

Computational Syntax

Денис Кирьянов, Sberdevices,

denkirjanov@gmail.com,

05/03/2020

О лекторе

- **Сбербанк**

SberDevices, NLP platform for goal-oriented chat-bots, OKKO smartbox and many more devices to come...

- **Angry Analytics**

мониторинг отзывов, сентимент, антиспам и др.

- **Double Data**

поиск по людям в соцсетях и скоринг профилей

- **Мое дело**

кастомный поисковик

- Филфак СПбГУ (бакалавриат)

- ФГН ВШЭ (магистратура, комплингвистика)

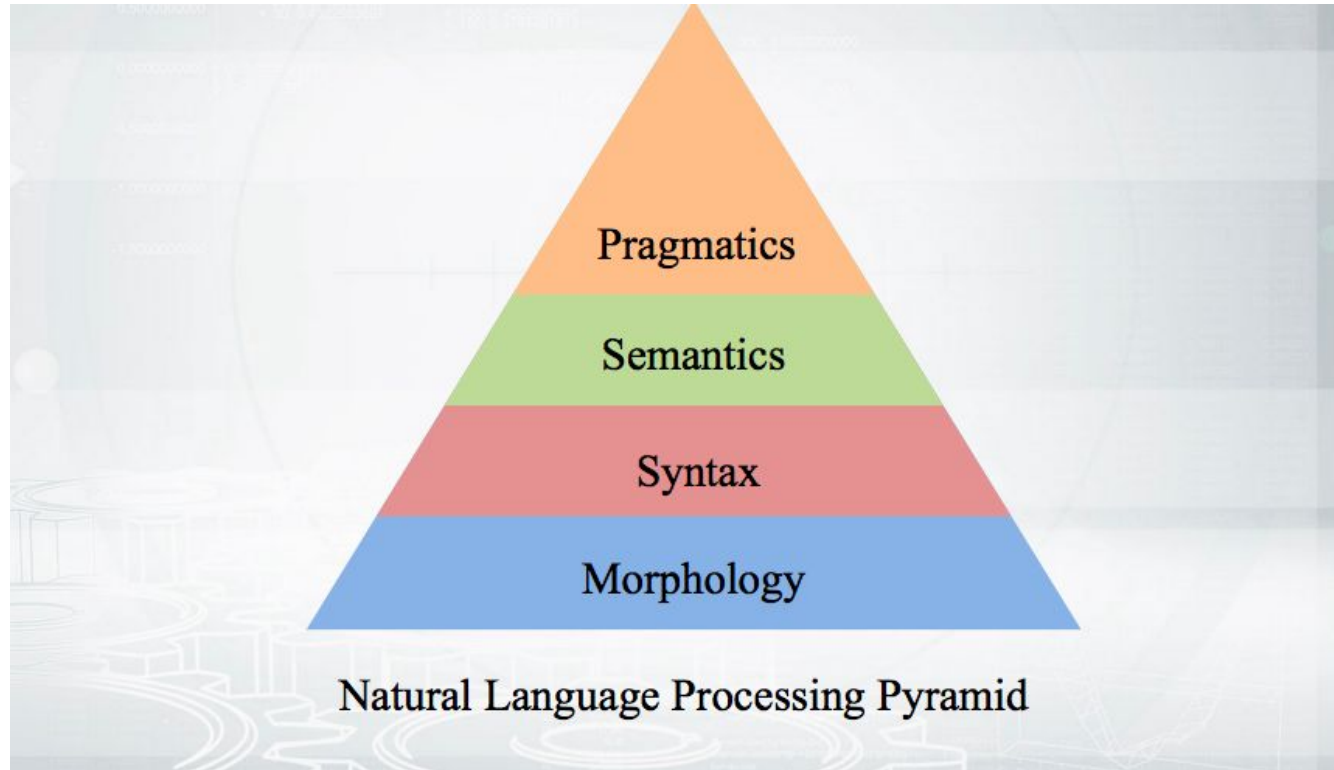
Disclaimers

- Практическое применение vs **теоретическое устройство парсеров**
- Ходите по [ссылкам!](#)
- Вопросы погромче :)

Содержание

1. **Что такое синтаксис и зачем он нужен**
2. Теоретические фреймворки
3. Dependency parsing
4. Метрики и соревнования
5. Инструменты
6. Varia

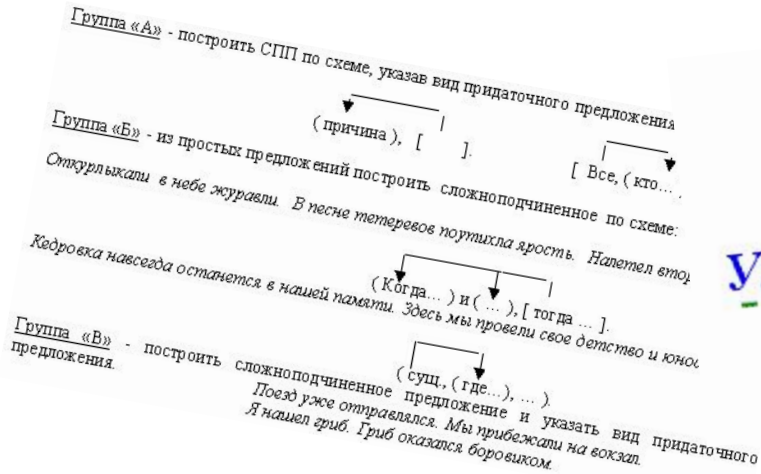
Где мы вообще?



Где мы вообще?

машинный анализ структуры текста, в т.ч. структуры предложения

Мы все это делали в школе, и машине это зачастую тоже под силу



Зачем это нужно: теория

- Человечество активно строит новые эмбединги
- Оценивается на датасетах, включающих в себя синтаксические задачи (см. датасет [GLUE](#))
- И пытается понять, какие знания о языке лежат внутри этих эмбедингов
- В связи с чем смотрит на синтаксические структуры, см. далее
- Активный рост статей о синтаксисе в последние пару лет

Зачем это нужно: практика

- «Банкомат съел карту» vs «карта съела банкомат»;
- Определение правильности грамматики фразы (при порождении речи);
- Question answering;
- Машинный перевод;
- Information extraction (*напомни мне сделать икс*);
- Синтаксическая роль токена как метрика его важности (подлежащее важнее определения), использование весов в классификаторе.

