Синтаксическая разметка текста

Математические методы анализа текстов осень 2021

Попов Артём Сергеевич

9 ноября 2021

Представление текста

Какие представления текста мы уже рассмотрели:

Представление текста

Какие представления текста мы уже рассмотрели:

- мультимножество слов / n-грамм
  (мешок слов, tf-idf, агрегация эмбеддингов слов)

- последовательность слов / символов / буквенных n-грамм
  (RNN, CNN, Transformer)

Пусть текст — направленный граф $G = (V, E)$:

- V — слова предложения (+ вспомогательные сущности)
- $E \subseteq V \times V \times C$ — связи между словами в предложении
- C — типы связей

Что мы хотим от представления графом?

Хотим одинаковые представления для амонимичных по смыслу последовательностей, без потери связей между словами:

- В телеграме появились рекламные объявления.
- Рекламные объявления появились в телеграме.
- В телеграме рекламные объявления появились.

	представления одинаковы?	сохранили связь между словами
мультмножество	да	нет
последовательность	нет	да
граф (теория)	да	да

Что мы хотим от представления графом?

После лемматизации разные последовательности токенов могут стать одинаковыми:

- Заблокирована моя карта $\rightarrow$ заблокировать мой карта
- Заблокируйте мою карту $\rightarrow$ заблокировать мой карта

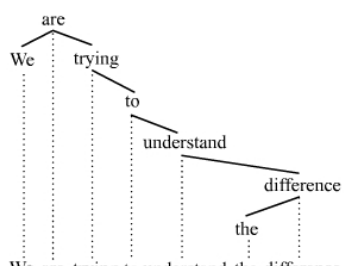
Хотим решить проблему через графовое представление:

1. Строим граф по исходному предложению до лемматизации
2. Лематизируем слова в вершинах графа
3. Хотим уметь восстанавливать разницу в предложениях по типам связей между словами

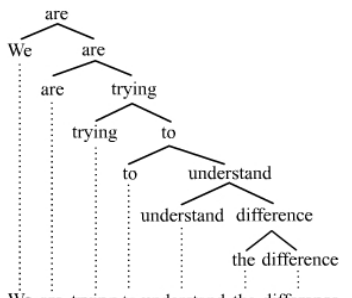
Синтаксический разбор предложения и его виды

Синтаксический разбор — естественный способ построения графа для текста. Две основных парадигмы синтаксиса:

- Грамматика составляющих (constituency, phrases)
- Грамматика зависимостей (dependency)



Dependency



Constituency (BPS)

Грамматика составляющих

- S — линейно упорядоченное множество слов.
- Система составляющих на S — множество C отрезков S.
- C содержит S и каждое слово, входящее в S.
- Любые два отрезка, входящие в C, либо не пересекаются, либо один из них содержится в другом.
- Элементы множества C — составляющие.

Пример разбора на составляющие

S — исходное предложение

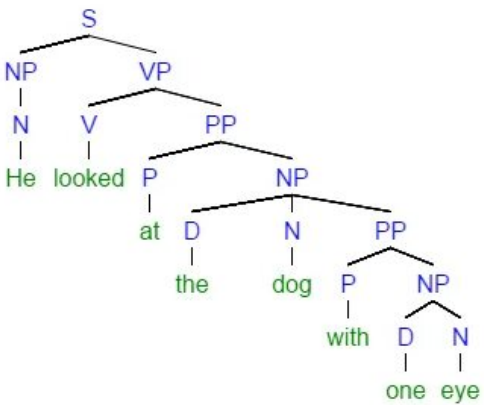
VP — глагольная группа, verb phrase (глагол + зависимые)

NP — именная группа, noun phrase (существительное)

PP — предложная группа, prepositional phrase

AP — группа прилагательного, adjective phrase

D (Det) — детерминативы (артикли и т.п.)



Применение составляющих для аугментации текста

1. Составляющие можно перемещать в рамках предложений:
   John talked [to the children] [about rules].
   John talked [about rules] [to the children].
   (запрещено) *John talked rules to the children about.*

2. Составляющие можно заменять на похожие:
   I sat [on the box / on top of the box / in front of you].
   Важно. Строить парафраз для короткого текста проще чем для длинного.

Резюме по составляющим

- Популярный подход в лингвистике
- Лучше описан в «учебной» литературе[1]
- Плохо применим для языков, в которых может быть произвольный порядок слов (например, для русского языка)
- Часто описывается контекстно-свободными языками
- Для построения разбора может использоваться алгоритм CYK (Cocke-Younger-Kasami)

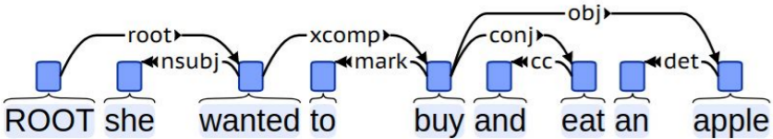
[1] Jurafsky and Martin, Speech and Language Processing (три главы)

