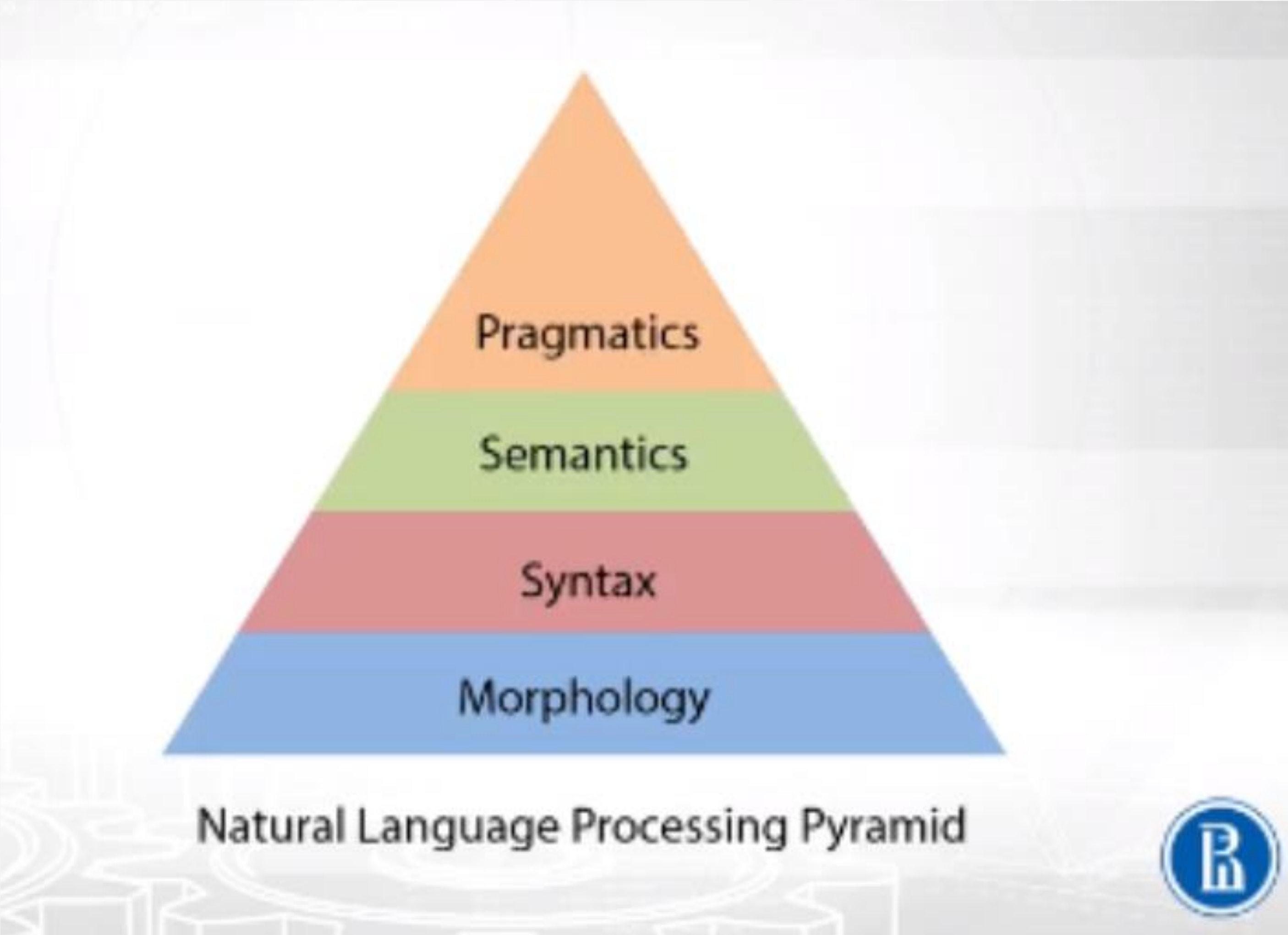
Curso de pocket NLP

Diego Lira
Alex Mansano
Pablo da costa
Vinicius F. Caridá

Natural Language Processing



Natural Language ML Model

API

Review

I had a great experience. The grocer was really helpful. One thing I would recommend is putting the price of the market price items online so I can look them up as they change.

Food & Grocery Retailers	0.53
Hospitality Industry > Food Service	0.53

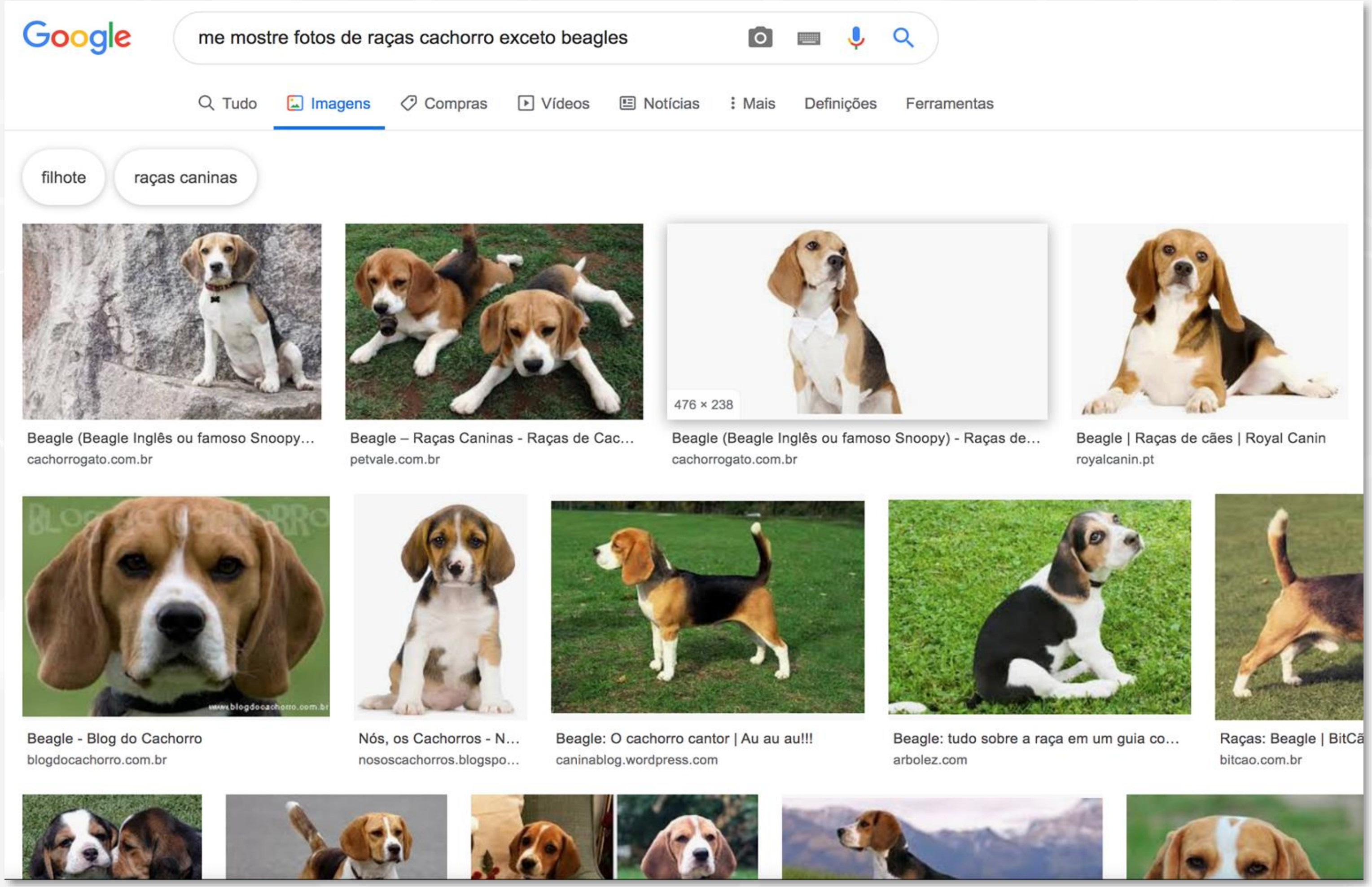
Custom

Review

I had a great experience. The grocer was really helpful. One thing I would recommend is putting the price of the market price items online so I can look them up as they change.

