PRE-TRAINING IN NLP

Piotr Rybak

Agenda

What is Transfer learning?

Agenda

What is **Transfer learning**?

How does **BERT** work?

Agenda

What is Transfer learning?

How does BERT work?

Evaluation of pre-trained models

Agenda

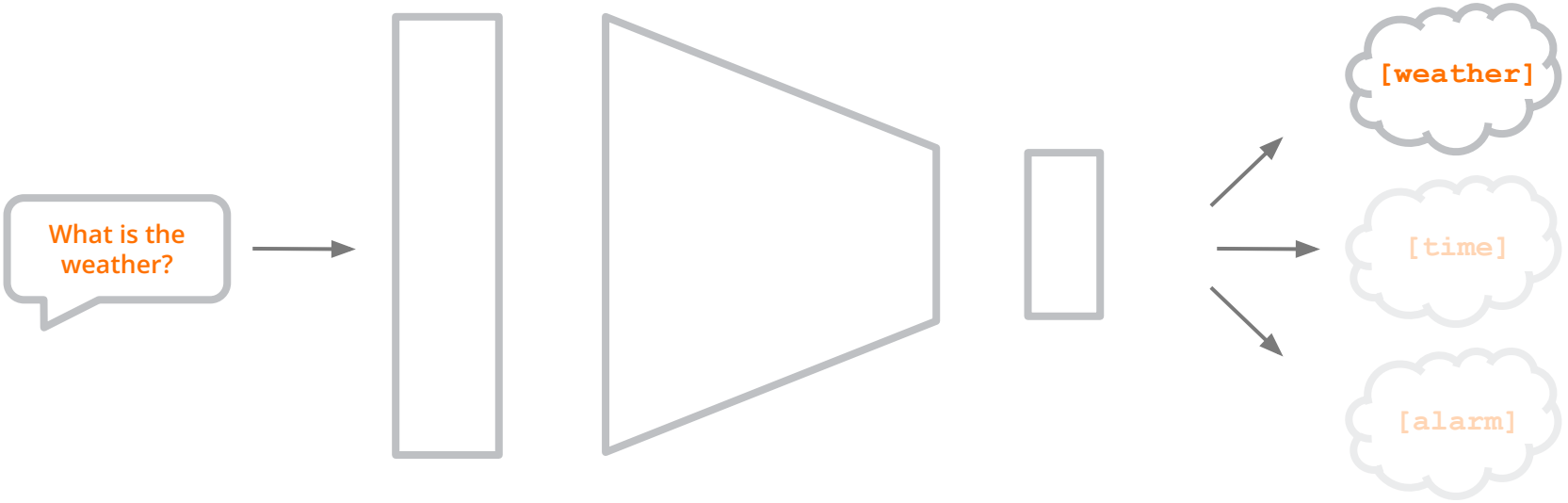
What is Transfer learning?

How does BERT work?

