Настоящее и будущее алгоритмов на текстах: NLP и production

Штех Геннадий
NAUMEN
19.04.2019

#DUMP2019

https://github.com/ShT3ch/public_workshop

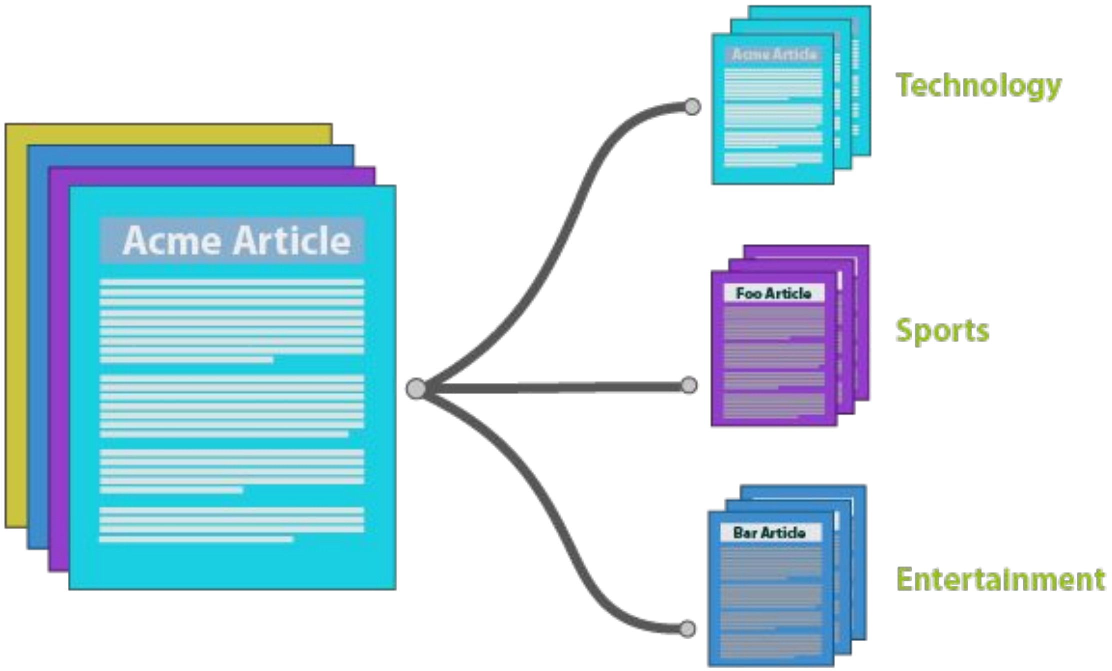
О чем поговорим

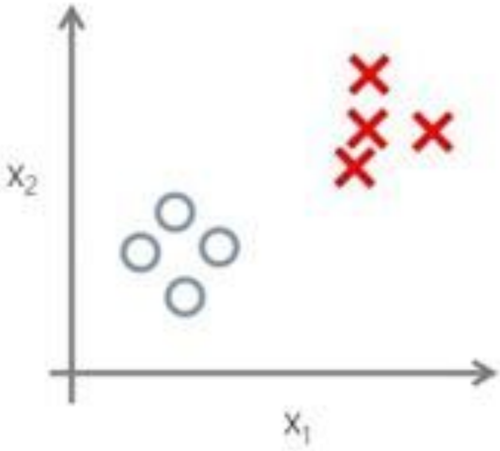
1. Об NLP и задачах, которые он решает
2. NLU: Эволюция подходов к решению задач
3. Методы тестирования NLU
4. Выбор моделей в продакшн
5. Пара слов о будущем

NLP-задачи

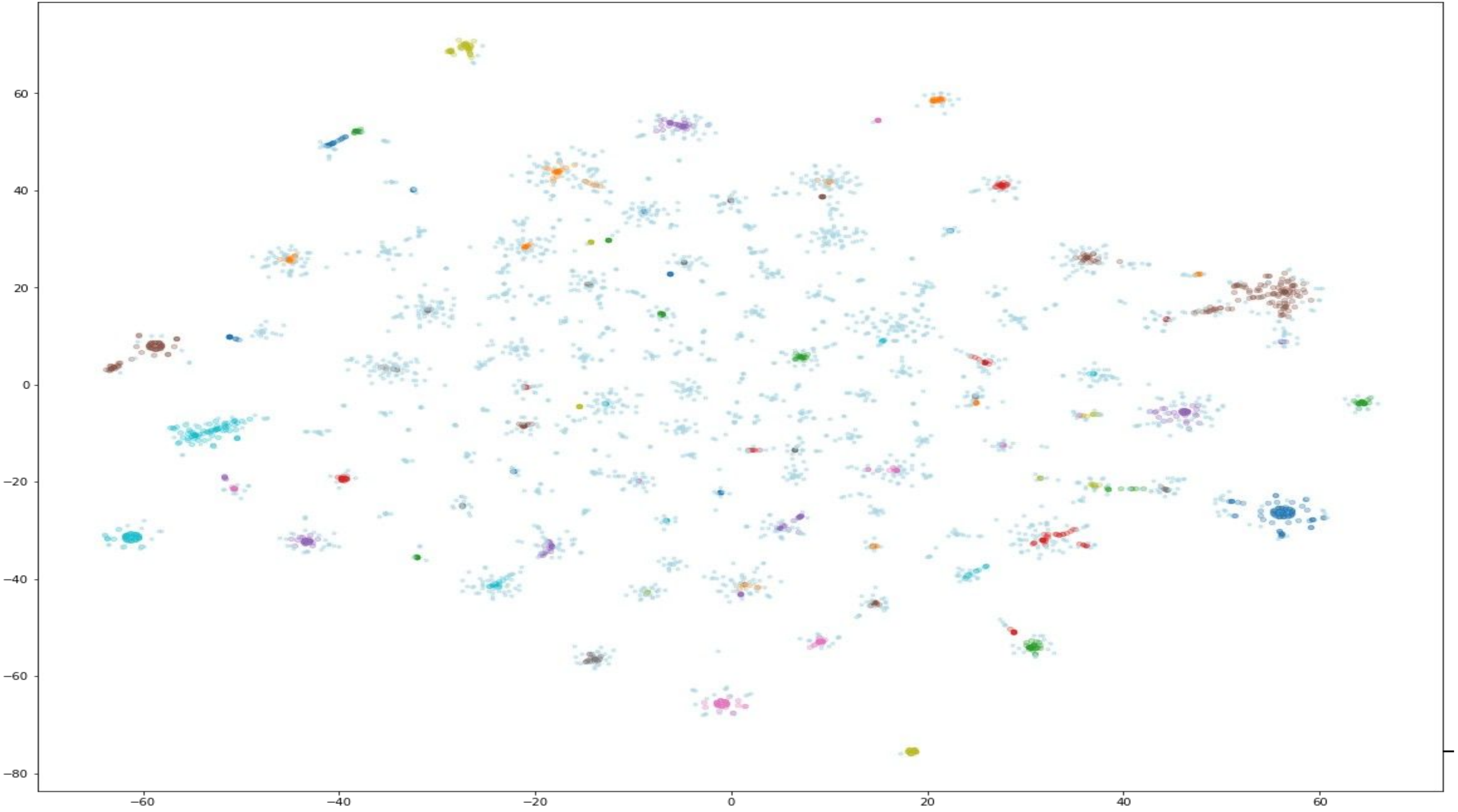
- Классификация
- Кластеризация
- Заполнение форм
- Машинный перевод
- Поиск
- Лингвистические задачи