Great Service	0.88
Suggestion	0.84
Info Request	0.79

Entender não é tão simples



Entender não é tão simples

"Eu vi um homem na montanha com um telescópio"

1 — *Eu vi um homem. O homem estava na montanha. Eu estava com o telescópio.*

2 — *Eu vi um homem. Eu estava na montanha. O homem estava com o telescópio.*

3 — *Eu vi um homem. O homem estava na montanha. O homem estava com o telescópio.*

4 — *Eu vi um homem. Eu estava na montanha. Eu estava com o telescópio.*

Reconhecimento de Entidade Nomeada

Identificar entidades em um dados não estruturados

José trabalha para o Itaú, o maior banco da América Latina

pessoa

organização

organização

José trabalha para o **Itaú**, o maior **banco** da **América Latina**

lugar

Análise de Sentimento

Entender o sentimento expressado no texto

Excelente Atendimento!	Positivo
Resolveu meu problema, nada excepcional	Neutro
Péssima experiência	Negativo

Chat Bots

Sistemas capazes de interagir com usuário conversacionalmente

- Bom dia, Itaú. Quanto tenho de saldo na conta?
- Seu saldo é de R$ 1300, 00

Recuperação de Informação

Encontrar a resposta a uma pergunta em um texto ou base de conhecimento

Tradução

Traduzir textos de um idioma a outro

Português

Está bem, chega de exemplos sobre PLN...

↓

All right, enough examples about NLP...

Inglês

Descrição de Imagens

Descrever em texto o conteúdo de uma imagem

EXPLICAR PROCESSO DE TOKENIZACAO

Natural Language Processing

['Natural', 'Language', 'Processing']

Natural Language ML Model

- 👍 Good price! Quality not bad! I'm happy I bought it.
- 👎 Bad quality! I'm sad! I bought it I will return it.

Natural Language ML Model

Good price! Quality not bad! I'm happy I bought it.

Bad quality! I'm sad! I bought it I will return it.

fail	good	card	price	quality	bad	not	I	am	it	bought	return	happy	sad	will
0	1	0	1	1	1	1	1	1	1	1	0	1	0	0
0	0	0	0	1	1	0	1	1	1	1	1	0	1	1

Natural Language ML Model

- 👍 Good price! Quality not bad! I'm happy I bought it.
- 👎 Bad quality! I'm sad! I bought it I will return it.
- 👎 Price not good. Quality bad! I'm not happy I bought it.

Natural Language ML Model

- Good price! Quality not bad! I'm happy I bought it.
- Bad quality! I'm sad! I bought it I will return it.
- Price not good. Quality bad! I'm not happy I bought it.

fail	good	card	price	quality	bad	not	I	am	it	bought	return	happy	sad	will
0	1	0	1	1	1	1	1	1	1	1	0	1	0	0
0	0	0	0	1	1	0	1	1	1	1	1	0	1	1
0	1	0	1	1	1	1	1	1	1	1	0	1	0	0

Como representar um texto?

“O menino viu a menina com o binóculo”

Vetores binários

o	menino	viu	menina	binoculos	andar	fazer	correr
0	1	0	1	1	0	0	0

Frequência de termos

o	menino	viu	menina	binoculos	andar	fazer	correr
0.003	0.023	0.025	0.024	0.001	0.0	0.0	0.0

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

$$idf(w) = log(\frac{N}{df_t})$$

Problemas?

- casa → [0 0 0 0 0 0 0 0 0 1]
- apartamento → [0 1 0 0 0 0 0 0 0 0]

AND = 0

Representações binárias não permitem combinações complexas:

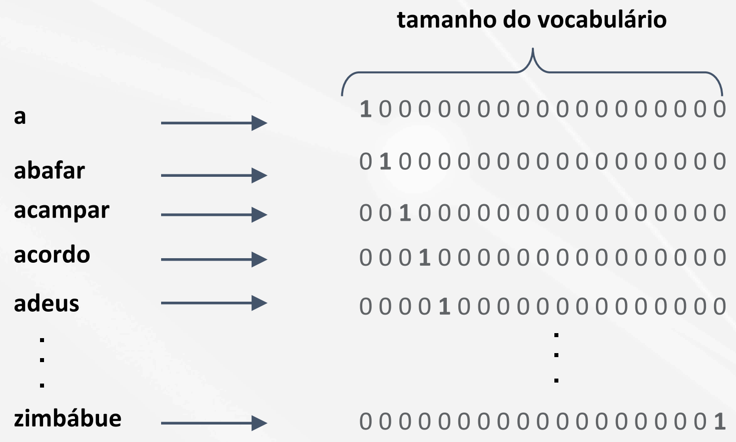
- Operações lógicas básicas como “and” e “not”, não são possíveis de serem operacionalizadas
- Não é possível manter a semântica das palavras em diferentes cenários de combinação;
- Alta esparsidade nos dados;

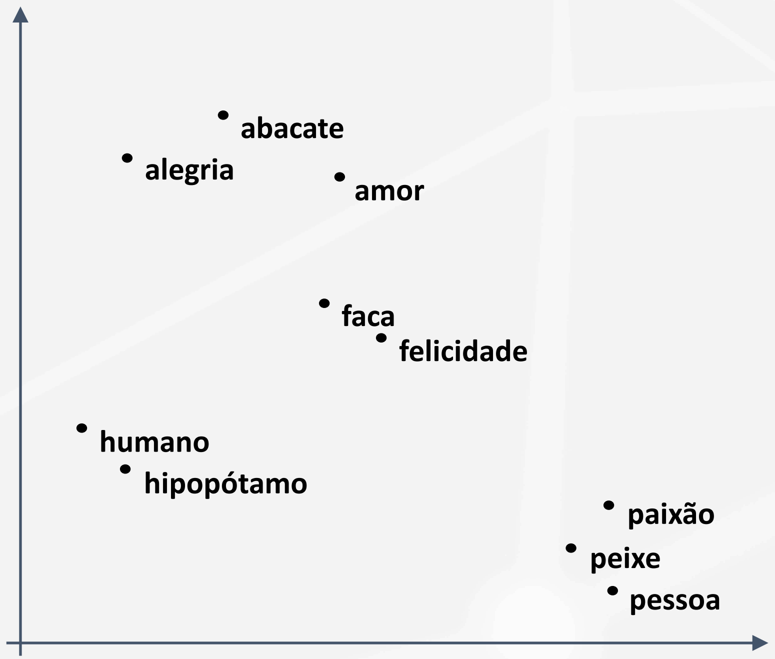
Solução?