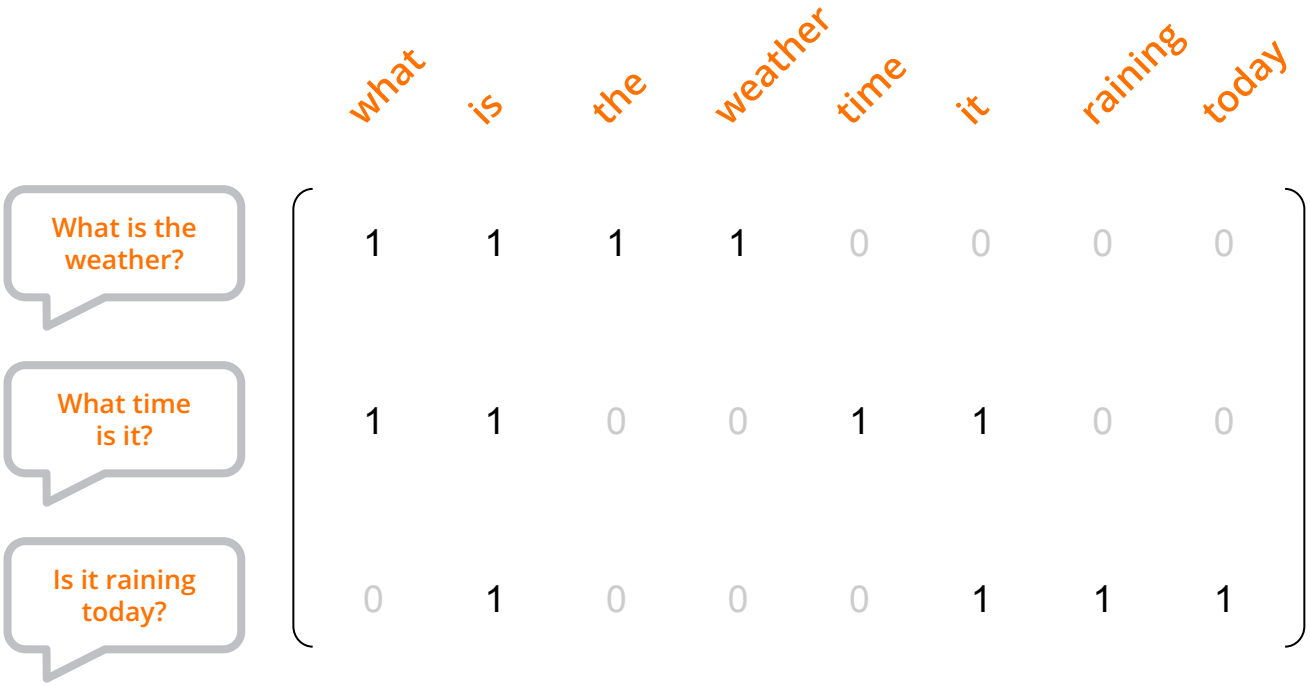
Evaluation of pre-trained models

Many improvements of BERT model

Transfer Learning


What is the weather?
[weather]
[time]
[alarm]
Input
Encoder
Classifier

Bag-of-words



Bag-of-words

Bag-of-words

Synonyms: weather vs temperature

Bag-of-words

Synonyms: weather vs temperature

Large vocabulary: many weights to learn

Bag-of-words

Synonyms: weather vs temperature