Классификация





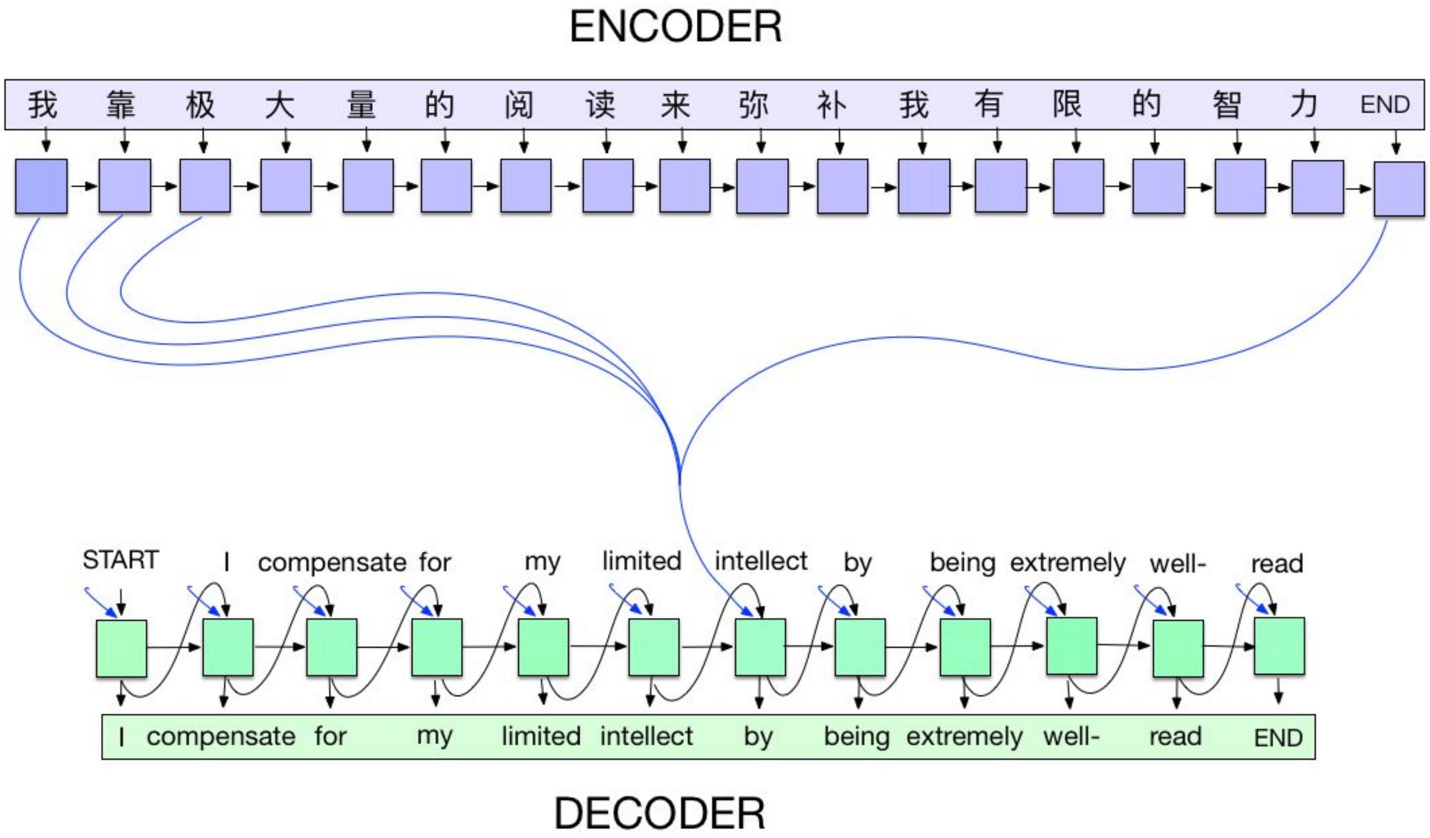
Кластеризация



Заполнение форм

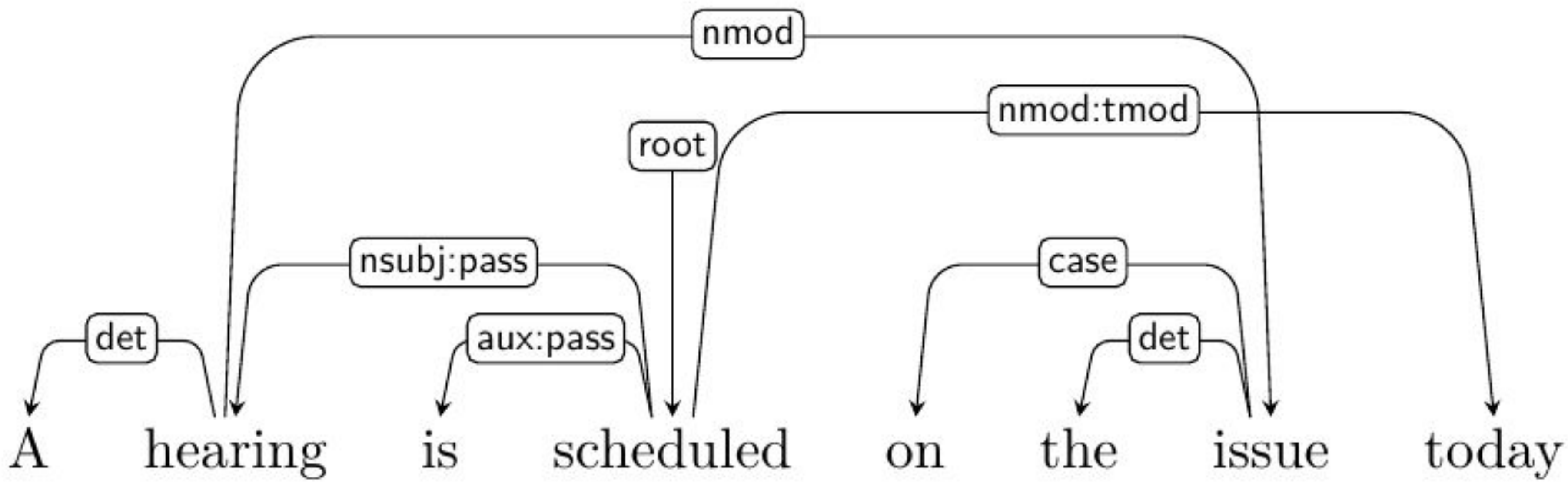
contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