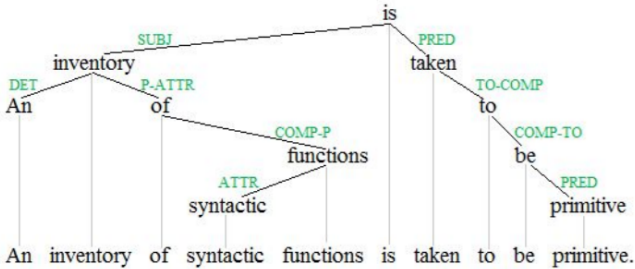
Грамматика зависимостей

Дерево зависимостей — направленный граф, такой что:

- Граф является деревом
- Вершины графа — слова и [ROOT]
- В каждую вершину, кроме [ROOT], входит одно ребро
- В [ROOT] не входит ни одно ребро
- Рёбра графа описывают зависимость одного слова от другого.
- Рёбра могут иметь «тип» связи.

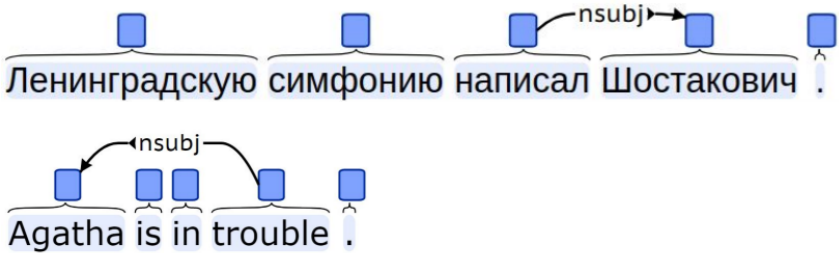
Пример разбора на зависимости





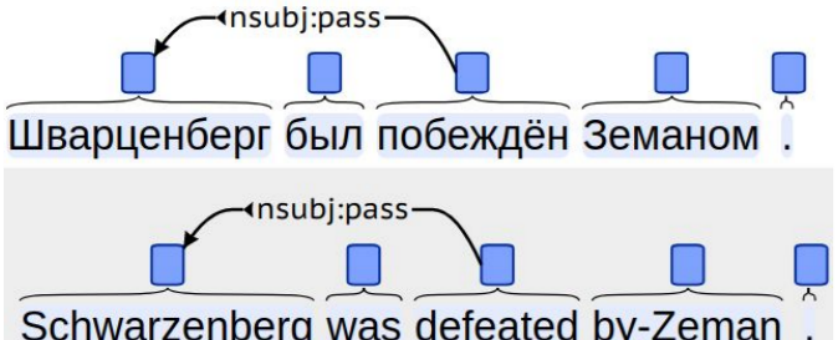
Пример связей: nsubj

Кто совершает действие? Кто субъект действия?



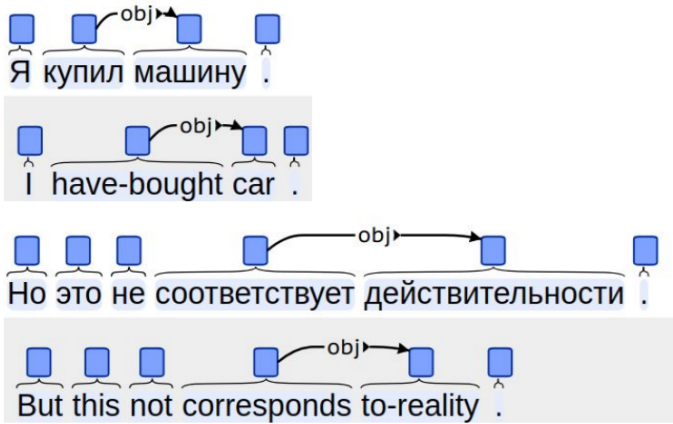
Пример связей: nsubj:pass

Кто совершает действие? Кто субъект действия?
+ пассивный залог



Пример связей: obj

Над кем совершают действие? Кто объект действия?



Пример связи: det

Связь объекта и указывающего местоимения

Свойство проективности

Предложение называется проективным, если:

1. ни одна из стрелок не пересекает другую стрелку
2. нет стрелок накрывающих корневую

а) проективное

б) непроективное — нарушено 1

в) непроективное — нарушено 2

Зачем нужна проективность?

- Слова близкие друг к другу синтаксически, обычно близки по положению в тексте.
- Проективность — «естественность» синтаксической структуры.
- Непроективность иногда несёт особую информацию, которая не будет содержаться в соответствующем проективном аналоге:
  Кубок все выиграть мечтают / Все мечтают выиграть кубок
  Очень они хорошие были / Они были очень хорошие

Таким образом, последовательность нельзя полноценно заменить графом...

Неоднозначность синтаксического разбора

Если для предложения нет контекста, синтаксический разбор может быть неоднозначен!

Он сам увидел их семью

Он сам увидел их семью своими глазами

Он сам увидел их семью (маму, папу и дочку)

Эти типы стали есть в цехе

Эти странные люди решили начали есть в цехе

Эти типы стали (углеродистые) есть в цехе