Large vocabulary: many weights to learn

Word **order** matters

Bag-of-words

Synonyms: weather vs temperature

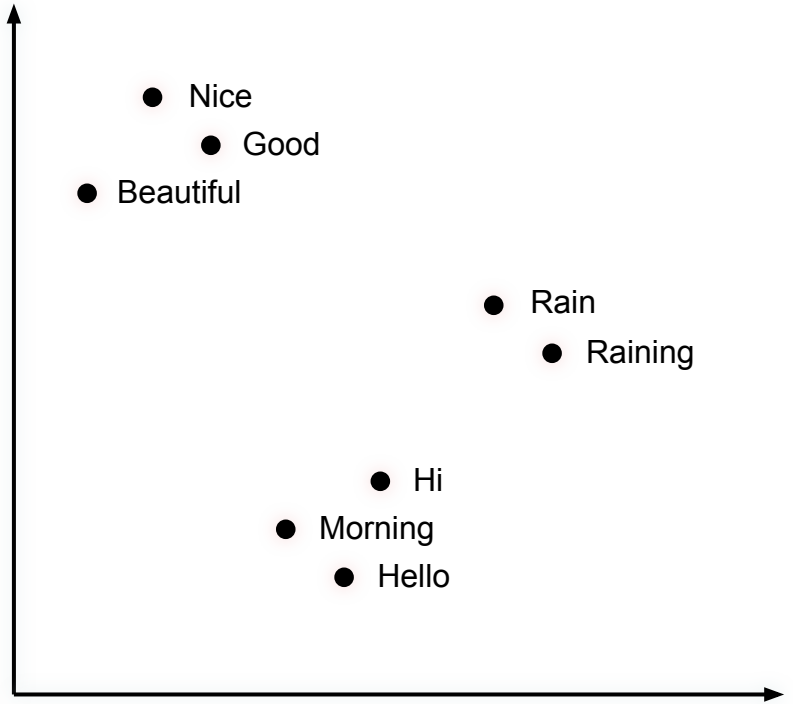
Large vocabulary: many weights to learn

Word **order** matters

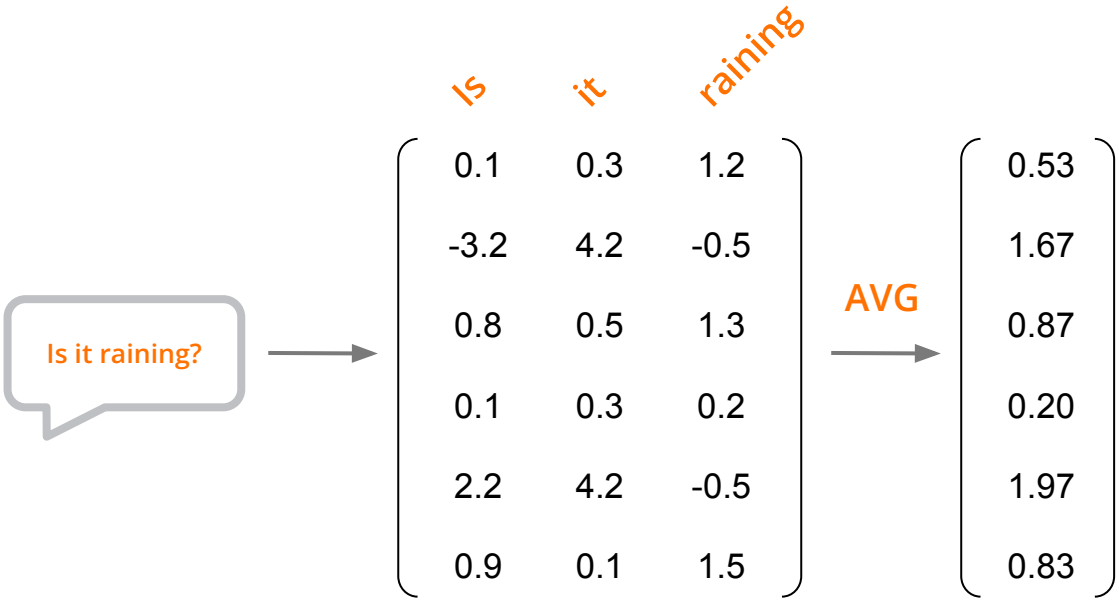
No reusability

Word Embeddings

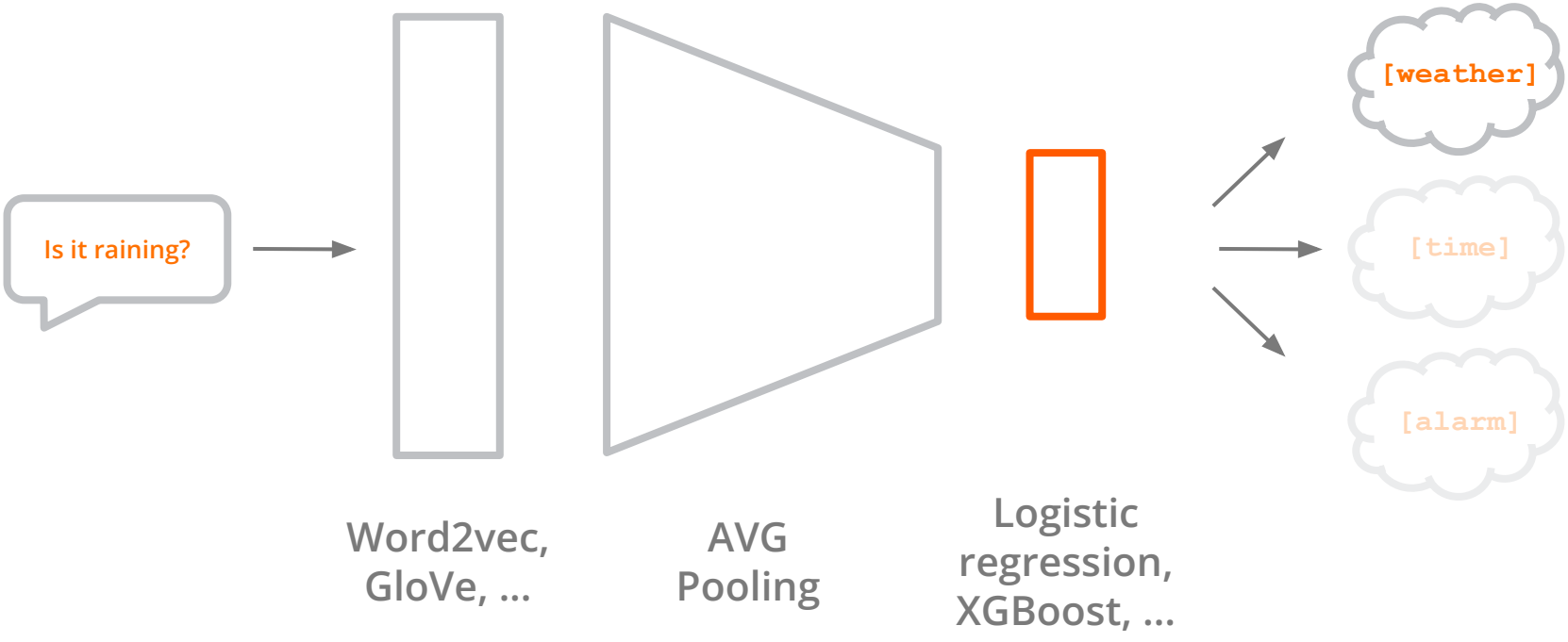
Is	it	raining
0.1	0.3	1.2
-3.2	4.2	-0.5
0.8	0.5	1.3
0.1	0.3	0.2
2.2	4.2	-0.5
0.9	0.1	1.5