Машинный перевод



Поиск

Лингвистика, POS tagging, dependency parsing



Эволюция инструментов

NLU

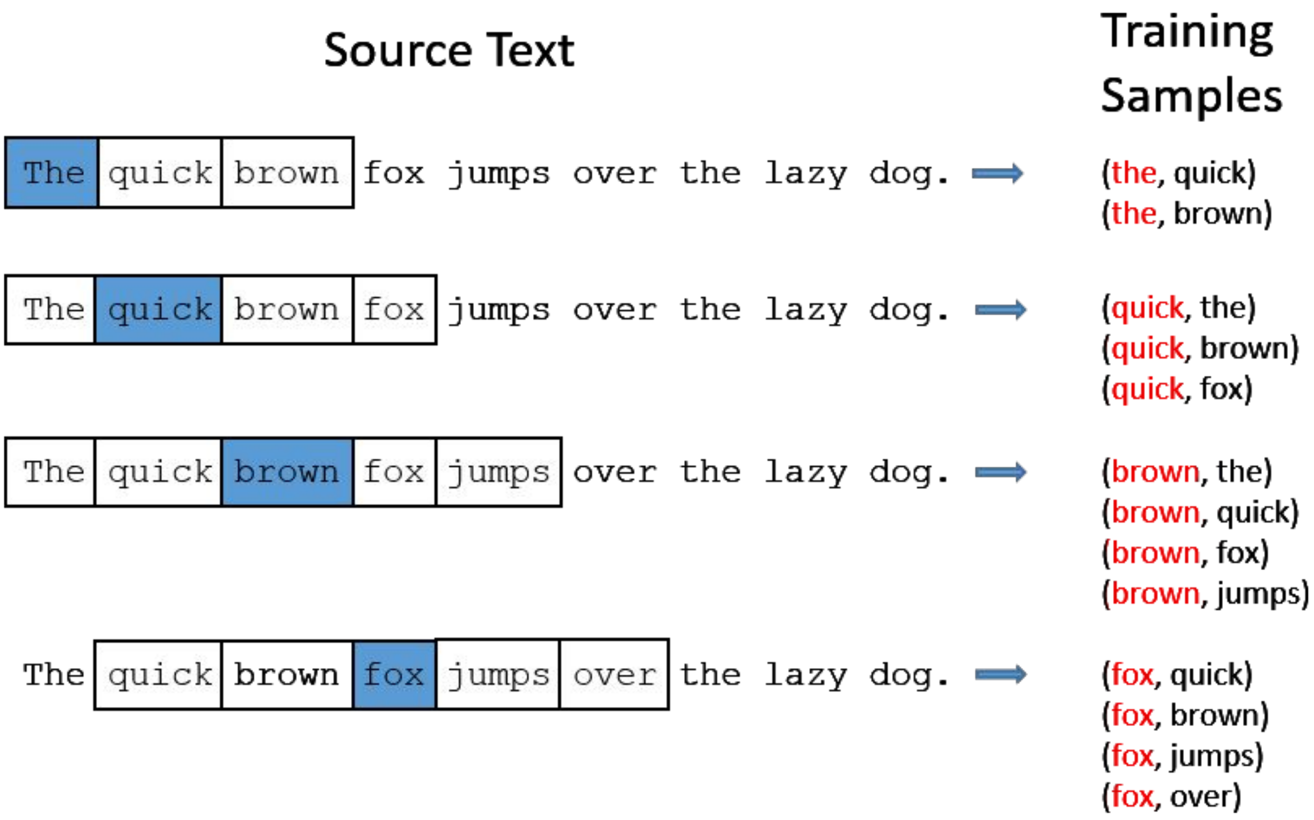
NLU

В чем сдвиг парадигмы?

Методы получения NLU-моделей

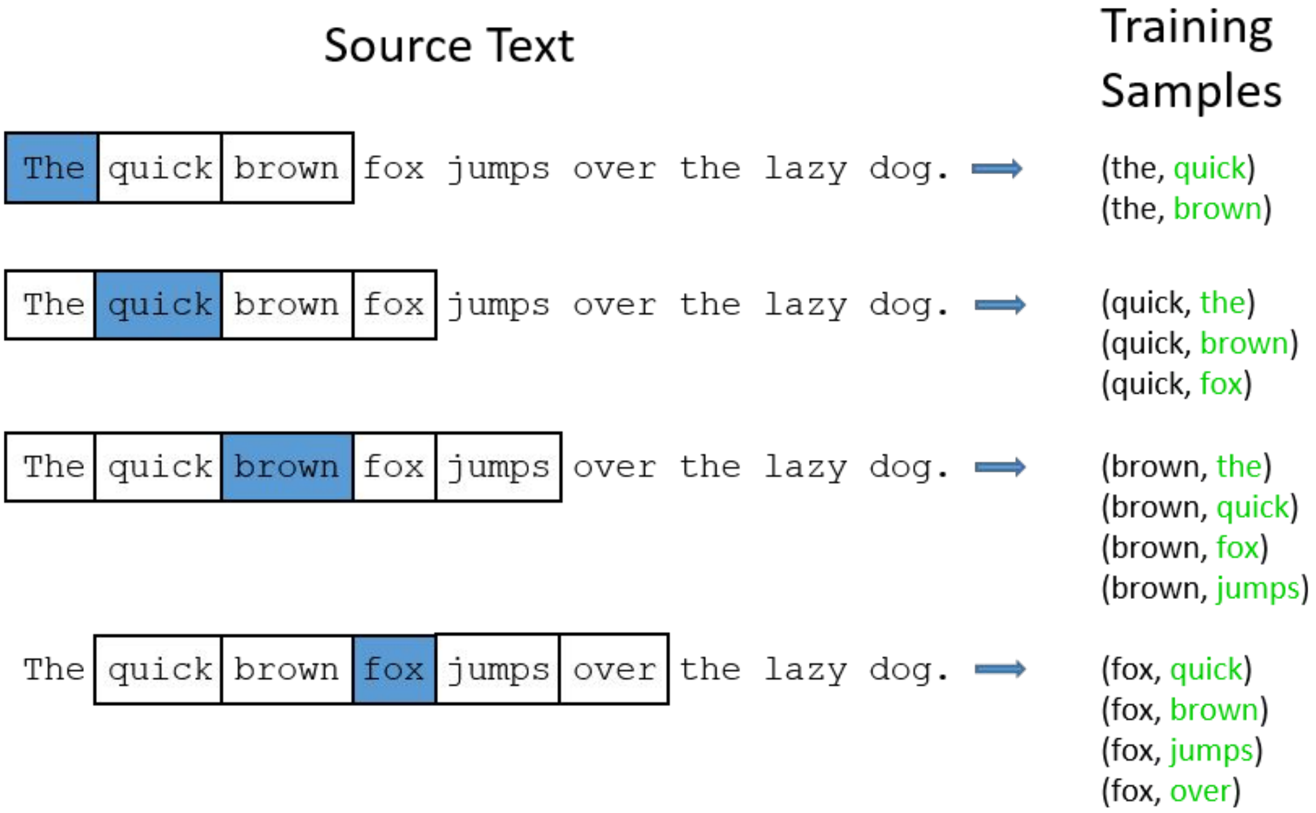
- Skip-Gram (CBOW)
- Language Modeling
- Masking
- Skip-thoughts
- Multi task
- Autoencoder