Содержание

1. Что такое синтаксис и зачем он нужен
- 2. Теоретические фреймворки**
3. Dependency parsing
4. Метрики и соревнования
5. Инструменты
6. Varia

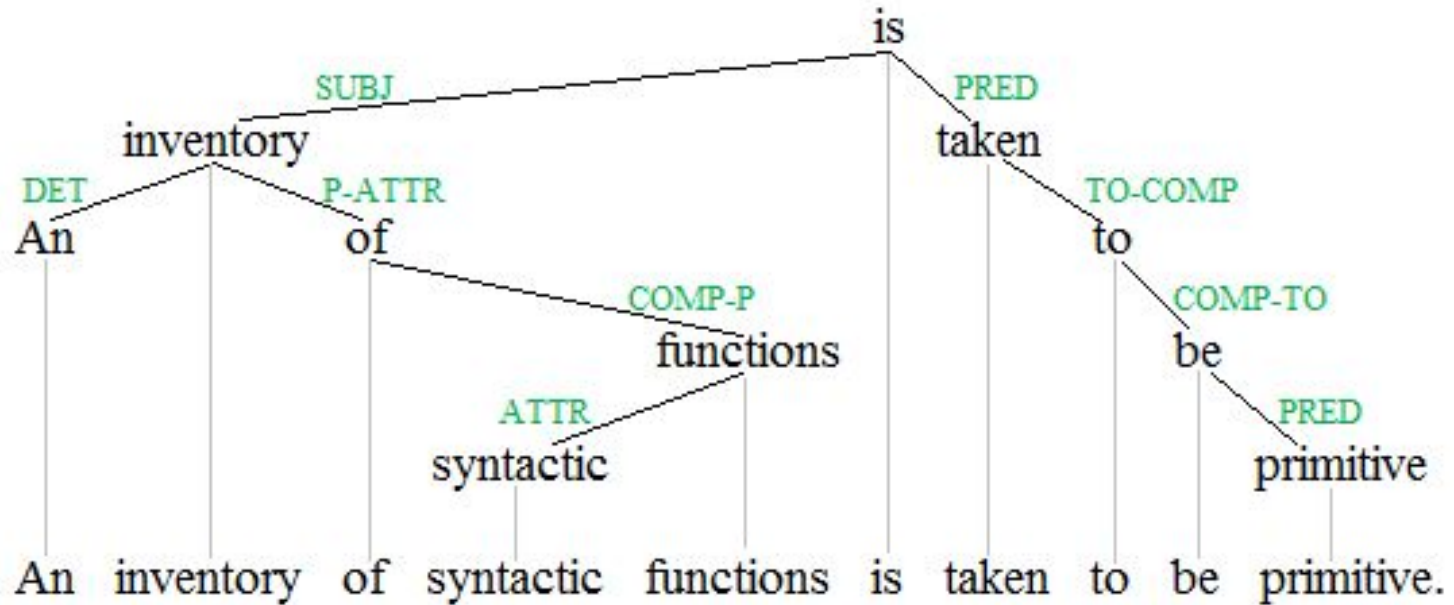
Грамматика непосредственно составляющих (constituency) и грамматика зависимостей (dependency)

●  Моя мама мыла грязную раму

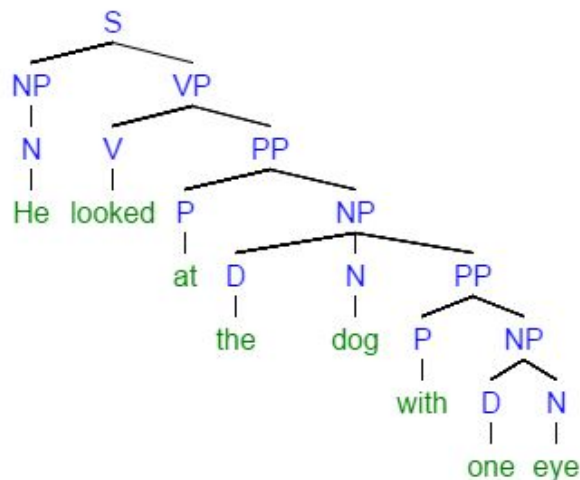
● [[моя мама] [мыла] [грязную раму]]

- В автоматическом парсинге для РЯ нет phrase-based парсеров (проблемы со “свободным” порядком слов);
- У многих других языков [получше](#).

Пример разбора: зависимости



Пример разбора: непосредственно составляющие



<https://i.imgur.com/ShMtNEy.png>

VP - глагольная группа, verb phrase

(грубо: состоит из глагола и зависимых;
но не подлежащее)

NP - именная группа, noun phrase

(грубо: вершина — существительное)

PP - предложная группа, prepositional phrase

AP - группа прилагательного, adjective phrase

D (Det) - детерминативы: артикли, указ., притяж.,
определительные местоимения, квантификаторы,
числительные, вопросительные слова

...

Phrases vs Dependencies

- Хорошо разработанные в лингвистике теории;
- Отчасти (!) формально (!) взаимозаменяемые — грубо говоря, обе основаны на правилах взаимодействия частей речи;
- Нет главной и нет вторичной (хотя ГЗ слегка устарела);
- Как водится, много проблем на периферии у обеих, см. ([Тестелец 2001](#));
- [SoTA работа](#) (2019) по автоматическому парсингу умеет за раз в оба фреймворка (правда, только на английском)

А в чем проблемы?

- Эллипсис: *Вася любит Машу, а Петя -- Катю*
- Сочинение: ***Петя и Маша*** пошли в магазин
- Предикаты подъема/контроля: *Маша хочет спать* — есть ли подлежащее у *спать*?
- Относительные местоимение: *Петя, который любит спать* — от кого зависит *который*?
- Синтаксическая омонимия: *Эти типы стали есть в цехе, подпись руководителя группы или командированного лица*

(Не)проективность (формализуя “нехорошее”)

Предложение называется **проективным**, если $\langle \dots \rangle$:

а) ни одна из стрелок не пересекает другую стрелку

б) никакая стрелка не накрывает корневую (\sim сказуемое \rightarrow подлежащее)

[[Тестелец 2001](#): 95]

Все мечтают выиграть кубок.

← Проективное

б.

Кубок все выиграть мечтают.

← НЕпроективное — нарушен принцип **пересечения**

в.

Кубок все мечтают выиграть.

← НЕпроективное (слабо проективное) — нарушен принцип **обрамления**

- Non-projectivity in natural languages

Language	Trees	Arcs
Arabic [Hajič et al. 2004]	11.2%	0.4%
Basque [Aduriz et al. 2003]	26.2%	2.9%
Czech [Hajič et al. 2001]	23.2%	1.9%
Danish [Kromann 2003]	15.6%	1.0%
Greek [Prokopidis et al. 2005]	20.3%	1.1%
Russian [Boguslavsky et al. 2000]	10.6%	0.9%
Slovene [Džeroski et al. 2006]	22.2%	1.9%
Turkish [Oflazer et al. 2003]	11.6%	1.5%

Table from invited talks by Prof. Joakim Niver: [Beyond MaltParser - Advances in Transition-Based Dependency Parsing](#)

Грамматика зависимостей

- Активное развитие в computational сфере;
- Лучше применима к парсингу русского языка;
- Всё не успеть за одну пару;
- Субъективный выбор лектора;
- Constituency parsing освещен в литературе, особенно см. [три главы учебника Журафского и Мартина](#).