Word Embeddings



Word Embeddings



Word Embeddings

Word **order** matters

Word Embeddings

Word **order** matters

Words have **multiple meanings**

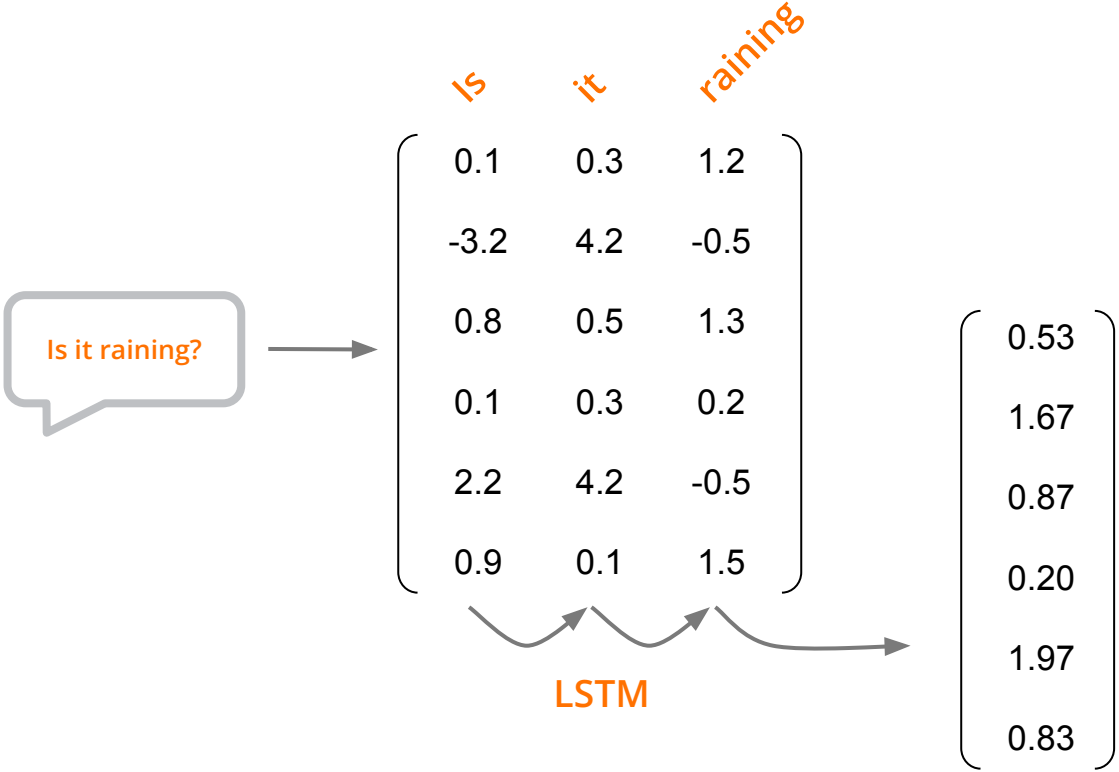
Word Embeddings

Word order matters

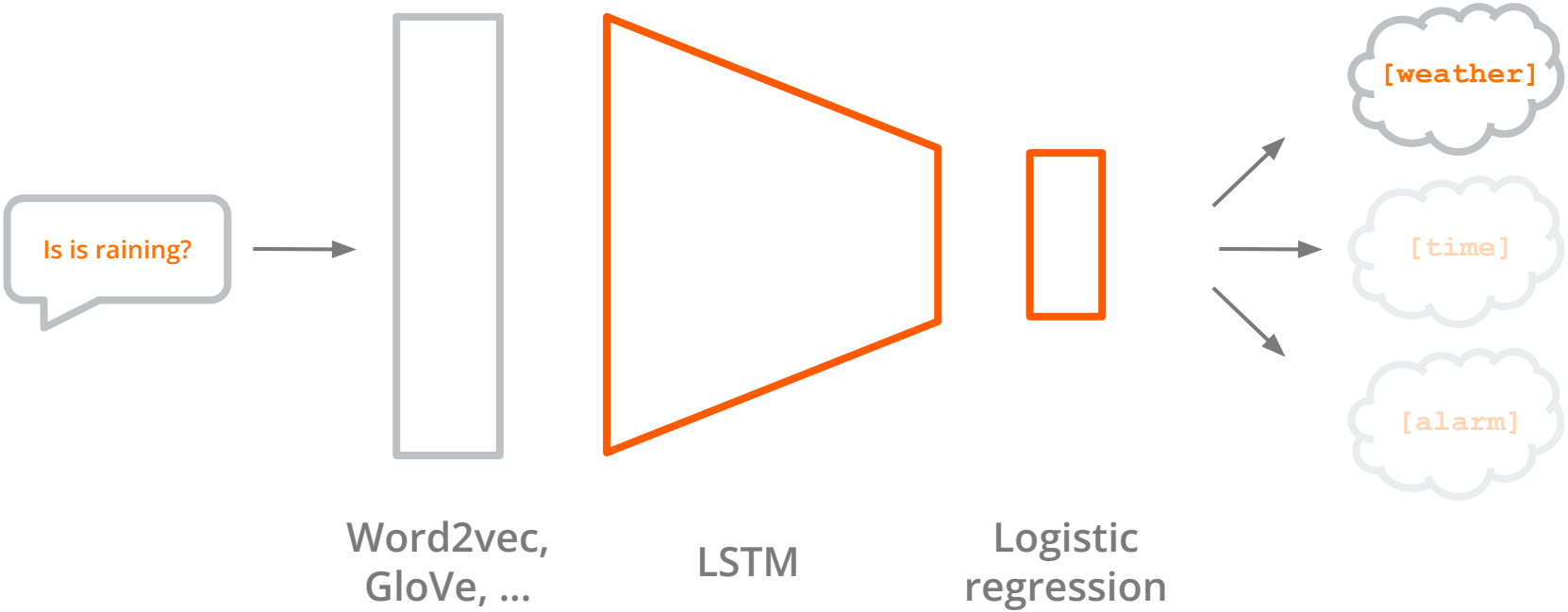
Words have multiple meanings

Some reusability