Skip-gram



https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Skip-gram



https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

NLU

Хроника появления решений

2013

Эмбеддинги слов

- \+ Моделируют язык
- \+ Являются хорошими признаками слов

- — Не являются алгоритмом для эмбеддинга сразу нескольких слов (текста)
- — Недостаточно выразительны, происходит смешение контекстов

Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

2014

Методы работы с текстами на LSTM

+ Позволяют работать с текстами, как с последовательностями

— Работают достаточно медленно
— Требуют большого количества данных
— Плохо работают на достаточно длинных последовательностях

http://proceedings.mlr.press/v32/le14.pdf

2015

Методы работы с текстами на GRU, CNN

- \+ Позволяют работать с текстами как с последовательностями
- \+ Работают быстрее LSTM

- — Требуют значительного количества данных
- — Плохо работают на достаточно длинных последовательностях

https://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf, http://papers.nips.cc/paper/5782-character-level-convolutional-networks-fo

2016

Attention и дополненные LSTM/GRU

+ Позволяют работать с текстами, как с последовательностями
+ Хорошо работают на длинных текстах

– Требуют значительного количества данных

http://www.aclweb.org/anthology/P16-2034

2017

Transformer

+ Побил по качеству многие известные алгоритмы
+ Не зависит от предобученных эмбеддингов
+ Моделирует тексты более естественным образом

– Требует много данных

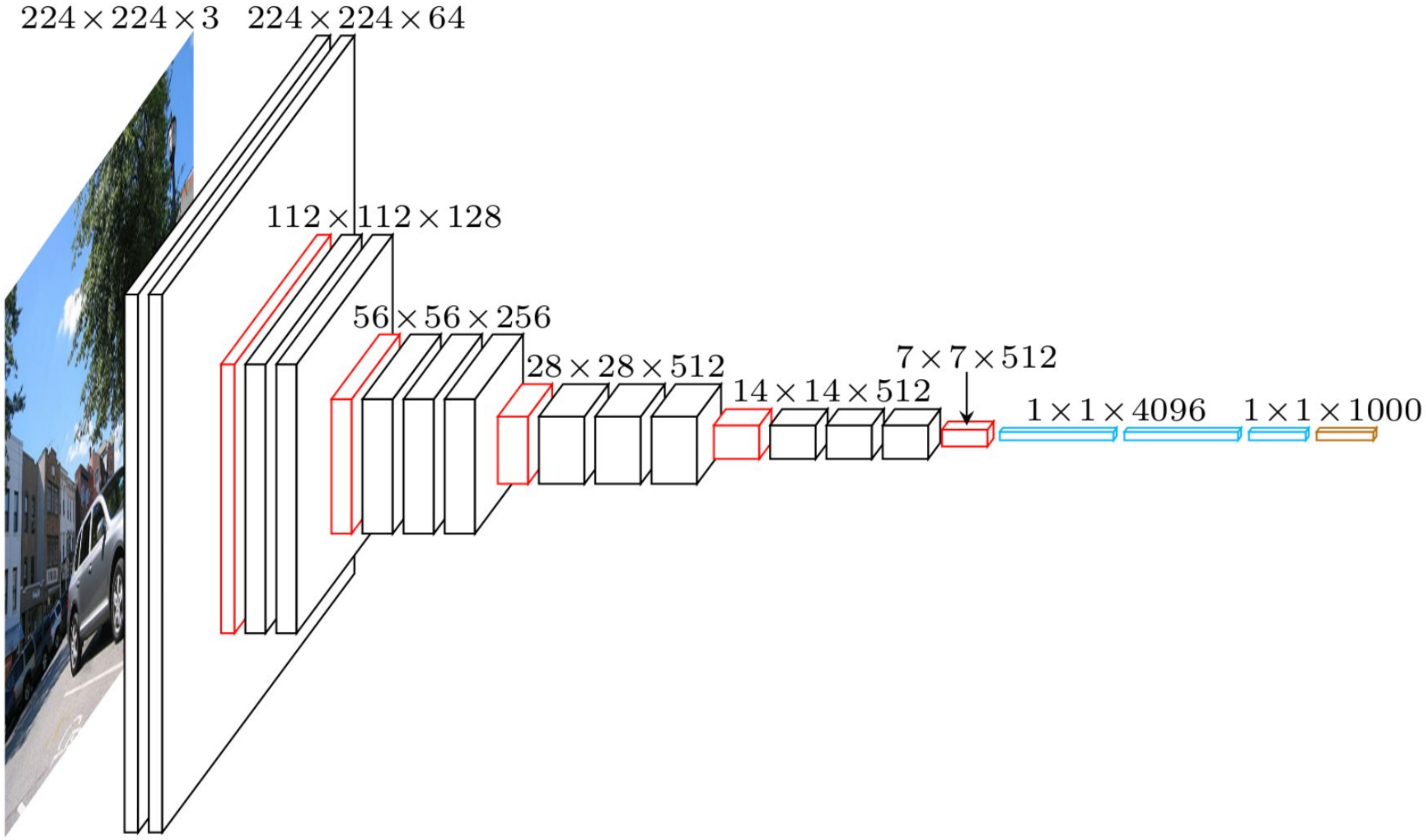
https://arxiv.org/abs/1706.03762, http://jalammar.github.io/illustrated-transformer/,
https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need/#.XIWlzBNKjOR

2018

Transfer Learning 2: ULMfit

+ Почти не требует размеченных данных

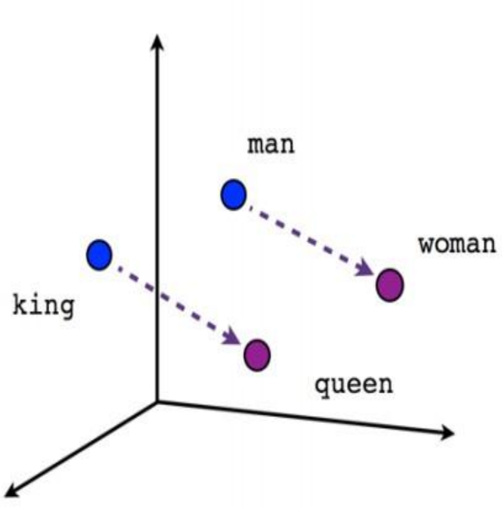
https://arxiv.org/abs/1801.06146



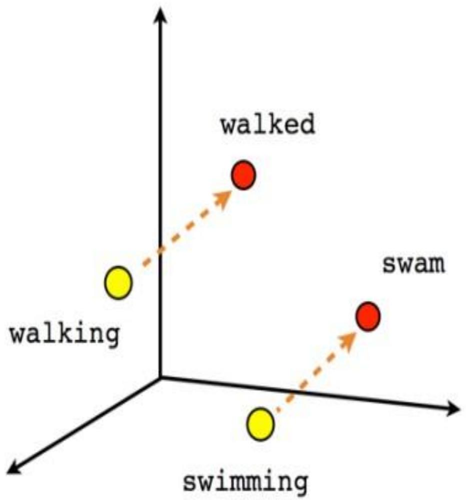
$224 \times 224 \times 3$
$224 \times 224 \times 64$
$112 \times 112 \times 128$
$56 \times 56 \times 256$
$28 \times 28 \times 512$
$14 \times 14 \times 512$
$7 \times 7 \times 512$
$1 \times 1 \times 4096$
$1 \times 1 \times 1000$