Содержание

1. Что такое синтаксис и зачем он нужен
2. Теоретические фреймворки
3. **Dependency parsing**
4. Метрики и соревнования
5. Инструменты
6. Varia

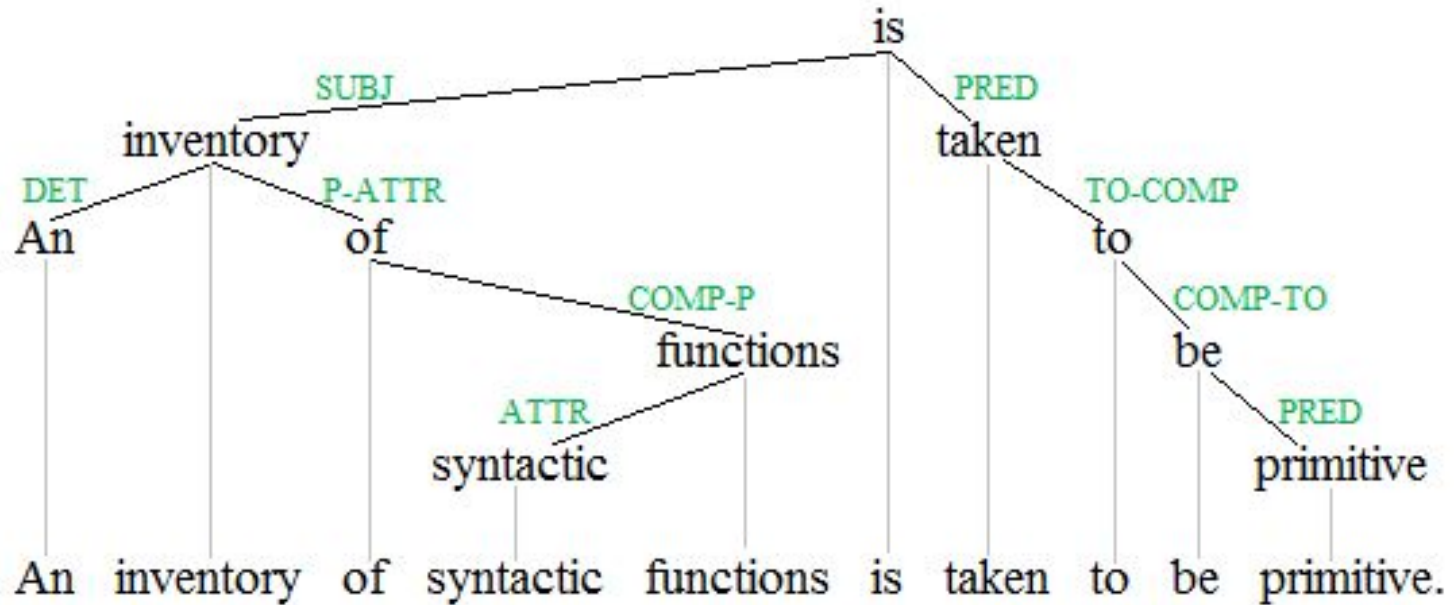
Дерево зависимостей [[Jurafsky & Martin 2019](#)]

(пока забудем про эллипсис и др.)

“Dependency tree is a directed graph that satisfies the following constraints:

- There is a single designated **root node that has no incoming arcs.**
- With the exception of the root node, **each vertex has exactly one incoming arc.**
- There is a **unique path from the root node to each vertex in V .**”

Пример разбора: зависимости



Dependency parsing: алгоритмы

Построение дерева зависимостей по предложению

Два основных подхода (supervised learning; еще есть их микс, но это довольно маргинально, хотя см. [[Li et al. 2020](#)])

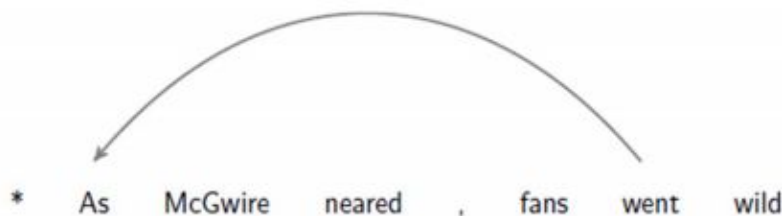
- **transition-based**

жадно набираем дерево (см. далее)

- **graph-based**

*ищем минимальное остовное дерево (minimum spanning tree; MST)
в полном графе всех возможных связей*

Features for one dependency

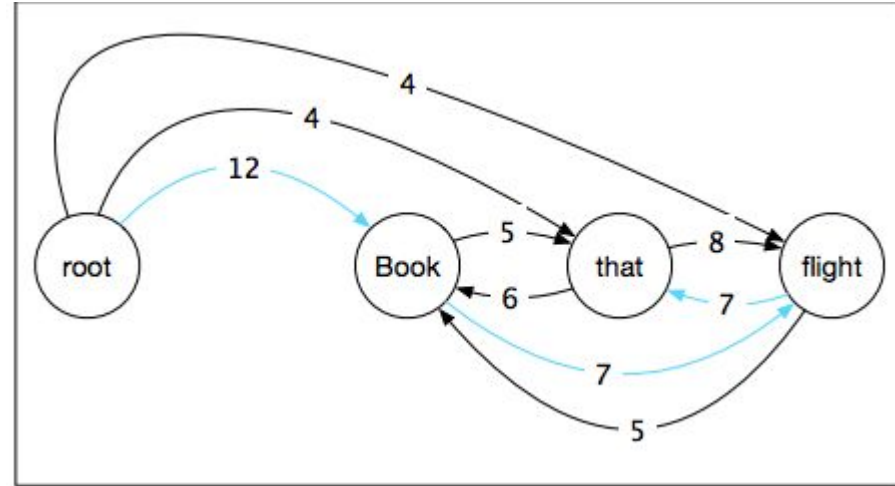


Example from slides of Rush and Petrov (2012)