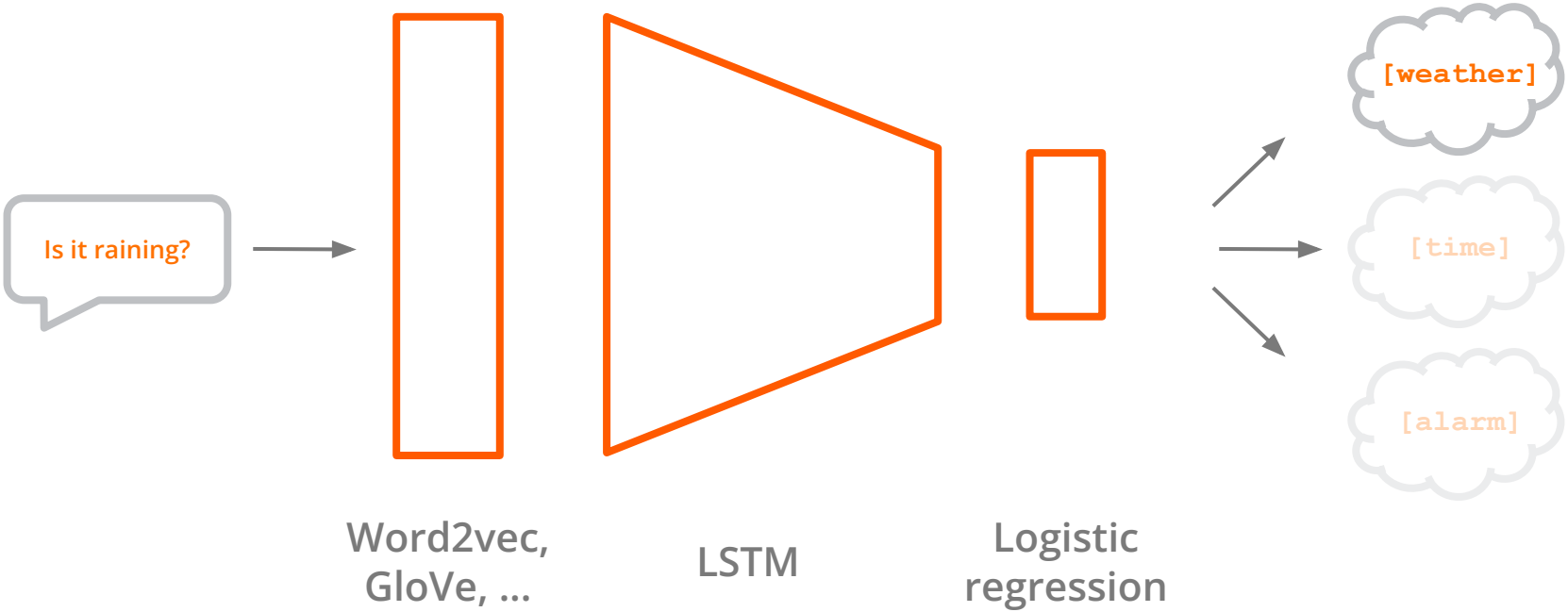
Word Embeddings + LSTM



Word Embeddings + LSTM



Fine-tuned Word Embeddings + LSTM



Fine-tuned Word Embeddings + LSTM

Even more weights to train

Fine-tuned Word Embeddings + LSTM

Even more weights to train

Easy to **overfit**

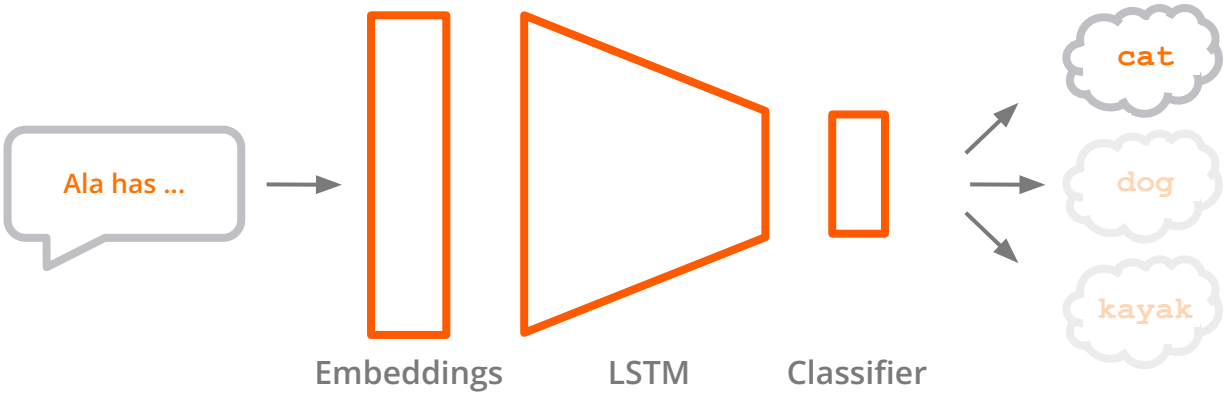