- Representação densa e vetorial;
- Capturar representações distribuídas das palavras através de deep learning;

Problemas

- **Sem representatividade semântica**
- Vetores esparsos e de altíssima dimensão
- Necessidade de **mais dados rotulados** para generalizar
- **Não identifica** similaridade em palavras fora do vocabulário





Informação contextual

- Você pode capturar muita informação do contexto, em outras palavras uma média das palavras do contexto.
- “Você pode conhecer uma palavra pela companhia que ela mantem” (J. R. Firth 1957: 11)

government debt problems turning into **banking** crises as has happened in saying
that Europe needs unified **banking** regulation to replace the hodgepodge

essas palavras representam o contexto da palavra banco

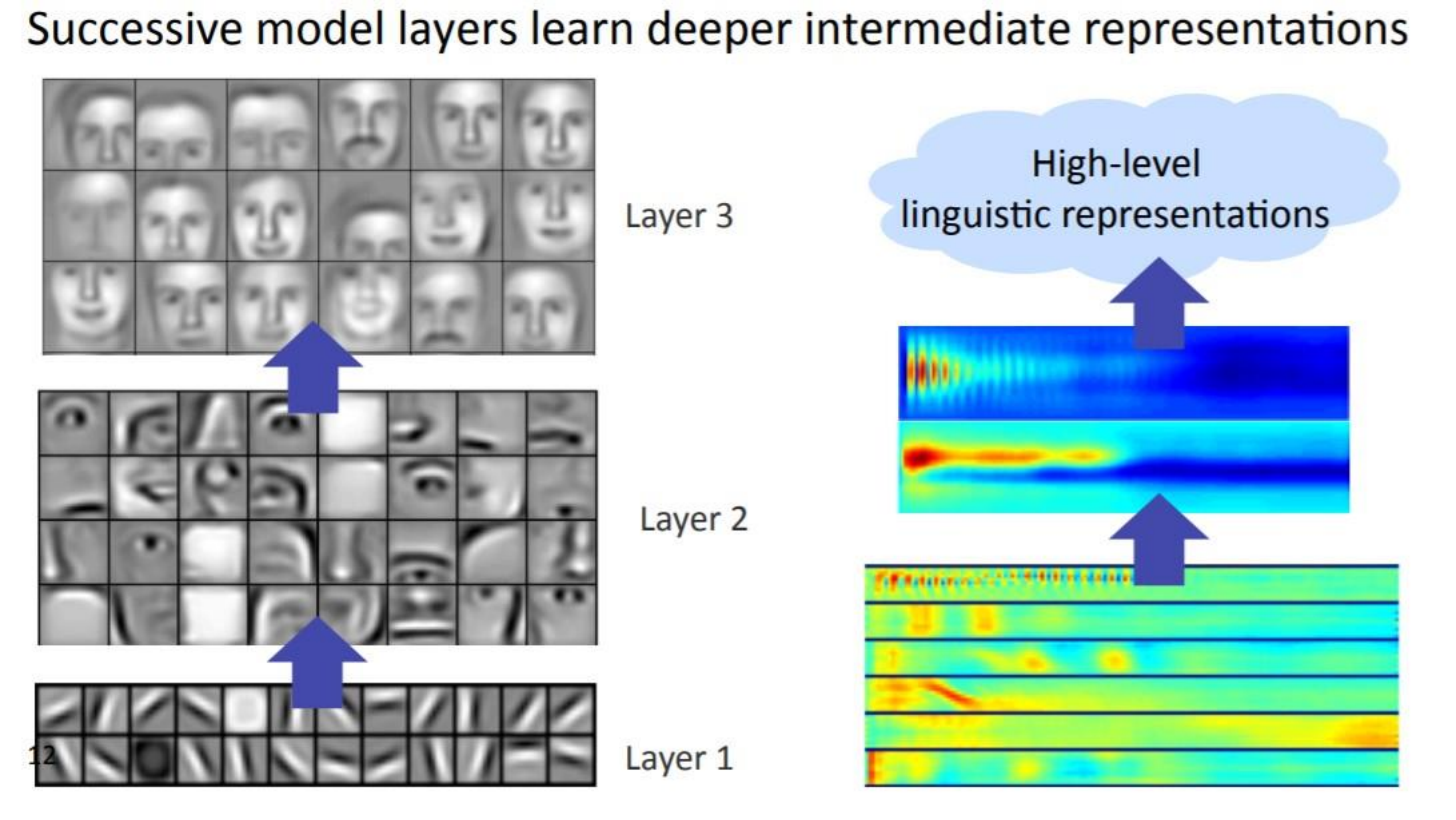
Informação contextual

- Representação **semântica**
- Vetores densos
- Generaliza palavras morfologicamente distintas

tamanho do embedding

a	.1234	.4763	.0948	.9384	.7363
abafar	.6463	.2347	.9434	.3232	.5832
acampar	.5546	.7293	.1039	.3220	.1342
acordo	.9482	.3293	.4201	.4403	.8494
adeus	.9584	.7564	.9924	.2134	.1123
⋮			⋮		
zimbábue	.9857	.7746	.2234	.3342	.9948