[went]	[VBD]	[As]	[ADP]	[went]
[VERB]	[As]	[IN]	[went, VBD]	[As, ADP]
[went, As]	[VBD, ADP]	[went, VERB]	[As, IN]	[went, As]
[VERB, IN]	[VBD, As, ADP]	[went, As, ADP]	[went, VBD, ADP]	[went, VBD, As]
[ADJ, *, ADP]	[VBD, *, ADP]	[VBD, ADJ, ADP]	[VBD, ADJ, *]	[NNS, *, ADP]
[NNS, VBD, ADP]	[NNS, VBD, *]	[ADJ, ADP, NNP]	[VBD, ADP, NNP]	[VBD, ADJ, NNP]
[NNS, ADP, NNP]	[NNS, VBD, NNP]	[went, left, 5]	[VBD, left, 5]	[As, left, 5]
[ADP, left, 5]	[VERB, As, IN]	[went, As, IN]	[went, VERB, IN]	[went, VERB, As]
[JJ, *, IN]	[VERB, *, IN]	[VERB, JJ, IN]	[VERB, JJ, *]	[NOUN, *, IN]
[NOUN, VERB, IN]	[NOUN, VERB, *]	[JJ, IN, NOUN]	[VERB, IN, NOUN]	[VERB, JJ, NOUN]
[NOUN, IN, NOUN]	[NOUN, VERB, NOUN]	[went, left, 5]	[VERB, left, 5]	[As, left, 5]
[IN, left, 5]	[went, VBD, As, ADP]	[VBD, ADJ, *, ADP]	[NNS, VBD, *, ADP]	[VBD, ADJ, ADP, NNP]
[NNS, VBD, ADP, NNP]	[went, VBD, left, 5]	[As, ADP, left, 5]	[went, As, left, 5]	[VBD, ADP, left, 5]
[went, VERB, As, IN]	[VERB, JJ, *, IN]	[NOUN, VERB, *, IN]	[VERB, JJ, IN, NOUN]	[NOUN, VERB, IN, NOUN]
[went, VERB, left, 5]	[As, IN, left, 5]	[went, As, left, 5]	[VERB, IN, left, 5]	[VBD, As, ADP, left, 5]
[went, As, ADP, left, 5]	[went, VBD, ADP, left, 5]	[went, VBD, As, left, 5]	[ADJ, *, ADP, left, 5]	[VBD, *, ADP, left, 5]
[VBD, ADJ, ADP, left, 5]	[VBD, ADJ, *, left, 5]	[NNS, *, ADP, left, 5]	[NNS, VBD, ADP, left, 5]	[NNS, VBD, *, left, 5]
[ADJ, ADP, NNP, left, 5]	[VBD, ADP, NNP, left, 5]	[VBD, ADJ, NNP, left, 5]	[NNS, ADP, NNP, left, 5]	[NNS, VBD, NNP, left, 5]
[VERB, As, IN, left, 5]	[went, As, IN, left, 5]	[went, VERB, IN, left, 5]	[went, VERB, As, left, 5]	[JJ, *, IN, left, 5]
[VERB, *, IN, left, 5]	[VERB, JJ, IN, left, 5]	[VERB, JJ, *, left, 5]	[NOUN, *, IN, left, 5]	[NOUN, VERB, IN, left, 5]

Graph-based dependency parsing

- Изначально имеем полный орграф;
- Все ребра и все типы связей;
- При обучении учимся скорить связи;
- Фичи те же, что у transition-based;
- + могут быть фичи про порядок слов;
- Постпроцессинг: фильтр на циклы



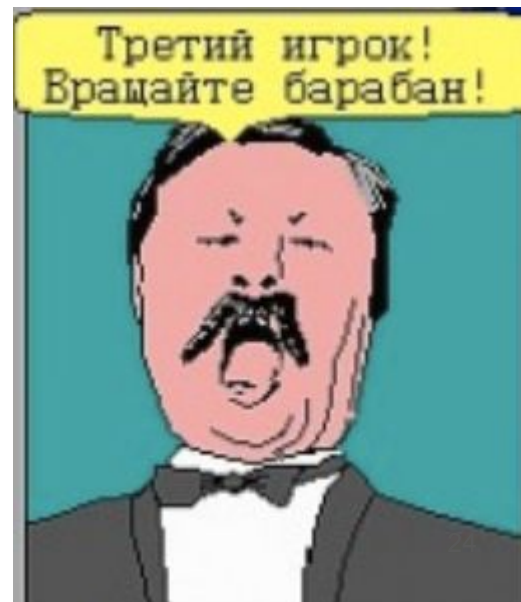
Проще справляться с непроективностью: без этого ограничения нам нужно меньше постпроцессинга, просто берем топ-кандидата, не отбирая именно топ-проективного

Transition-Based DP: интуиция

Представим, что нам Леонид Якубович открывает по одному слову, а мы строим разбор предложения на лету

Book...

Окей, что-то про книгу, книга может быть и подлежащим

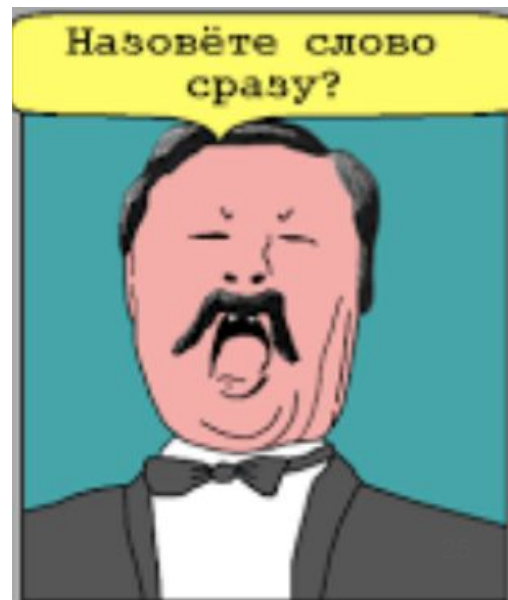


Transition-Based DP: интуиция

Представим, что нам Леонид Якубович открывает по одному слову, а мы строим разбор предложения на лету

Book me...

ан нет! “забронируй меня” или “забронируй мне”,
но **me** явно зависимое



Transition-Based DP: интуиция

Представим, что нам Леонид Якубович открывает по одному слову, а мы строим разбор предложения на лету

Book me the...

пока непонятно, но всё-таки это просьба
забронировать **что-то**



Transition-Based DP: интуиция

Представим, что нам Леонид Якубович открывает по одному слову, а мы строим разбор предложения на лету

Book me the morning...

“забронируй мне утро?” странновато, конечно (парсеру зависимостей это м.б. и не важно), но — **morning** может зависеть от **book**:
“забронируй мне утро у стоматолога”



Transition-based dependency parsing: интуиция

Представим, что нам Леонид Якубович открывает
по одному слову, а мы строим разбор предложения на лету

Book me the morning flight.

А вот теперь всё ясно!