2013

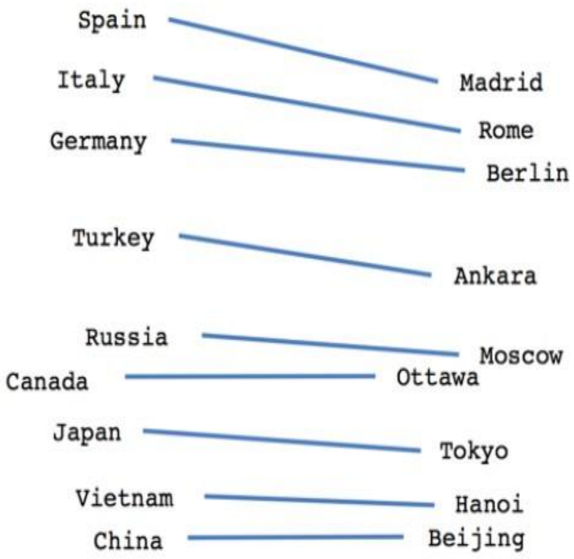
Эмбеддинги слов



Male-Female



Verb tense



Country-Capital

2018

Контекстно-зависимые эмбеддинги

+ Знают, что Apple бывает разный
+ Универсальны для дальнейшего применения
+ Дают хорошую базу для работы остальных алгоритмов

— Медленно работают

https://allennlp.org/elmo

Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

WORD	$p(z)$	NEAREST NEIGHBOURS
python	0.33	monty, spamalot, cantsin
	0.42	perl, php, java, c++
	0.25	molurus, pythons
apple	0.34	almond, cherry, plum
	0.66	macintosh, iifx, iigs
date	0.10	unknown, birth, birthdate
	0.28	dating, dates, dated
	0.31	to-date, stateside
	0.31	deadline, expiry, dates
bow	0.46	stern, amidships, bowsprit
	0.38	spear, bows, wow, sword
	0.16	teign, coxs, evenlode
mass	0.22	vespers, masses, liturgy
	0.42	energy, density, particle
	0.36	wholesale, widespread

2018

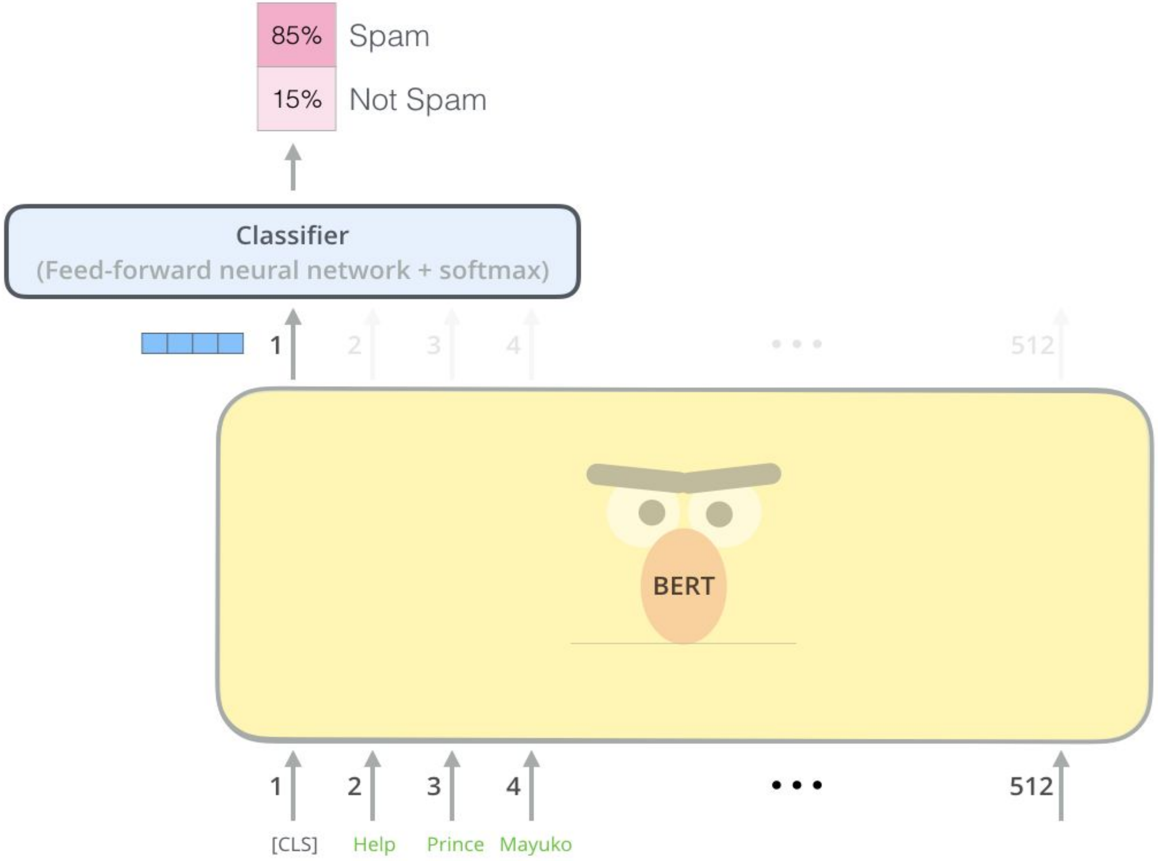
BERT

+ Почти не требует данных
+ Это трансформер
+ По-настоящему глубокая нейросеть
+ Бьёт все остальные архитектуры

— Медленный, да

http://jalammar.github.io/illustrated-bert/

Трансформеры



Методы тестирования NLU-моделей

https://arxiv.org/abs/1804.07461

Рассмотрим рост метрик подробнее

Model	Score
GLUE Human Baselines	87.1
BERT: 24-layers, 16-heads, 1024-hidc	80.5
Singletask Pretrain Transformer	72.8
BiLSTM+ELMo+Attn	70.0
BiLSTM+ELMo	67.7
BiLSTM+Attn	65.6
BiLSTM	64.2
CBOW	58.6

SINGLE SENTENCE TASKS

CoLA: The Corpus of Linguistic Acceptability (Warstadt et al., 2018)

SST-2: The Stanford Sentiment Treebank (Socher et al., 2013)

CoLA

1 John fed the baby up with rice.
0 John fed the baby rice up.

1 Spray all the paint onto the wall completely.
0 Spray the wall with all the paint.

1 The man who I gave John a picture of was bald.
0 The man who I gave John Ed's picture of was bald.
0 The man who I gave John this picture of was bald.