Mergulhando nas profundezas

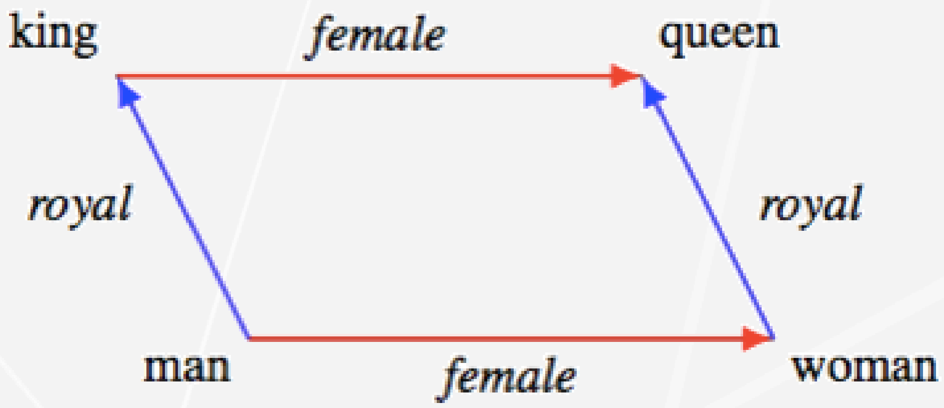


Modelos de Embedding

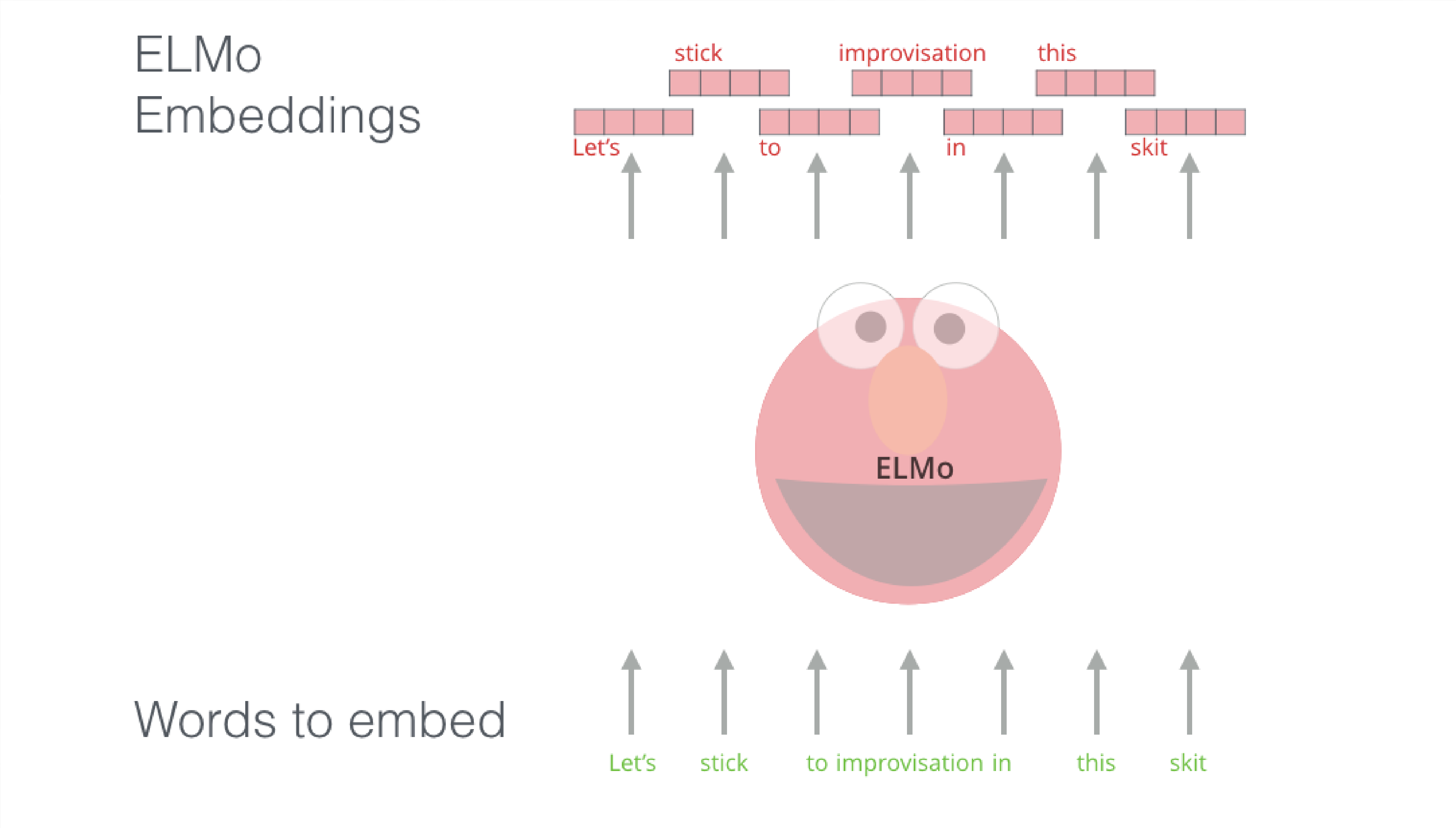
- **Word2Vec** — Mikolov et al, 2013
- **Glove** — Pennington et al., 2014
- **CoVe** — McCann et al., 2017
- **ULMFiT** — Howard et al.. 2018
- **ELMo** — Peters et al, 2018
- **GPT** — Radford et al.. 2018
- **BERT** — Devlin et al, 2018
- **GPT-2** — Radford et al.. 2019
- **XLM** — Lample et al.. 2019
- **XLNet** — Yang et al, 2019
- **RoBERTa** — Liu t al, 2019