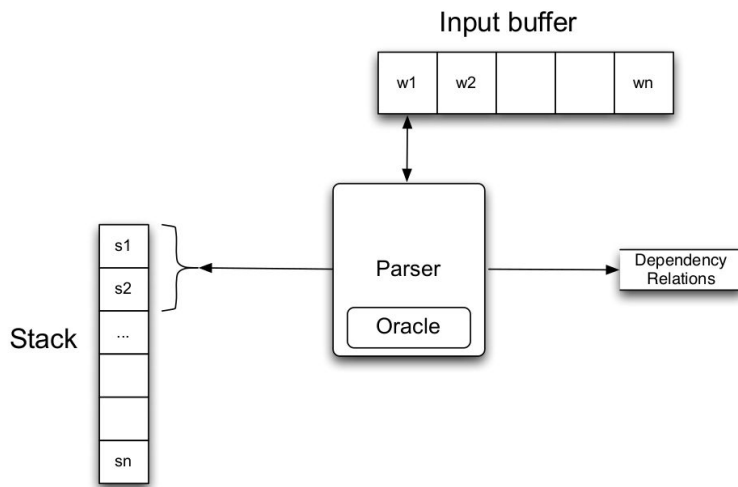
Transition-based (arc-standard) dependency parsing

Есть список токенов, стек (изначально содержит только root) и конфигурация (изначально пустая). Три дефолтных способа изменить конфигурацию:

- **LeftArc** [применим, если второй элемент стека не ROOT]
проводим зависимость между токеном на верхушке стека и вторым + выкидываем второй из стека
- **RightArc**
то же, но зависимость в другую сторону, и выкидываем верхушку стека
- **Shift**
переносим очередное слово из буфера в стек

Transition-based dependency parsing

Ключевое понятие: “**конфигурация**” = **состояние** процесса разбора:
входящие токены, верхушка стека и набор уже построенных отношений
(то, что мы “держали в уме”, когда играли; да, аналогия не вполне точна)



Потому и transition-based -- мы сейчас будем **переходить** из состояния в состояние системы по правилам

Псевдокод

function DEPENDENCYPARSE(*words*) **returns** dependency tree

state \leftarrow { [root], [*words*], [] } ; initial configuration

while *state* **not** final

t \leftarrow ORACLE(*state*) ; choose a transition operator to apply

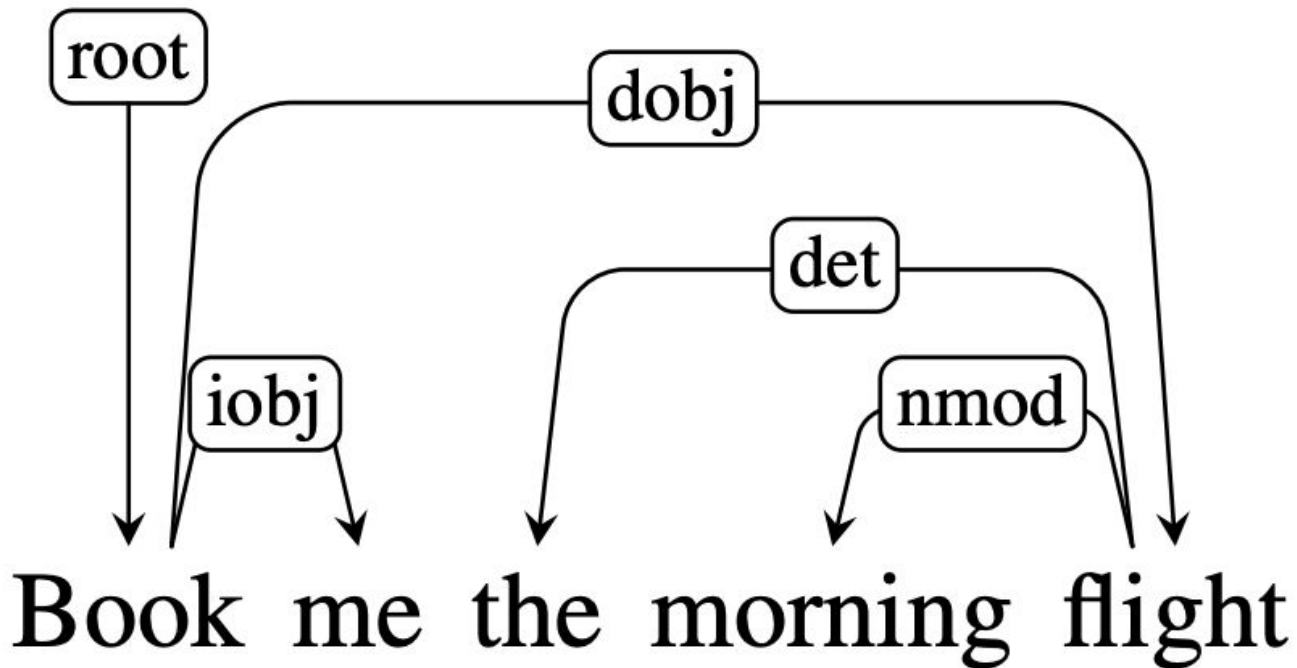
 state \leftarrow APPLY(*t*, *state*) ; apply it, creating a new state

return *state*

Пример работы

Step	Stack	Word List	Action	Relation Added
0	[root]	[book, me, the, morning, flight]	SHIFT	(book → me)
1	[root, book]	[me, the, morning, flight]	SHIFT	
2	[root, book, me]	[the, morning, flight]	RIGHTARC	
3	[root, book]	[the, morning, flight]	SHIFT	
4	[root, book, the]	[morning, flight]	SHIFT	
5	[root, book, the, morning]	[flight]	SHIFT	(morning ← flight)
6	[root, book, the, morning, flight]	[]	LEFTARC	
7	[root, book, the, flight]	[]	LEFTARC	
8	[root, book, flight]	[]	RIGHTARC	
9	[root, book]	[]	RIGHTARC	
10	[root]	[]	Done	(root → book)

Пример работы [[Jurafsky & Martin 2019](#)]



А что тут с непроективностью?

- Есть математическое доказательство, что сферический в вакууме transition based-parser может разобрать любое проективное предложение
- А непроективное?



Нарушен принцип пересечения — предложение непроективное.

Базовый ТВ-parser и непроективность

Я мальчика вижу красивого

Шаг	Стэк	Буфер
0	[root]	[я, мальчика, вижу, красивого]
1	[root, я]	[мальчика, вижу, красивого]
2	[root, я, мальчика]	[вижу, красивого]
3	[root, я, мальчика, вижу]	[красивого]
4	[root, я, мальчика, вижу, красивого]	[]

Цугцванг? Или можно соединить “вижу” и “мальчика”?..