1 The noise gave Terry a headache.
0 The noise gave a headache to Terry.

Метрики, SINGLE SENTENCE TASKS

Model	Score	CoLA	SST-2
GLUE Human Baselines	87.1	66.4	97.8
BERT: 24-layers, 16-heads, 1024-hidc	80.5	60.5	94.9
Singletask Pretrain Transformer	72.8	45.4	91.3
BiLSTM+ELMo+Attn	70.0	33.6	90.4
BiLSTM+ELMo	67.7	32.1	89.3
BiLSTM+Attn	65.6	18.6	83.0
BiLSTM	64.2	11.6	82.8
CBOW	58.6	0.0	80.0

SIMILARITY AND PARAPHRASE TASKS

MRPC: The Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005)

QQP: The Quora Question Pairs

STS-B: The Semantic Textual Similarity Benchmark (Cer et al., 2017)

The Quora Question Pairs

question1	question2	is_duplicate
What is the step by step guide to invest in share market in india?	What is the step by step guide to invest in share market?	0
What is the story of Kohinoor (Koh-i-Noor) Diamond?	What would happen if the Indian government stole the Kohinoor (Koh-i-Noor) diamond back?	0
How can I increase the speed of my internet connection while using a VPN?	How can Internet speed be increased by hacking through DNS?	0
Why am I mentally very lonely? How can I solve it?	Find the remainder when [math]23^{24}[/math] is divided by 24,23?	0
Which one dissolve in water quikly sugar, salt, methane and carbon di oxide?	Which fish would survive in salt water?	0
Astrology: I am a Capricorn Sun Cap moon and cap rising...what does that say about me?	I'm a triple Capricorn (Sun, Moon and ascendant in Capricorn) What does this say about me?	1

Метрики, SIMILARITY AND PARAPHRASE TASKS

Model	Score	MRPC	STS-B	QQP
GLUE Human Baselines	87.1	86.3/80.8	92.7/92.6	59.5/80.4
BERT: 24-layers, 16-heads, 1024-hidc	80.5	89.3/85.4	87.6/86.5	72.1/89.3
Singletask Pretrain Transformer	72.8	82.3/75.7	82.0/80.0	70.3/88.5
BiLSTM+ELMo+Attn	70.0	84.4/78.0	74.2/72.3	63.1/84.3
BiLSTM+ELMo	67.7	84.7/78.0	70.3/67.8	61.1/82.6
BiLSTM+Attn	65.6	83.9/76.2	72.8/70.5	60.1/82.4
BiLSTM	64.2	81.8/74.3	70.3/67.8	62.5/84.2
CBOW	58.6	81.5/73.4	61.2/58.7	51.4/79.1

INFERENCE TASKS

MNLI: The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018)

QNLI: The Stanford Question Answering Dataset (Rajpurkar et al. 2016)

RTE: The Recognizing Textual Entailment

WNLI: The Winograd Schema Challenge (Levesque et al., 2011)

The Multi-Genre Natural Language Inference Corpus

The Old One always comforted Ca'daan, except today.	***neutral***	Ca'daan knew the Old One very well.
Your gift is appreciated by each and every student who will benefit from your generosity.	***neutral***	Hundreds of students will benefit from your generosity.

The Multi-Genre Natural Language Inference Corpus

yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual	***contradiction***	August is a black out month for vacations in the company.
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	***entailment***	People formed a line at the end of Pennsylvania Avenue.

Метрики, INFERENCE TASKS

Model	Score	MNLI-mm	QNLI	RTE	WNLI
GLUE Human Baselines	87.1	92.8	91.2	93.6	95.9
BERT: 24-layers, 16-heads, 1024-hidc	80.5	85.9	92.7	70.1	65.1
Singletask Pretrain Transformer	72.8	81.4	87.4	56.0	53.4
BiLSTM+ELMo+Attn	70.0	74.5	79.8	58.9	65.1
BiLSTM+ELMo	67.7	67.9	75.5	57.4	65.1
BiLSTM+Attn	65.6	68.3	74.3	58.4	65.1
BiLSTM	64.2	66.1	74.6	57.4	65.1
CBOW	58.6	56.4	72.1	54.1	62.3

SWAG

Situations With Adversarial Generations

On stage, a woman takes a seat at the piano. She a) sits on a bench as her sister plays with the doll. b) smiles with someone as the music plays. c) is in the crowd, watching the dancers. **d) nervously sets her fingers on the keys.**
A girl is going across a set of monkey bars. She a) jumps up across the monkey bars. b) struggles onto the monkey bars to grab her head. **c) gets to the end and stands on a wooden plank.** d) jumps up and does a back flip.
The woman is now blow drying the dog. The dog **a) is placed in the kennel next to a woman's feet.** b) washes her face with the shampoo. c) walks into frame and walks towards the dog. d) tried to cut her face, so she is trying to do something very close to her face.

Table 1: Examples from **SWAG**; the correct answer is **bolded**. Adversarial Filtering ensures that stylistic models find all options equally appealing.

SWAG

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

#DUMP2019

Image-caption retrieval

"A group of people on some horses riding through the beach."

Выбираем модели в продакшн

RNN VS CNN

			performance
TextC	SentiC (acc)	CNN	82.38
		GRU	**86.32**
		LSTM	84.51
	RC (F1)	CNN	68.02
		GRU	**68.56**
		LSTM	66.45
SemMatch	TE (acc)	CNN	77.13
		GRU	**78.78**
		LSTM	77.85
	AS (MAP & MRR)	CNN	(**63.69**,**65.01**)
		GRU	(62.58,63.59)
		LSTM	(62.00,63.26)
	QRM (acc)	CNN	**71.50**
		GRU	69.80
		LSTM	71.44

https://arxiv.org/pdf/1702.01923.pdf, https://arxiv.org/pdf/1707.05589.pdf

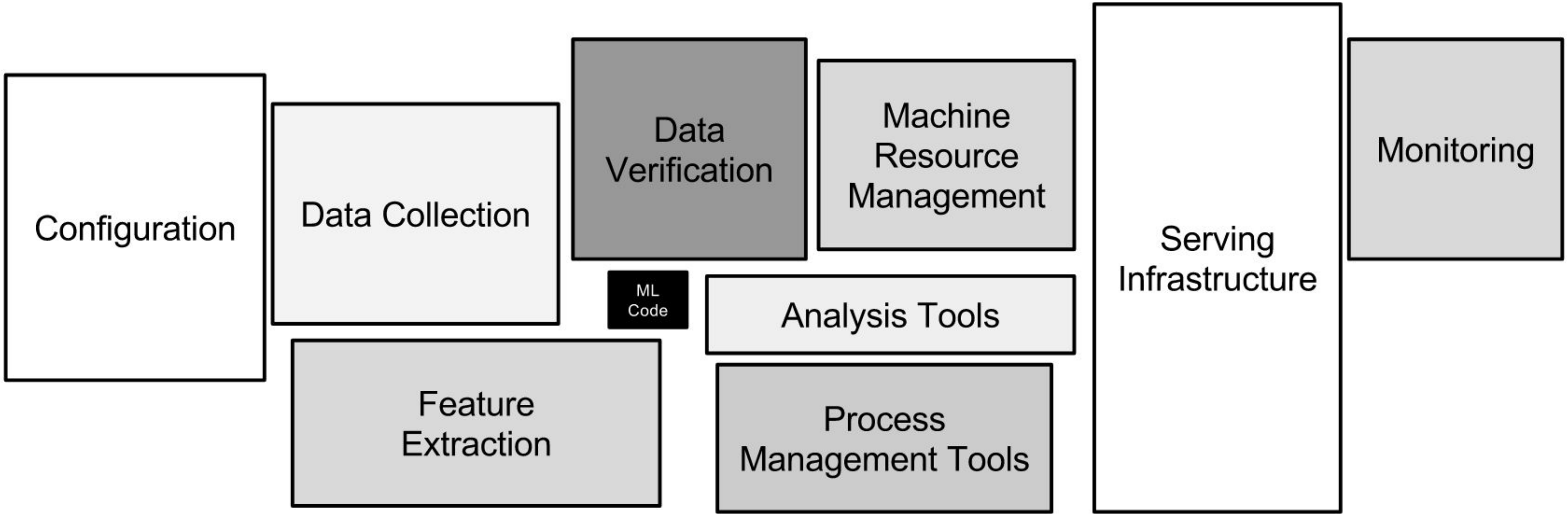
Владения инструментами недостаточно для построения эффективных решений