Representações profundas

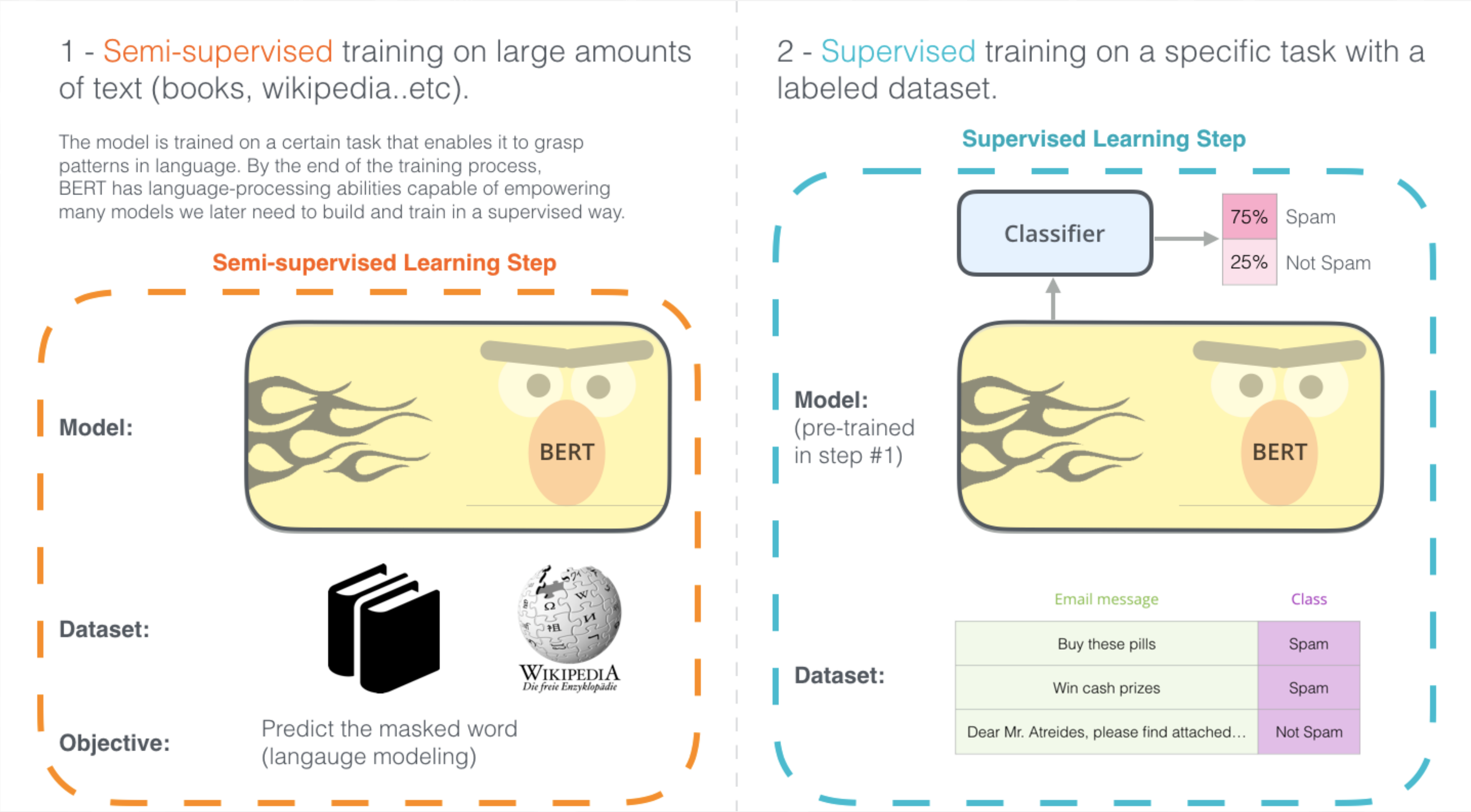
king − man + woman ≈ queen



Mergulhando nas profundezas



Mergulhando nas profundezas



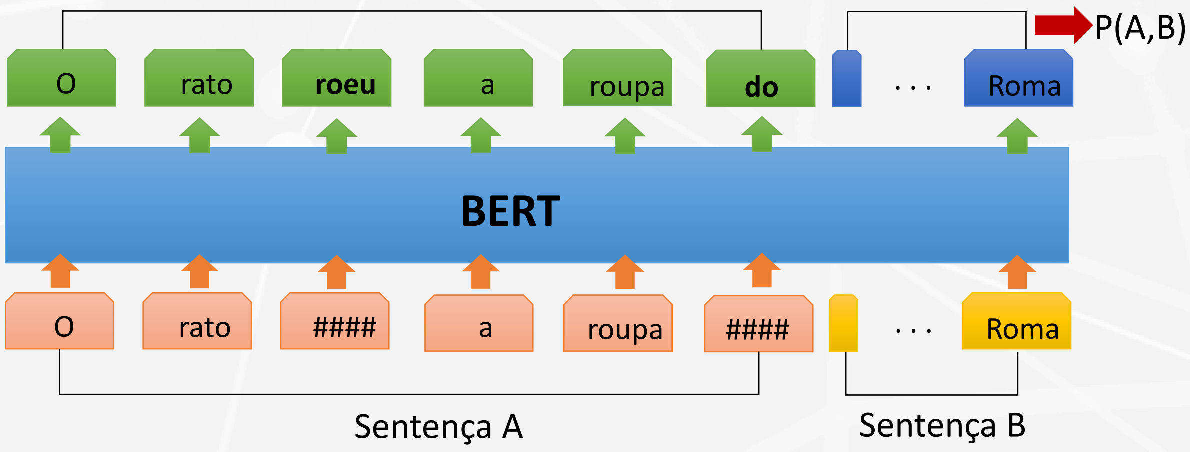
BERT

Dividido em duas subtarefas

- Pré-training
- Fine-tuning

235 milhões de parâmetros

Treinar em dados não rotulados

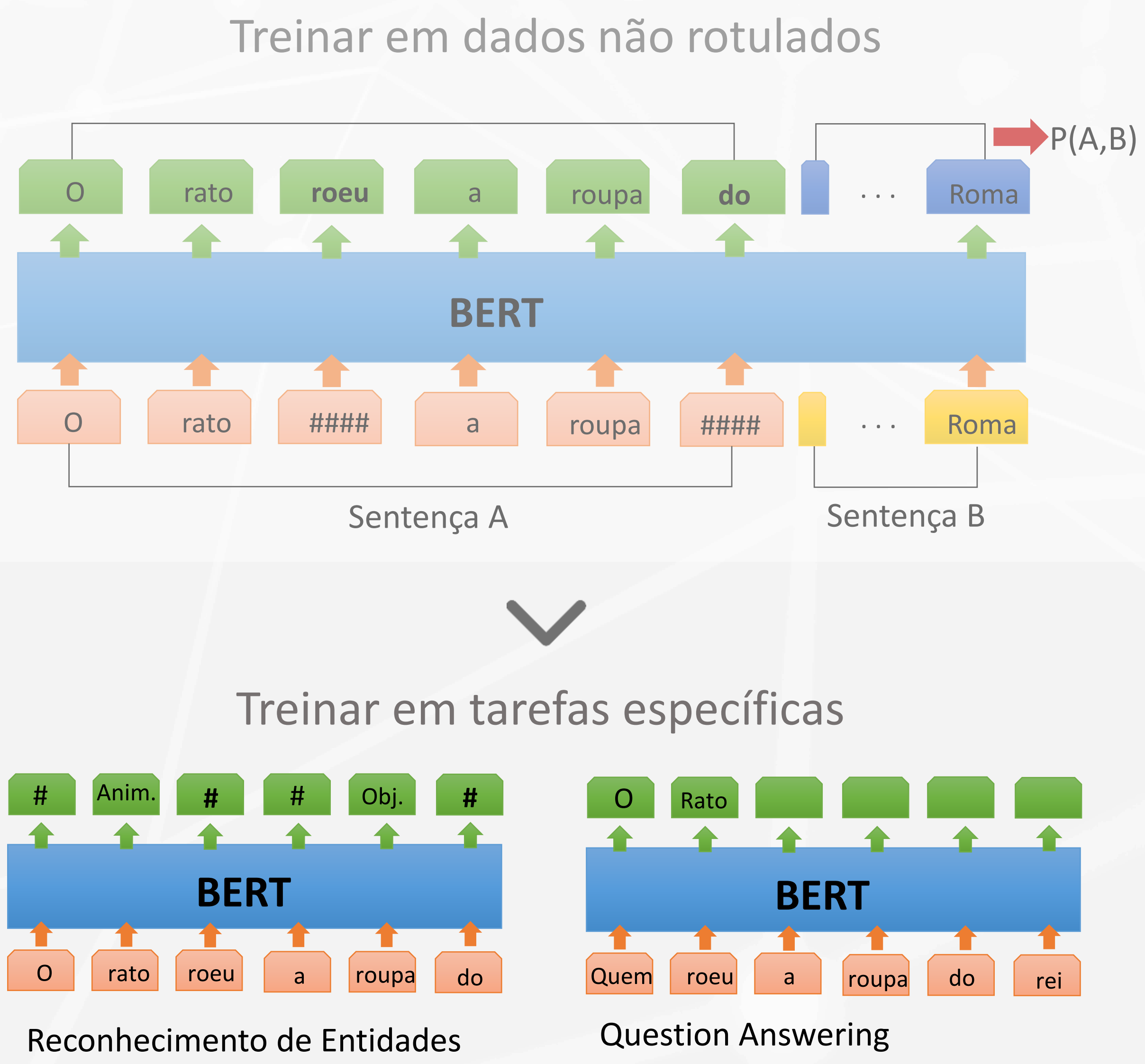


BERT

Dividido em duas subtarefas

- Pré-training
- Fine-tuning

235 milhões de parâmetros



BERT

Estado da arte em 11 tarefas de NLP

Question Answering **88.5 %**

Evolução **5 %**

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 May 21, 2019	XLNet (single model) *XLNet Team*	86.346	89.133
5 Apr 13, 2019	SemBERT(ensemble) *Shanghai Jiao Tong University*	86.166	88.886
5 May 14, 2019	SG-Net (ensemble) *Anonymous*	86.211	88.848
6 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
7 May 14, 2019	SG-Net (single model) *Anonymous*	85.229	87.926
8 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715
9 Apr 16, 2019	Insight-baseline-BERT (single model) *PAII Insight Team*	84.834	87.644
9 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615
9 Mar 13, 2019	BERT + ConvLSTM + MTL + Verifier (single model) *Layer 6 AI*	84.924	88.204
10 Apr 11, 2019	SemBERT (single model) *Shanghai Jiao Tong University*	84.800	87.864
11 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
12 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
12 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
13 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
13 Mar 20, 2019	Bert-raw (ensemble) *None*	83.604	86.036
14 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035
15 May 14, 2019	ATB (single model) *Anonymous*	82.882	86.002
15 Jan 14, 2019	BERT + MMFT + ADA (single model) *Microsoft Research Asia*	83.040	85.892
16 Jan 10, 2019	BERT + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	82.972	85.810
16 Feb 26, 2019	BERT with Something (ensemble) *Anonymous*	83.051	85.737
17 Feb 15, 2019	BERT + NeurQuRI (ensemble) *2SAH*	82.803	85.703
17 Feb 16, 2019	Bert-raw (ensemble) *None*	83.175	85.635
18 Dec 16, 2018	PAML+BERT (single model) *PINGAN GammaLab*	82.577	85.603
18 Mar 11, 2019	Bert-raw (ensemble) *None*	83.119	85.510
19 May 13, 2019	BERT-Base + QA Pre-training (single model) *Anonymous*	82.724	85.491
19 Feb 27, 2019	BERT + NeurQuRI (ensemble) *2SAH*	82.713	85.584
20 Nov 16, 2018	AoA + DA + BERT (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.374	85.310
20 Mar 03, 2019	Unnamed submission by null	82.431	85.178
21 Dec 12, 2018	BERT finetune baseline (single model) *Anonymous*	82.126	84.820
21 Feb 28, 2019	BERT_s (single model) *Anonymous*	81.979	84.846
21 Dec 10, 2018	Candi-Net+BERT (ensemble) *42Maru NLP Team*	82.126	84.624
22 Feb 28, 2019	BERT-large+UBFT (single model) *anonymous*	81.573	84.535
23 Feb 25, 2019	BERT with Something (single model) *Anonymous*	81.110	84.386
23 Feb 15, 2019	BERT + NeurQuRI (single model) *2SAH*	81.257	84.342
24 Nov 16, 2018	AoA + DA + BERT (single model) *Joint Laboratory of HIT and iFLYTEK Research*	81.178	84.251
25 Mar 07, 2019	BERT + UnAnsQ (single model) *Anonymous*	80.749	83.851
25 Mar 20, 2019	Bert-raw (single) *None*	80.693	83.922
25 Apr 07, 2019	BERT + AL (single model) *Anonymous*	80.715	83.827
26 Dec 19, 2018	Candi-Net+BERT (single model) *42Maru NLP Team*	80.659	83.562
27 Jan 09, 2019	Unnamed submission by null	80.512	83.539
28 Mar 11, 2019	Bert-raw (single) *None*	80.411	83.457
28 Jan 22, 2019	BERT + NeurQuRI (single model) *2SAH*	80.591	83.391