Базовый TV-parser и непроективность

Шаг	Стэк	Буфер	Операция
0	[root]	[я, мальчика, вижу, красивого]	SHIFT
1	[root, я]	[мальчика, вижу, красивого]	SHIFT
2	[root, я, мальчика]	[вижу, красивого]	SHIFT
3	[root, я, мальчика, вижу]	[красивого]	LeftArc
4	[root, я, вижу]	[красивого]	LeftArc
5	[root, вижу]	[красивого]	SHIFT
6	[root, вижу, красивого]	[]	RightArc
7	[root, вижу]	[]	RightArc
8	[root]	[]	

Модификации для непроективности

via [[Kuhlmann, Nivre 2010](#)]

- Pseudo-projective parsing [[Nivre, Nilsson 2005](#)]
- Non-adjacent arc-transitions [[Attardi 2006](#)]
(и далее [[Cohen et al. 2011](#)], [[Gómez-Rodríguez et al. 2014](#)])
- Online reordering [[Nivre 2009](#)]

Сравнивались на корпусах немецкого, чешского и английского языков

Спойлер: последнее чуть лучше остальных...

Online reordering

- Концепция: добавить **четвертую операцию SWAP** — вернуть второй элемент стека обратно в буфер, тем самым изменив порядок токенов в буфере (т.е. исходный порядок слов!)
- Лучше ее применять с пониженным весом (только если других вариантов больше не осталось), см. [[Nivre et al. 2009](#)]

Online reordering

- Пусть есть предложение “A B”
- Тогда я могу вернуть A в буфер, и если я потом его верну снова в стек, то как бы получу строку “B A”
- Поэтому online **reordering**: де-факто я могу рассматривать строку с другим порядком токенов.
- Хочется из “я мальчика вижу красивого” получить хотя бы “я мальчика красивого вижу”...

Online reordering

Я мальчика вижу красивого

Шаг	Стэк	Буфер	Операция
0	[root]	[я, мальчика, вижу, красивого]	SHIFT
1	[root, я]	[мальчика, вижу, красивого]	SHIFT
2	[root, я, мальчика]	[вижу, красивого]	SHIFT
3	[root, я, мальчика, вижу]	[красивого]	SHIFT
4	[root, я, мальчика, вижу, красивого]	[]	SWAP
5	[root, я, мальчика, красивого]	[вижу]	RightArc

Online reordering

Я мальчика вижу красивого

Шаг	Стек	Буфер	Операция
6	[root, я, мальчика]	[вижу]	SHIFT
7	[root, я, мальчика, вижу]	[]	LeftArc
8	[root, я, вижу]	[]	LeftArc
9	[root, вижу]	[]	RightArc
10	[root]	[]	

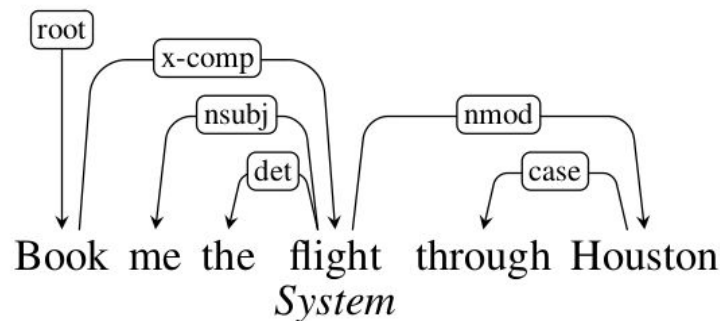
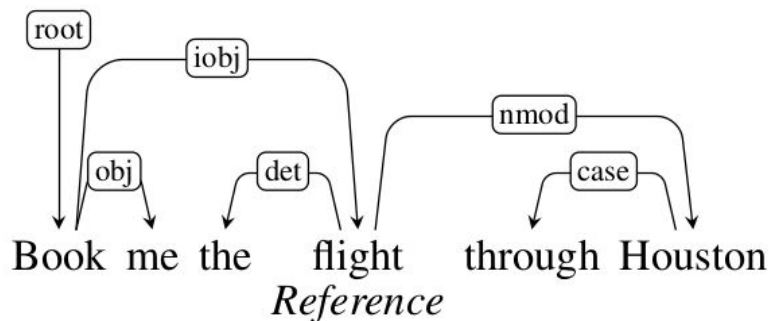
Transition-based vs graph-based

- Обе известны давно, “нейронная революция” улучшила сами классификаторы, но не поменяла архитектуры
- В GB лучше видим глобальное, все связи, работает медленнее, но лучше с длинными предложениями и непроективностью
- В TB видим следующий шаг, локальное, зато работает быстрее
- До 2018 было больше работ по TB, потом резкий переход в сторону GB (см. далее), но ситуация всегда может меняться
- Возможно, со временем bert-based вытеснит их обеих; что-то могут поменять и enhanced dependencies, см. далее
- Возможно, что-то из этого — уже история науки...

Содержание

1. Что такое синтаксис и зачем он нужен
2. Теоретические фреймворки
3. Dependency parsing
4. **Метрики и соревнования**
5. Инструменты
6. Varia

Оценка качества



Unlabeled Attachment Score (UAS) = 5/6

(правильно приписана вершина)

Labeled Attachment Score (LAS) = 4/6

(правильно приписана вершина И тип метки)

+ macro-averaged vs micro-averaged (по предложениям vs независимо от предложений)

И что, берем accuracy?