Важно не забывать о процессах

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Актуальные алгоритмы

Представление

- Tf-idf, nPMI, hashing trick, BPE

Факторизация (декомпозиция)

- PCA, LSI-LSA, pLSA, nNMF

Тематическое моделирование

- pLSA, LDA, HDP, ARTM

Поиски

- BM25, HNSW, LSH

Эмбеддинги

- word2vec, glove, doc2vec, fasttext, poincaré, ELMO

Нейросетевые подходы

- LSTM, GRU, TCN, Attention, siamese network, similarity learning, Transformer, Augmented RNN

Полезный NLP-софт

Предобработка текста (нормализация, токенизация)

- pymorphy2(ru), snowball stemmer(en), Stanford NLP(en)

Фреймворки

- sklearn, NLTK, gensim, spaCy

Узкоспециализированные фреймворки

- BigARTM, Vowpal Wabbit, Fasttext, faiss, annoy, NMSLib, lucene, sphinx, elastic

Нейросетевые фреймворки

- Pytorch, HuggingFace, AllenNLP, torchtext

Подходы и данные для тестирования моделей

- https://github.com/facebookresearch/SentEval
- https://arxiv.org/pdf/1707.05589.pdf
- https://arxiv.org/pdf/1806.06259.pdf
- https://aclweb.org/anthology/D18-1009
- https://arxiv.org/pdf/1702.02170.pdf
- https://arxiv.org/pdf/1903.09442.pdf

- https://leaderboard.allenai.org/swag/submissions/public
- https://gluebenchmark.com/leaderboard

О прогрессе в НЛП

- https://nlpoverview.com/#3
- https://arxiv.org/pdf/1708.02709.pdf
- http://nlpprogress.com/english/language_modeling.html
- https://github.com/Separius/awesome-sentence-embedding

Посмотрим на будущее

Появятся совсем простые фреймворки для использования глубоких предобученных сетей

Появятся фреймворки для семантического поиска документов

Разовьётся подход к генерации контента на основе RL

Скорее всего сети на гиперболических пространствах взорвут

BERT “облегчат”

Контакты

Штех Геннадий *
@ NAUMEN
gshtekh@naumen.ru

Gennady Shtekh
shtechgen@gmail.com
t.me/sht3ch
github.com/ShT3cH

*R&D Data Usage Department Executive

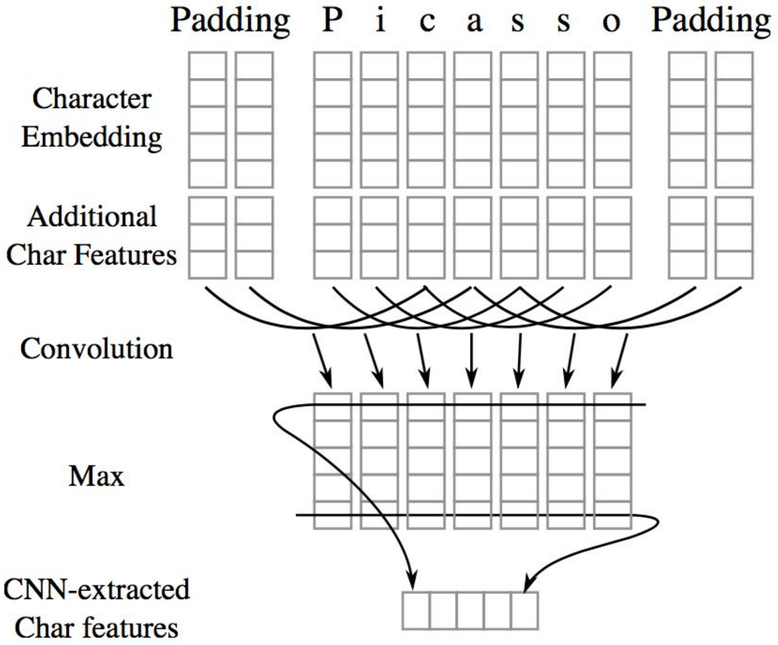
https://github.com/ShT3ch/public_workshop

Хроника появления решений

- Методы работы с текстами на LSTM
- Методы работы с текстами на GRU, CNN
- Attention и дополненные LSTM/GRU
- Transformer
- Transfer Learning
- Контекстно-зависимые эмбеддинги
- BERT

Подходы к решению OOV

Char-level Convolution



Проблема Out-Of-Vocabulary (OOV)