29 Apr 19, 2019	Unnamed submission by null	80.354	83.329
30 Feb 16, 2019	Bert-raw (single model) *None*	80.343	83.243
30 Jan 09, 2019	Unnamed submission by null	80.343	83.221
31 Feb 19, 2019	BERT + UDA (single model) *Anonymous*	80.005	83.208
31 Dec 03, 2018	PwP+BERT (single model) *AITRICS*	80.117	83.189
32 Apr 10, 2019	bert (single model) *vinda msqjmxx*	79.971	83.184
32 Apr 04, 2019	BISAN-CC (single model) *Seoul National University & Hyundai Motors*	80.208	83.149
32 Dec 05, 2018	Candi-Net+BERT (single model) *42Maru NLP Team*	80.388	82.908
32 Nov 08, 2018	BERT (single model) *Google AI Language*	80.005	83.061
33 Feb 12, 2019	BERT + Sparse-Transformer *single model*	79.948	83.023
34 Dec 06, 2018	NEXYS_BASE (single model) *NEXYS, DGIST R7*	79.779	82.912
34 Mar 07, 2019	BERT uncased (single model) *Anonymous*	79.745	83.020
35 Feb 01, 2019	{bert-finetuning} (single model) *ksai*	79.632	82.852
36 Mar 14, 2019	{Anonymous} (single model) *Anonymous*	78.876	82.524
36 Nov 09, 2018	L6Net + BERT (single model) *Layer 6 AI*	79.181	82.259
37 Mar 14, 2019	BISAN (single model) *Seoul National University & Hyundai Motors*	78.481	81.531
37 Apr 24, 2019	BERT + WIAN (ensemble) *Infosys Limited*	78.650	81.497
38 Jan 09, 2019	Unnamed submission by null	78.301	81.350
39 Dec 14, 2018	BERT+AC(single model) *Hithink RoyalFlush*	78.052	81.174
40 Nov 06, 2018	SLQA+BERT (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	77.003	80.209
41 Jan 05, 2019	synss (single model) *bert_finetune*	76.055	79.329
42 Dec 18, 2018	ARSG-BERT (single model) *TRINITI RESEARCH LABS, Active.ai* https://active.ai	74.746	78.227
42 Nov 05, 2018	MIR-MRC(F-Net) (single model) *Kangwon National University, Natural Language Processing Lab. & ForceWin, KP Lab.*	74.791	77.988
43 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	74.272	77.052
44 Dec 22, 2018	Unnamed submission by null	73.234	76.790
44 Dec 29, 2018	MMIPN *Single*	73.505	76.424
45 Apr 20, 2019	BERT-Base (single model) *Dining Philosophers*	73.099	76.236
46 Oct 12, 2018	YARCS (ensemble) *IBM Research AI*	72.670	75.507
47 Nov 14, 2018	BERT+Answer Verifier (single model) *Pingan Tech Olatop Lab*	71.666	75.457
47 Nov 10, 2018	Unnamed submission by null	72.580	75.075
48 Sep 17, 2018	Unet (ensemble) *Fudan University & Liulishuo Lab* https://arxiv.org/abs/1810.06638	71.417	74.869
49 Jan 19, 2019	{BERT-base} (single-model) *Anonymous*	70.763	74.449
49 Aug 28, 2018	SLQA+ (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	71.462	74.434
49 Apr 24, 2019	BERT-Base (single) *GreenflyAI* https://greenfly.ai	71.699	74.430
49 Aug 15, 2018	Reinforced Mnemonic Reader + Answer Verifier (single model) *NUDT* https://arxiv.org/abs/1808.05759	71.767	74.295
50 Sep 14, 2018	SAN (ensemble model) *Microsoft Business Applications AI Research* https://arxiv.org/abs/1712.03556	71.316	73.704

https://rajpurkar.github.io/SQuAD-explorer/, 2019-06-12

SQuAD

Stanford **Qu**estion **A**nswering **D**ataset: base de dados de leitura e compreensão de texto

100.000+ perguntas sobre artigos presentes na Wikipedia

Respostas são trechos da Wikipedia

https://rajpurkar.github.io/SQuAD-explorer/

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O 2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

The atomic number of the periodic table for oxygen?

Ground Truth Answers: 8

Which gas makes up 20.8% of the Earth's atmosphere?

Ground Truth Answers: Diatomic oxygen

Roughly, how much oxygen makes up the Earth crust?