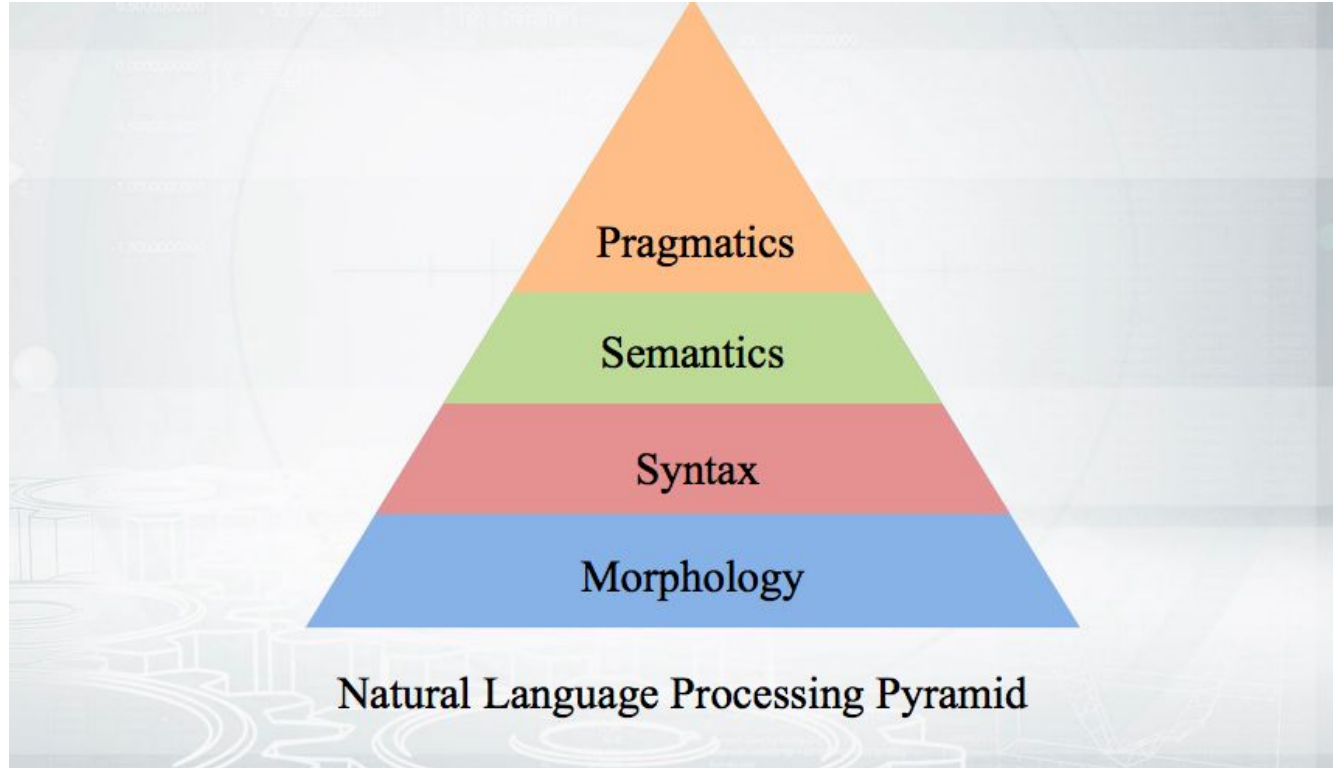
Проект Universal dependencies

- Лингвистическая проблема: несоответствие терминов и правил из грамматик зависимостей разных языков;
- Computational challenge: обучить синтаксический парсер для многих языков, включая low-resource languages;
=> <http://universaldependencies.org/> [Nivre et al. 2016]
- > 100 версионированных трибанков (синтаксически размеченных корпусов) для уже почти 100 (!!) языков, теги зависимостей унифицированы.

Соревнование пайплайнов

- [CONLL 2017 Shared Task](#) “from raw text to dependencies”
- 81 трибанк, 49 языков (82 трибанка, 57 языков в 2018)
- Парсинг сырого текста **vs** брать бейзлайн токенизацию и/или морфологию
- Один из лучших результатов среди всех команд и всех трибанков в 2017 был показан на корпусе русского языка:
94% UAS и 92,6% LAS (чуть меньше в 2018)

From raw text to dependencies



Соревнование пайплайнов: к дискуссии о метриках

- [Conll 2017 Shared Task](#)
- F-мера!
- Точность = количество точных попаданий/количество предсказаний;
- Полнота = количество точных попаданий/количество связей в размеченных данных;
- От чего зависит полнота?

Соревнование пайплайнов: к дискуссии о метриках

- При идеальной токенизации точность совпадает с полнотой, а значит:

$$F = 2PR/(P+R) = 2*x*x/(x+x) = x \text{ (=точность=полнота=accuracy)}$$

- При неидеальной токенизации значение F-меры меняется (точность \neq полнота);
- “Синтаксическая” метрика зависит от токенизации.

Соревнование пайплайнов: к дискуссии о метриках

- Не очень честное сравнение пайплайнов (кто-то мог не делать свою токенизацию);
- Отсутствие метрики, позволяющей чисто теоретически выделить идеальный сферический парсер в вакууме;
- Отсутствие единой метрики, позволяющей сравнить весь пайплайн целиком (from raw text to dependencies)
- Но как делать иначе — не очень понятно...

Соревнование пайплайнов: [Conll 2018 Shared Task](#)








- **LAS** (labeled attachment score) will be computed the same way as in the 2017 task so that results of the two tasks can be compared.
- **MLAS** (morphology-aware labeled attachment score) is inspired by the CLAS metric computed in 2017, and extended with evaluation of POS tags and morphological features.
- **BLEX** (bi-lexical dependency score) combines content-word relations with lemmatization (but not with tags and features).

Соревнование пайплайнов: [IWPT 2020](#)

- Эллипсис, сочинение и др. (т.н. *enhanced dependencies*): введем дополнительную колонку, где у некоторых слов будет несколько вершин (см. [полный список](#) было-стало)
- Более точные отношение типа не прямое дополнение: *Петя и Вася -> conj: **and***
- Еще одна метрика: Enhanced LAS (ELAS): F1-мера по enhanced LAS, остальные метрики те же
- Значит, перед нами не просто дерево. Повлияет ли это на архитектуры? Или будут правила поверх деревьев?..

Данные: русский язык

Russian treebanks

▶	GSD	98K	(L)(F)	W		★★★★☆
▶	SynTagRus	1,107K	(L)(F)(D)			★★★★☆
▶	Taiga	38K	(L)(F)			★★★★☆
▶	PUD	19K	(L)(F)			★★★★☆

See [here](#) for comparative statistics of Russian treebanks.

=> Практически все эксперименты проводятся на Syntagrus

(корпуса для всех языков [лежат на гитхабе](#))

Содержание

1. Что такое синтаксис и зачем он нужен
2. Теоретические фреймворки
3. Dependency parsing
4. Метрики и соревнования
- 5. Инструменты**
6. Varia

Инструменты

- См. результаты дорожек [Conll-17](#) и [Conll-18](#);
- Осторожно: есть академические и закрытые разработки (а еще многое уже устарело...);
- UDPipe (TB; и есть еще [UDPipe-future](#), GB);
- Syntaxnet (no updates since 2017; transition-based)
- [Deeppavlov](#) (BERT as input + graph-based)
- С 2018 все пересаживаются на (графовую) архитектуру а-ля [[Dozat, Manning 2017](#)]: bi-LSTM with attention

UDPipe vs Syntaxnet vs DeepPavlov

	UDPipe (1.2)	UDPipe (2.0)	Syntaxnet (parseysaurus-17)	Deeppavlov
UAS (russian)	92.96%	94.92%	92.67%	95.2%
LAS (russian)	91.46%	93.68%	88.68%	93.7%
Время парсинга 1 предложения, с/и	~ 3 ms	??	~ 100 ms	~ 50 ms
Возможность “распила” пайплайна	+	+	-	+
Запуск напрямую без докера и др.	+	+	-	+ (??)

UDPipe (старый!!)

