень упрощенно:
Блюдце - тарелка с размером меньше среднего
Бежать - идти быстро

Смысл суммы ≠ сумме смыслов

- Устойчивые выражения:

  красна девица, rain cats and dogs

- Эвфемизмы: белая полярная лиса

- Коллокации:

  крепкий чай – *сильный чай

  принять душ – *получить душ

  powerful engine – *strong engine

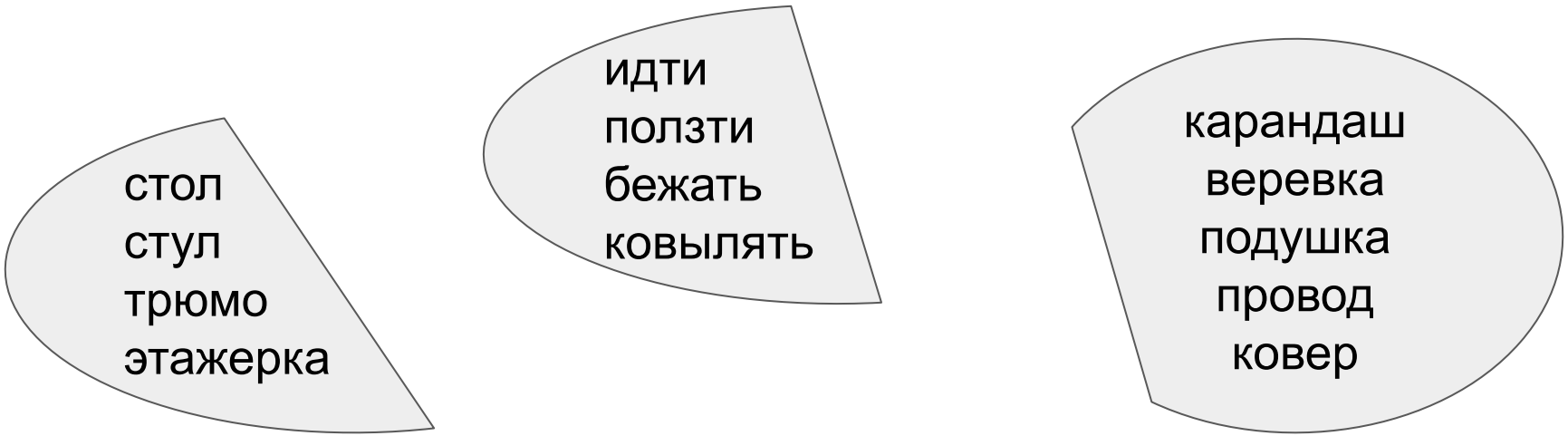
P(AB) > P(A)*P(B)

и многое другое

BIGDATA TEAM

Семантические классы слов

- Группы слов одной части речис общим компонентом значения



BIGDATA TEAM

Семантическое поле

Слова разных частей речи, объединенные общим компонентом значения

BIGDATA TEAM

Именованные сущности

Традиционное понимание: ~ имена собственные

- Персоны
- Названия организация
- Названия географических объектов
- Названия белков

……..

BIGDATA TEAM

Именованные сущности Custom edition

Современное понимание:

некоторый класс существительных + даты

- Названия политических должностей
- Термины родства
- Тип организации
- и т.д.

Семантические отношения

Основные отношения семантического поля

- Синонимия (милый – симпатичный)
- Антонимия (входить - выходить)
- Часть-целое (ветка – дерево, окно – дом)
- Гипонимия (~ включение в класс: человек - живые существа)
  - *человек* – гипоним, *живые существа* – гипероним
- Несовместимость (мать – отец)

BIGDATA TEAM

Семантические отношения

Любые отношения, которые можно формализовать

- Person **родиться в** Location
- Person **быть сотрудником** Organization
- Organization **быть акционером** Organization
- Location **находиться в** Location
- Person **быть родителем** Person