Fine-tuned Word Embeddings + LSTM

Even more weights to train

Easy to **overfit**

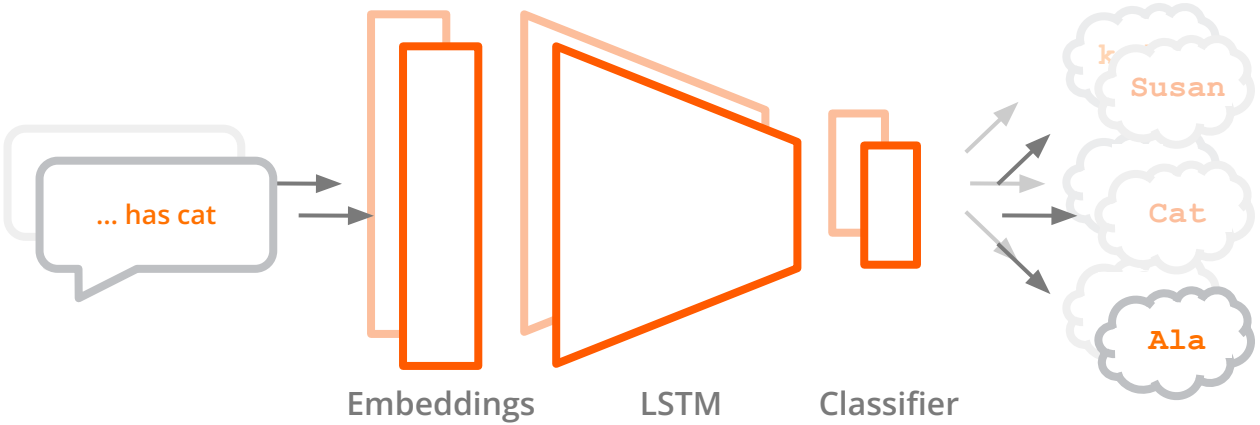
Still **not much reusability** due to too specific task

ELMo: Forward Language Modeling



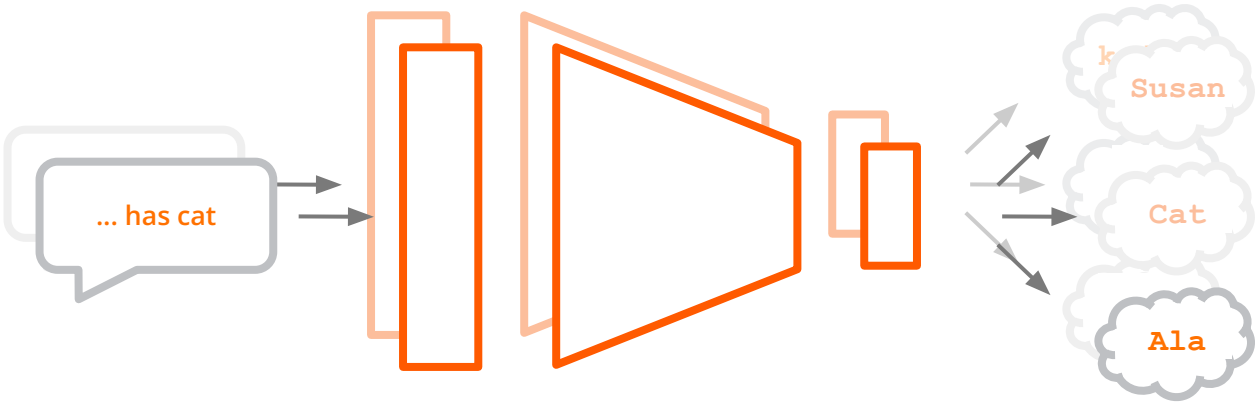
Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Backward Language Modeling