- UDPipe — пайплайн, обучаемый токенизации, лемматизации, морфологическому тэггингу и парсингу, основанному на грамматике зависимостей
- [Статья об архитектуре](#), [репозиторий с кодом обучения](#), [мануал](#)
- Есть готовые [модели](#) (в том числе и для РЯ)
- Подобранные для каждого корпуса параметры обучения [зарелизены](#)

UDPipe: архитектура

- **Совместное деление на слова и предложения:**
однослойная двухсторонняя GRU, для каждого символа предсказывающая, последний ли он в предложении и/или токене.
- **Теггер:** по последним четырем символам каждого слова генерируем триплеты (UPOS, XPOS, FEATS), при помощи перцептрона выбираем лучшего кандидата.

UDPipe: архитектура

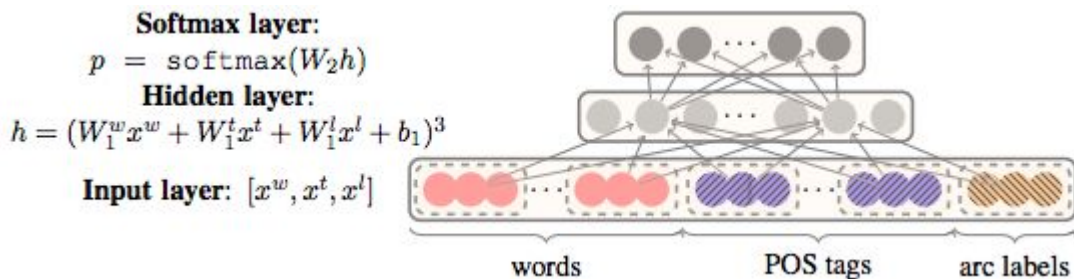
- **Лемматизатор:** генерируем пары (lemma rule, UPOS), лемму предсказываем, отрезая префиксы и суффиксы и генерируя новые на их место; перцептрон выбирает наилучшего кандидата
- **Раздельное предсказание тегов и лемм**
(2 модели, но можно соединить в одну)
- + можно подключить свой список лемм

UDPipe: архитектура

- Dependency parsing ([[Straka et al. 2015](#)]): transition-based arc-standard dependency parser;
- Один скрытый слой, нет рекуррентности, см. картинку;
- Mini-batched SGD при обучении;
- До 18 источников фич на вход: 3 элемента на вершине стека, 3 элемента на вершине буфера, первый и второй левый и правый потомки 2 элементов на вершине стека, и самый левый и самый правый потомок 2 элементов на вершине стека;
- Можно загрузить pre-trained семантические эмбединги форм или лемм (см. *мышь ест стол*).

UDPipe: фичи для парсера

- Each node is represented using distributed representations of its form, its POS tag and its arc label; the latter only if it has already been assigned;
- word2vec-like training
- Network like in [\[Chen 2014\]](#):



Содержание

1. Что такое синтаксис и зачем он нужен
2. Теоретические фреймворки
3. Dependency parsing
4. Метрики и соревнования
5. Инструменты
6. **Varia**

Varia

- А кто там щас SoTA?

- Все очень зависит от языка, SoTA из [таблички Рудера](#) проверялась только на английском и китайском (в обоих “жесткий” порядок слов)
- [Китайцы \(ой, уже не китайцы\)](#) поверх XLNet, но там довольно хитрый препроцессинг дерева в сторону хитрого формализма
- Они ближе к графовой архитектуре, но на самом деле там что-то третье
- Лидер меняется раз в два месяца...

- А можно unsupervised?

- [Можно](#), уже лет 10 [пытаются](#)! Но пока там своя лига...

Varia

- А что там BERT?

- На ранних слоях относительно компактно сосредоточено “знание” о синтаксисе
- Сырой BERT показывает очень неплохой (82,5%) скор по UUAS
- Правда, пока считали только на английском...
- Еще смотрели на влияние аттеншн хэдов, специальной “синтаксической” головы вроде нет. И это кажется логичным: разные синтаксические отношения выглядят по-разному и “весят” по-разному, поэтому попали в разные головы.
- Умрут ли в дальнейшем традиционные архитектуры?..

Varia

- А что делать, если у языка нет обучающего корпуса?
 - Взять корпуса родственных языков, взболтать, перемешать, обучить [delexicalized model](#)
- А другая морфология (кроме POS-тегов) влияет?
 - *Глокая куздра штеко будланула бокра и курдячит бокрѣнка*
 - Непохоже, см. [мою статью на хабре](#)
 - Но то, что влияет, — влияет на целевые примеры
 - Автор UDPipe [утверждает](#), что влияет :)
 - Но после появления BERT морфология не учитывается при обучении синтаксиса...

Varia

- А если у меня будет идеальная морфология, я смогу сделать идеальный парсер при помощи распространенного инструмента?
 - Попробуйте! Но вообще не в случае с русским, я [писал на хабре](#)
- А что делать с веб-текстами, в них же все по-другому, включая пунктуацию?
 - А чёрт его знает... Нужно собирать корпус...

Varia: GramEval-2020, русский язык

- Соревнование на Диалоге: на входе список токенов, на выходе леммы, морфологические и синтаксические теги.
- Кроме обычных текстов — соцсети, поэзия и тексты 17 века
- Победитель (90% LAS на обычных текстах, 80% поэзия, 66% 17 век);
- Большинство ошибок — на периферии ГЗ (сочинение, союзы)
- Берт + графовая архитектура (практически как у Deerpravlov)
- Похожая архитектура у второго места
- Результаты:
 - с одной стороны очевидно, что без берта (и графовой архитектуры) никуда;
 - с другой, учимся на “чистом” синтагматическом анализе, при переходе в другие домены (а чаты?) качество падает...

Полезные ссылки

- [Об архитектуре парсера в Spacy + библиография](#)
- [Программа воркшопа на EMNLP-18](#)
- [Материалы курса на ESSLI-18](#)
- [J. Nivre's workshop at EACL-2014](#)
- [Краткое саммари на сайте Себастиана Рудера про результаты в Dependency Parsing](#)
- [SyntaxRuEval-2012](#)

Заключение

- **Теоретические фреймворки:** ГЗ vs ГНС;
- **Dependency parsing:** minimum spanning tree (aka graph-based) vs transition-based подходы;
- **Метрики:** UAS, LAS, ELAS + более сложные для from raw text to dependencies;
- **Соревнования:** SyntaxRuEval-12, CoNLL-17, -18, dialogue-2019, -2020, IWPT 2020;
- **Данные:** корпуса Universal Dependencies (в формате conllu);
- **Инструменты:** Syntaxnet, UDPipe, Deeppavlov.

СПАСИБО ЗА ВНИМАНИЕ!

Денис Кирьянов, Sberdevices, denkirjanov@gmail.com

telegram: @kirdin

Автор благодарит за помощь в подготовке презентации
А. М. Алексеева, М. Ю. Князева и Е. Л. Артёмову.