Ground Truth Answers: almost half

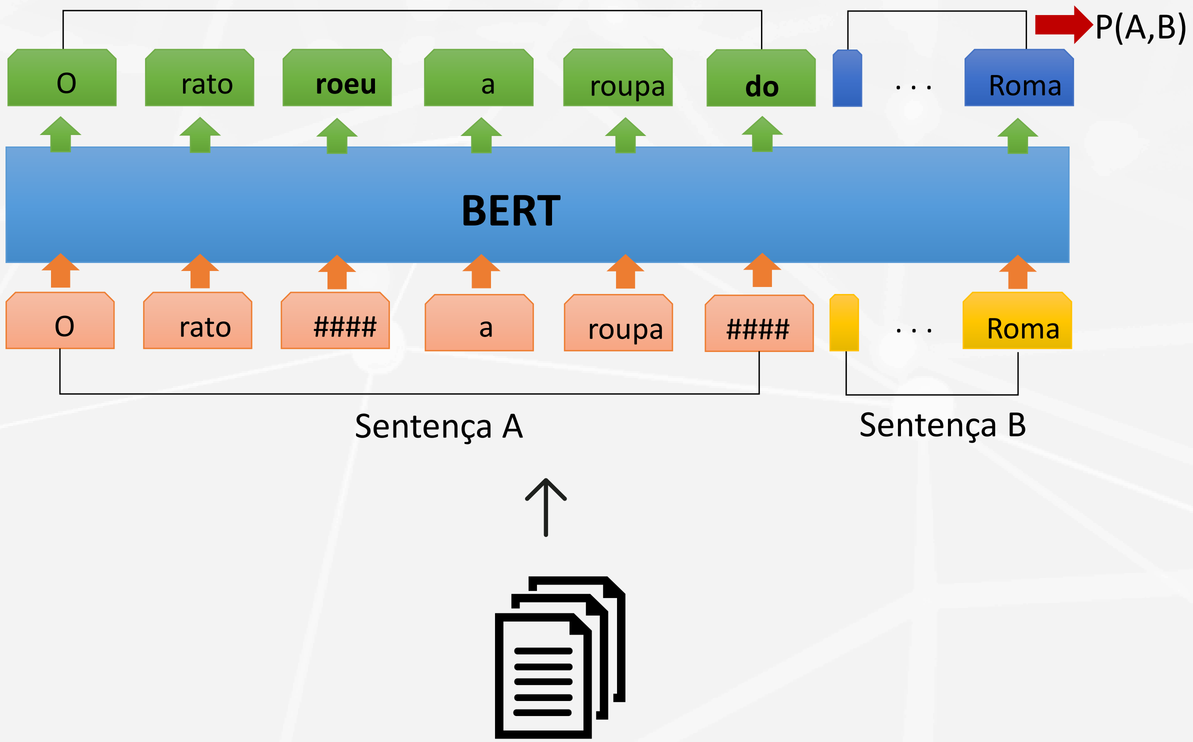
GPT-2

Estado da arte em geração de texto

Transformer com mais camadas e muito mais dados

Resultados promissores sem treinar em tarefas específicas

1.5 Bilhões de parâmetros



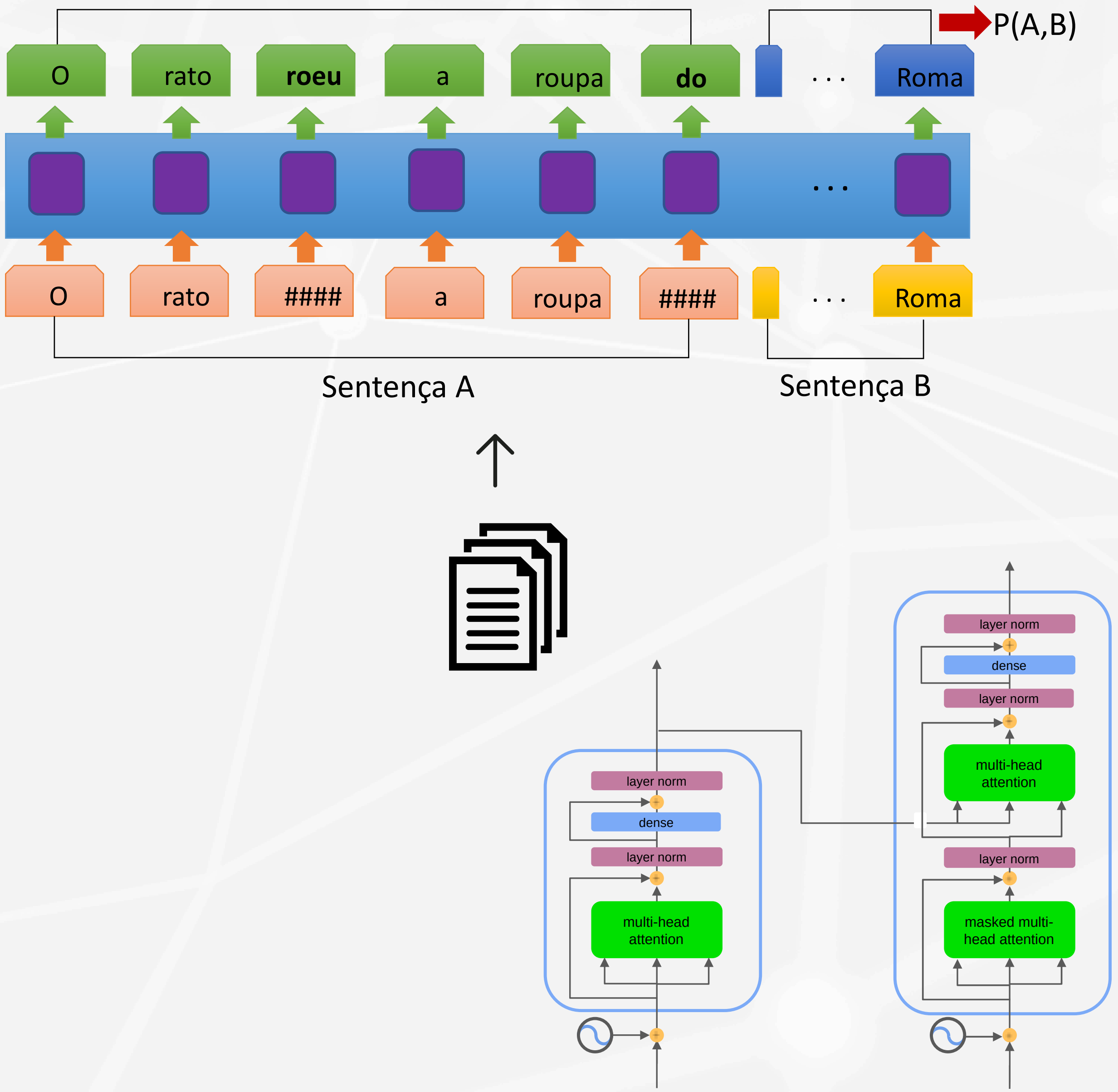
GPT-2

Estado da arte em geração de texto

Transformer com mais camadas e muito mais dados

Resultados promissores sem treinar em tarefas específicas

1.5 Bilhões de parâmetros



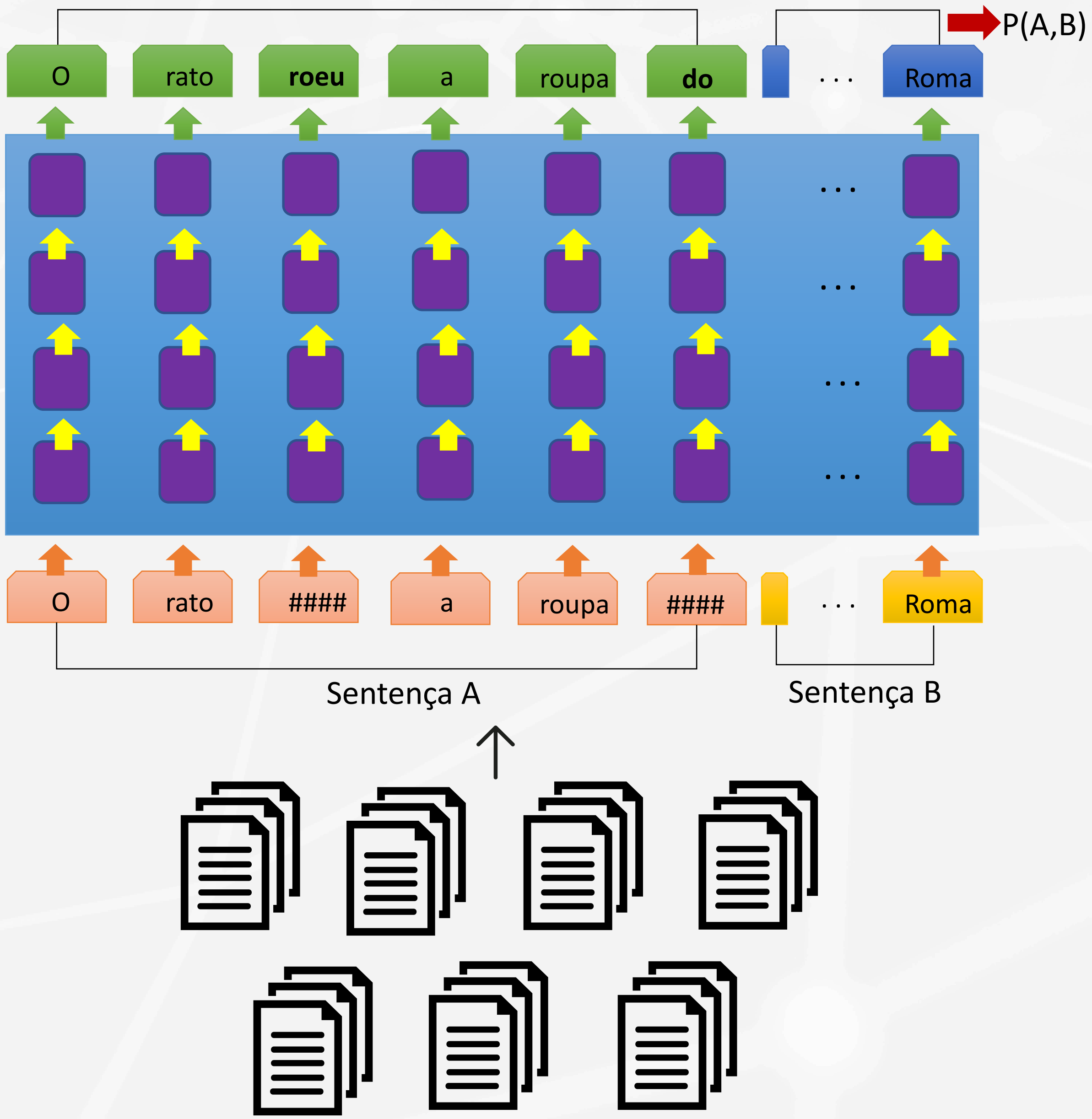
GPT-2

Estado da arte em geração de texto

Transformer com mais camadas e muito mais dados

Resultados promissores sem treinar em tarefas específicas

1.5 Bilhões de parâmetros



GPT-2

Estado da arte em geração de texto

Transformer com mais camadas e muito mais dados

Resultados promissores sem treinar em tarefas específicas

1.5 Bilhões de parâmetros

DATASET	TASK	SOTA	OURS
SNLI	Textual Entailment	89.3	**89.9**
MNLI Matched	Textual Entailment	80.6	**82.1**
MNLI Mismatched	Textual Entailment	80.1	**81.4**
SciTail	Textual Entailment	83.3	**88.3**
QNLI	Textual Entailment	82.3	**88.1**
RTE	Textual Entailment	**61.7**	56.0
STS-B	Semantic Similarity	81.0	**82.0**