Matthew E. Peters et al., *Deep contextualized word representations*, 2018

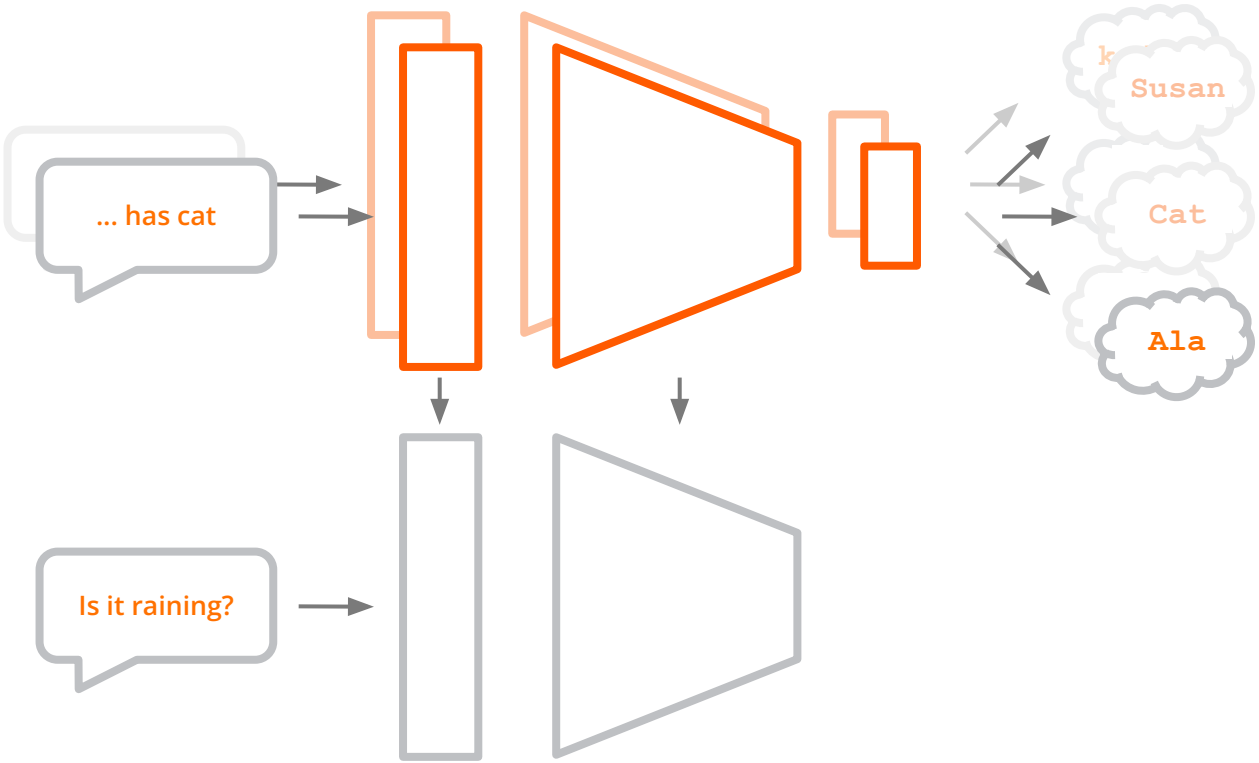
ELMo: Forward + Backward Language Modeling



Is it raining?

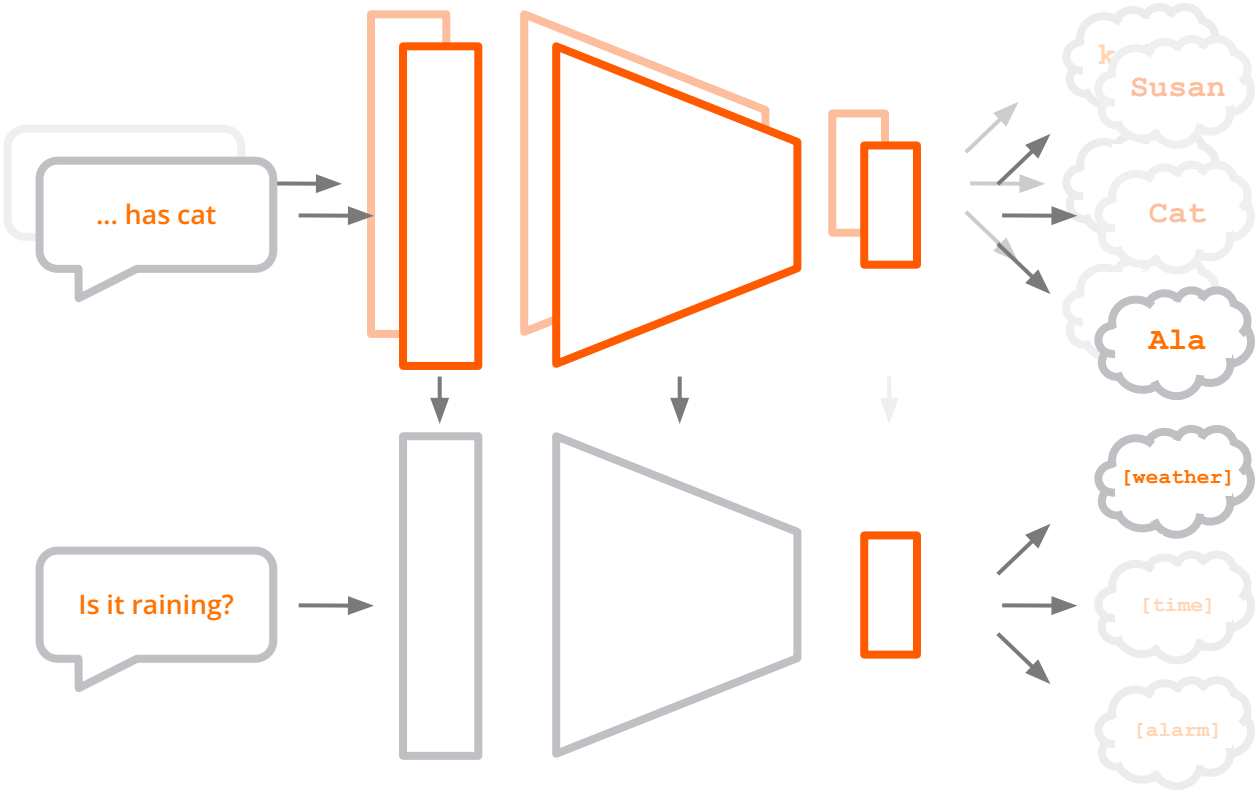
Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Forward + Backward Language Modeling