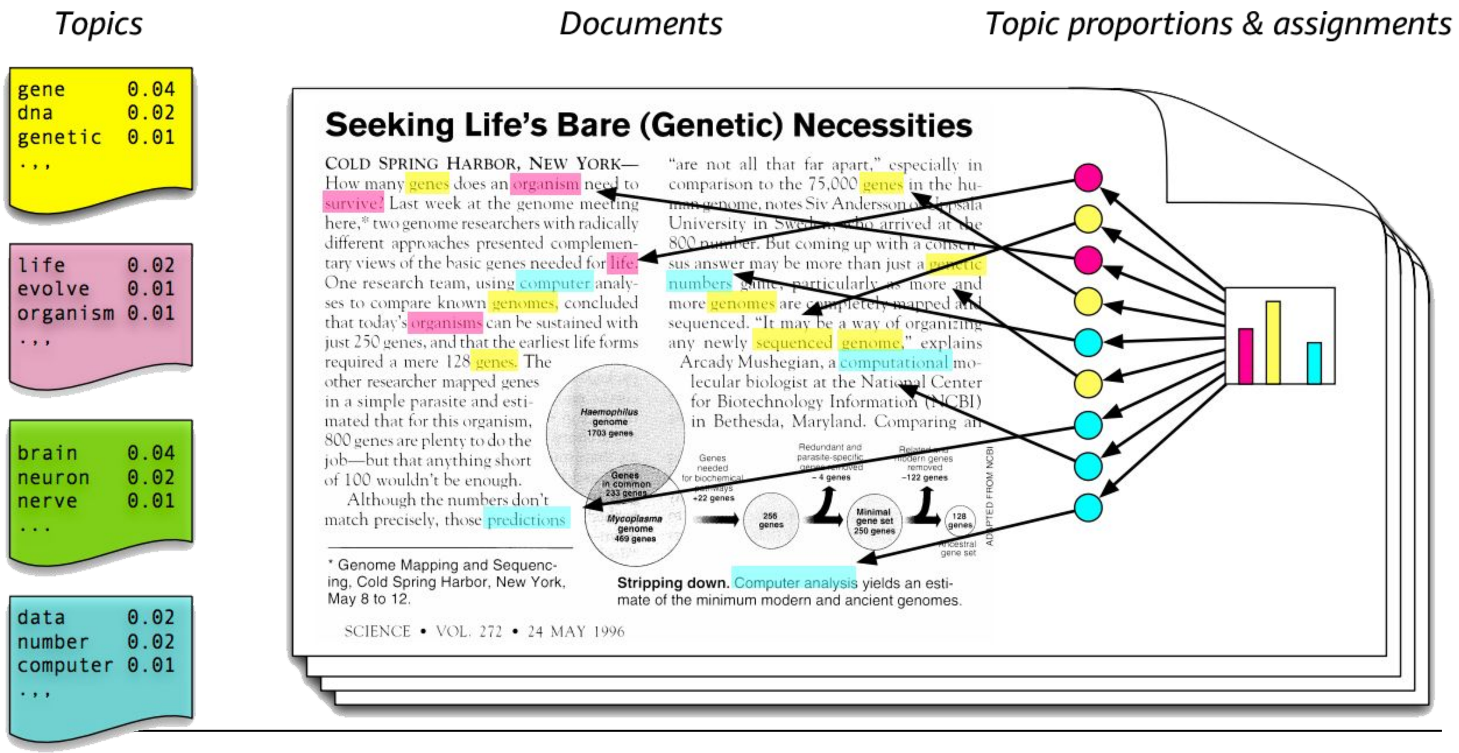
- Char-ngramm

<where>

<wh, whe, her, ere, re>

- Byte Pair Encoding

Тематическое моделирование

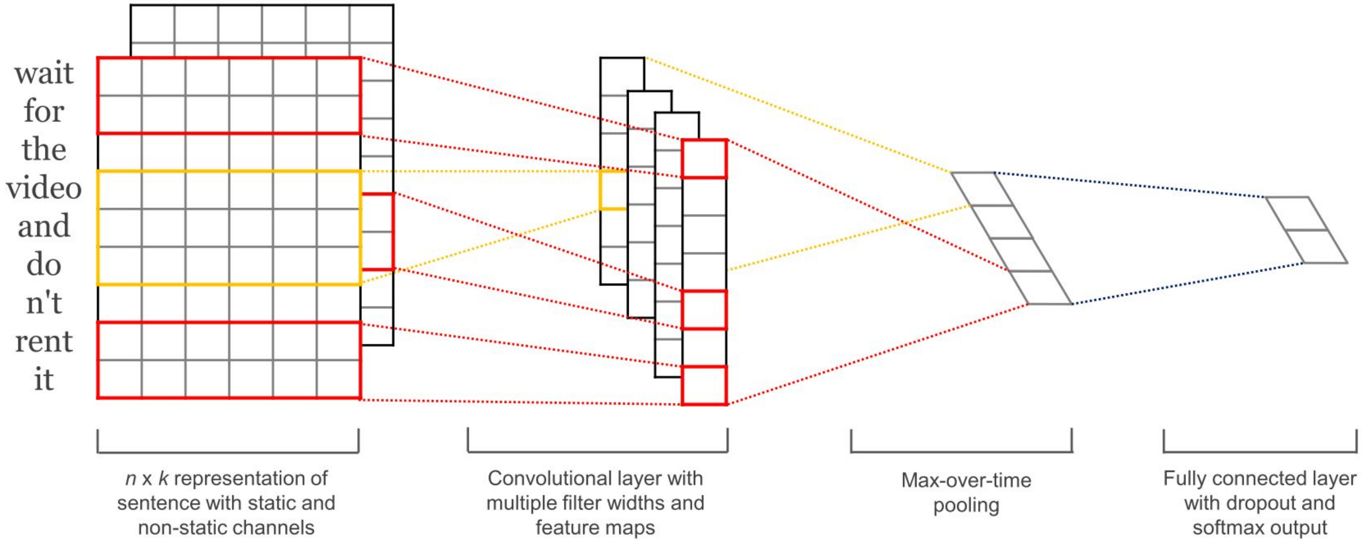


http://www.machinelearning.ru/wiki/images/6/6d/BigARTM-short-intro.pdf

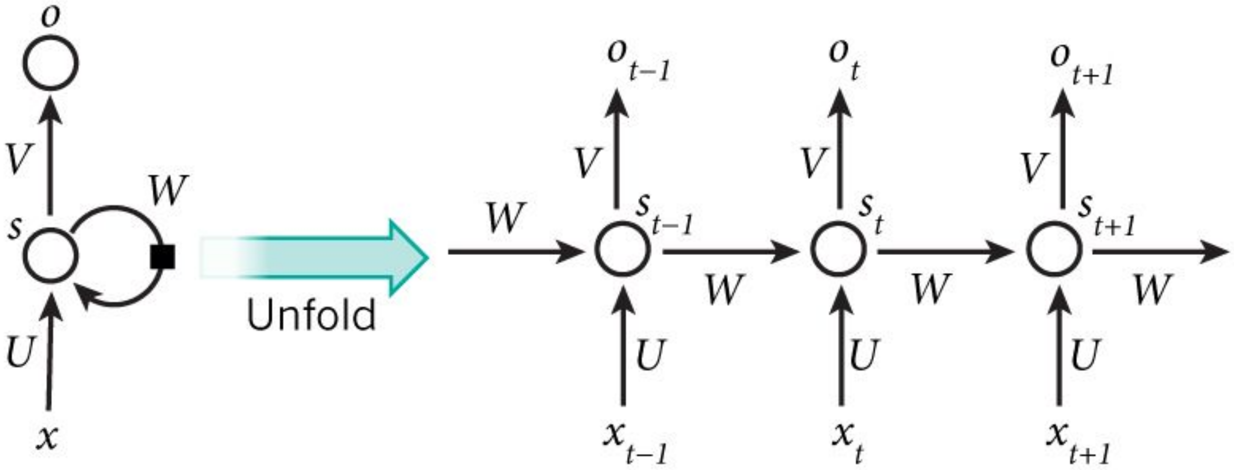
Пример тематической модели

#106: приложение + реклама + сервис + продукт + пользователь + платформа + ...
#107: проект + рамка + мрф + реализовать + кц + решение + данный + филиал + ...
#108: работа + затрата + качество + время + количество + сотрудник + расход + ...
#109: олег + александр + сергей + спасибо + тема + согласный + комментарий + ...
#110: приставка + компьютер + купить + пк + поставить + телевизор + питание + ...
#111: система + объект + управление + время + контроль + группа + прибор + ...
. . .

Convolutional Neural Network



Recurrent Neural Network



Доклад для разработчиков и бизнеса. Начнем с эволюции NLP: как произошел переход от Natural Language Processing к Natural Language Understanding, чему научились нейросети за 2018 год, и какие задачи над текстами ученые теперь могут решать автоматически. С разработчикам поговорим, как гуглить вопросы о машинной обработке текстов и сравним уже работающие методы NLP с самыми новыми. Для бизнеса расскажу, как включить критический подход в отношении машинного обучения, и как понять, нужно ли в оно в вашем бизнесе.