QQP	Semantic Similarity	66.1	**70.3**
MRPC	Semantic Similarity	**86.0**	82.3
RACE	Reading Comprehension	53.3	**59.0**
ROCStories	Commonsense Reasoning	77.6	**86.5**
COPA	Commonsense Reasoning	71.2	**78.6**
SST-2	Sentiment Analysis	**93.2**	91.3
CoLA	Linguistic Acceptability	35.0	**45.4**
GLUE	Multi Task Benchmark	68.9	**72.8**

https://openai.com/blog/language-unsupervised/

GPT-2

Estado da arte em geração de texto

Transformer com mais camadas e muito mais dados

Resultados promissores sem treinar em tarefas específicas

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

https://pbs.twimg.com/media/DzYpsJOU0AA1PO9.png

Teste em https://talktotransformer.com

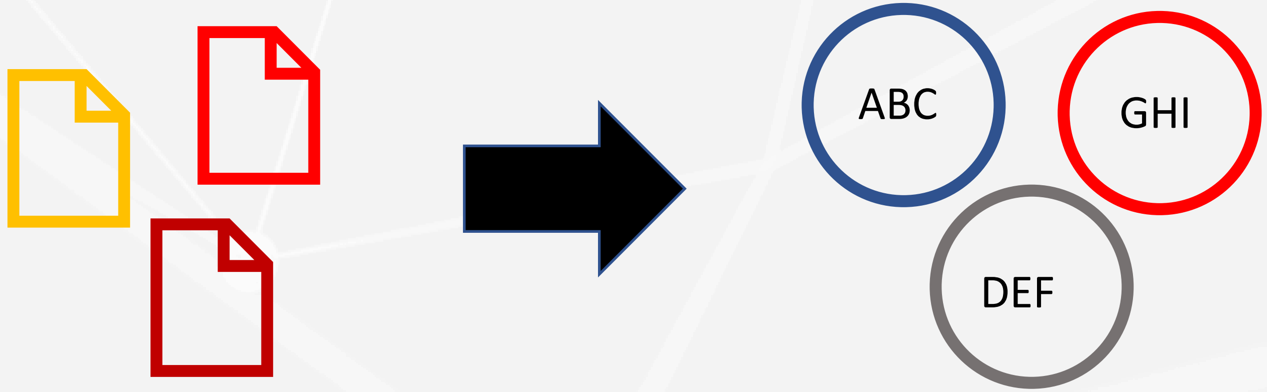
Modelagem de Tópicos - LDA

Modelagem de tópicos é um tipo de modelagem estatística para descobrir os tópicos abstratos que ocorrem em uma coleção de documentos.

Existem dois tipos de **abordagens para a modelagem** de tópicos:

Probabilística – Modela a aparição de palavras por tópicos por texto usando uma distribuição de probabilidade.

Matrix Factorization – Fatora as componentes utilizando uma decomposição de matrizes (similar a recomendação de filmes)



Modelagem de Tópicos – LDA – como funciona?