BIGDATA TEAM

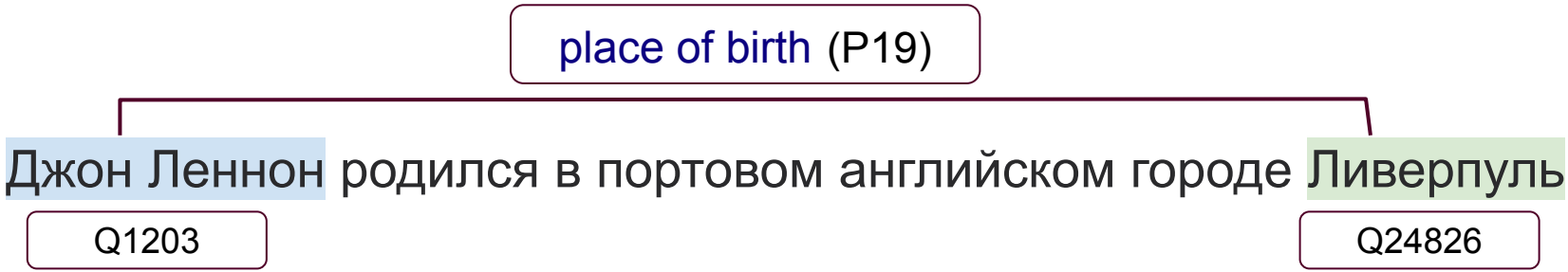
Онтологии и базы знаний

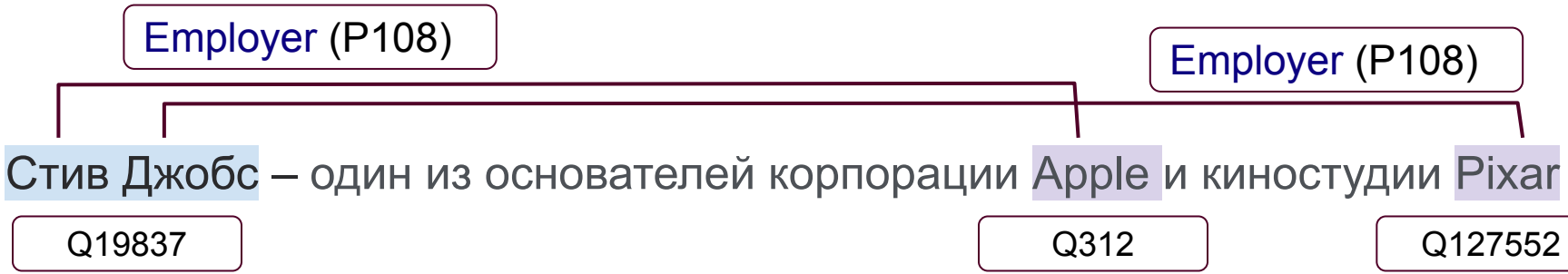
Понятия	Airbus	Bill Gates
Атрибуты	aerospace manufacturer	human
Отношения	member of -→ Linux Foundation	Spouse → Melinda Gates

BIGDATA TEAM

Извлечение отношений

Дистанционная разметка и Wikidata





Извлечение отношений

Проблемы. False positive

Джон Леннон уехал из Ливерпуля в 1960-х.

Уехать в город из аэропорта Ливерпуля «Джон Леннон» (LPL).

Разругавшись с Apple, Джобс создал фирму NeXT.

BIGDATA TEAM

Семантика: summary

- Компоненты значения (“атомы смыслы”)
- Многозначность и эмбеддинги
- Сумма смыслов ≠ смыслу суммы
- Семантические классы
- Семантические поля
- Именованные сущности
- Семантические отношения
- Использования баз знаний для извлечения отношений

Тексты и отношения в них

Референция и кореференция

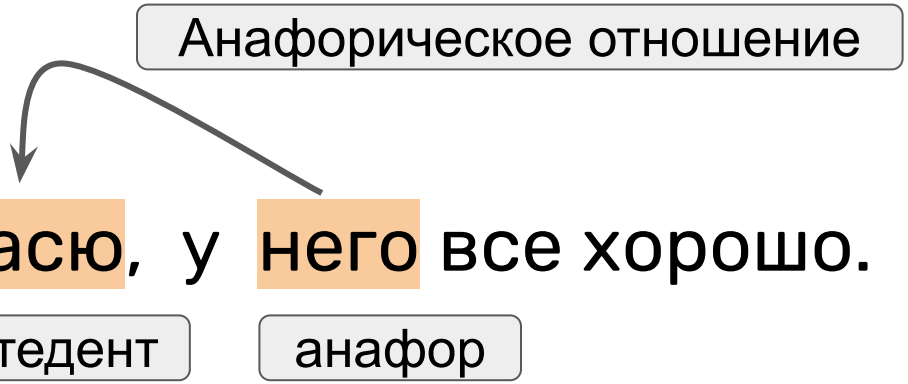
Референт - объект реального мира, называемый выражением

Кореференция - ситуация, когда несколько выражений в тексте, отсылают к одному объекту

> Дом на Бейкер стрит – легендарное жилище Шерлока Холмса. Именно в нем были найдены разгадки к самым запутанным делам.

Анафора (+ Катафора)

Для понимания значения элемента необходима отсылка к другому элементу из текста.



Синглтон - объект, который только один раз упоминается в тексте

Разная анафора

- Не только местоимения

IPO QIWI состоялось сегодня на бирже Nasdaq. Торги расписками компании стартовали в 17:00.

- Не только кореференция

В Уругвае прошли выборы. Новый президент – Луис Альберто Локалле

Связность текста

Coherence

- Coherence relations: отношения причины, контраста ...
- Entity-based coherence: есть ядерные (salient) слова вокруг которых строится текст
- Topic-based coherence: предложения об одной теме обычно оказываются рядом
- Global coherence: у каждого типа текста есть ожидаемая структура

Чем тексты могут отличаться?

- Измеряемые параметры: длина текста, длины предложений и т. п.
- Сбалансированность лексики
- Структура текста
- Стиль текста

BIGDATA TEAM

Как выбрать правильный корпус?

Главный вопрос всего:

С какими текстами предстоит работать в “боевых условиях” ?

Дискурс: summary

- Текст состоит из единиц, связанных отношениями
- Понятие связности текста
- Понятие структуры текста
- Нужно внимательно выбирать тексты, на которых вы учитесь

Thank you! Any questions?

Feedback: https://forms.gle/ZvjrXYt9UKD8BkK48

Ася Ройтберг, asya.roytberg@bigdatateam.org LinkedIn
ML Instructor at BigData Team, http://bigdatateam.org/

https://www.facebook.com/bigdatateam/