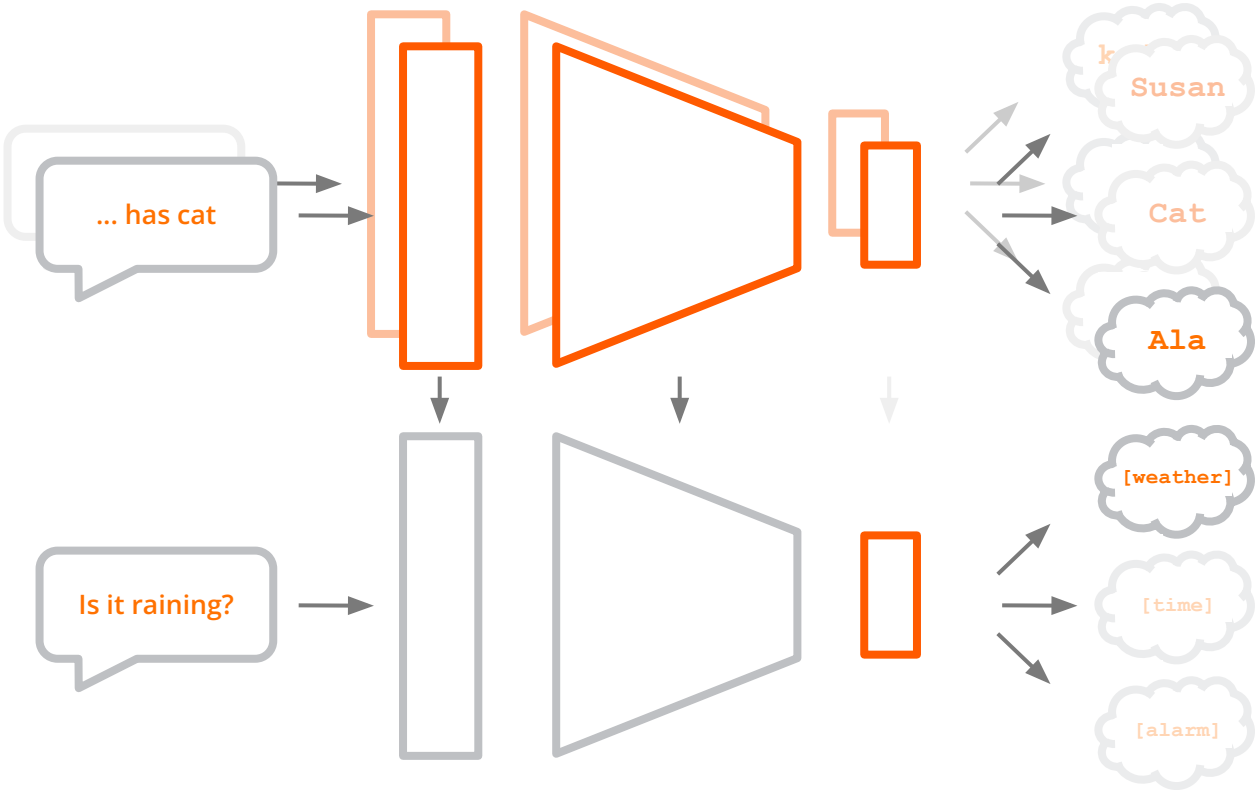
Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Forward + Backward Language Modeling



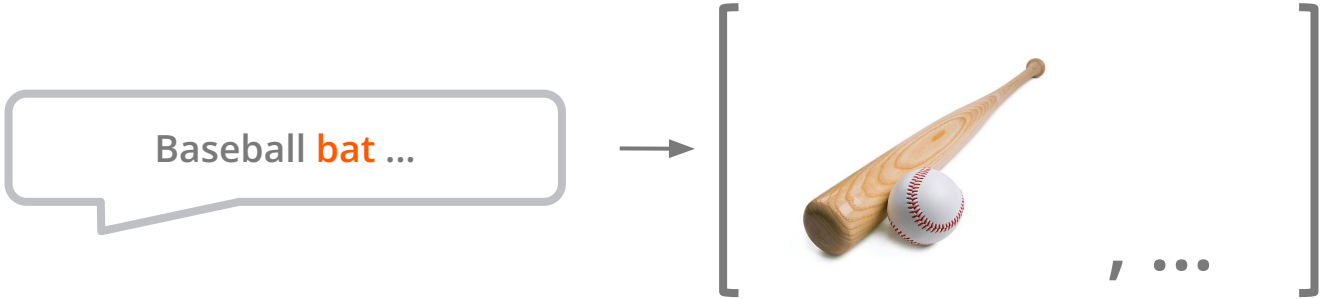
Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Forward + Backward Language Modeling

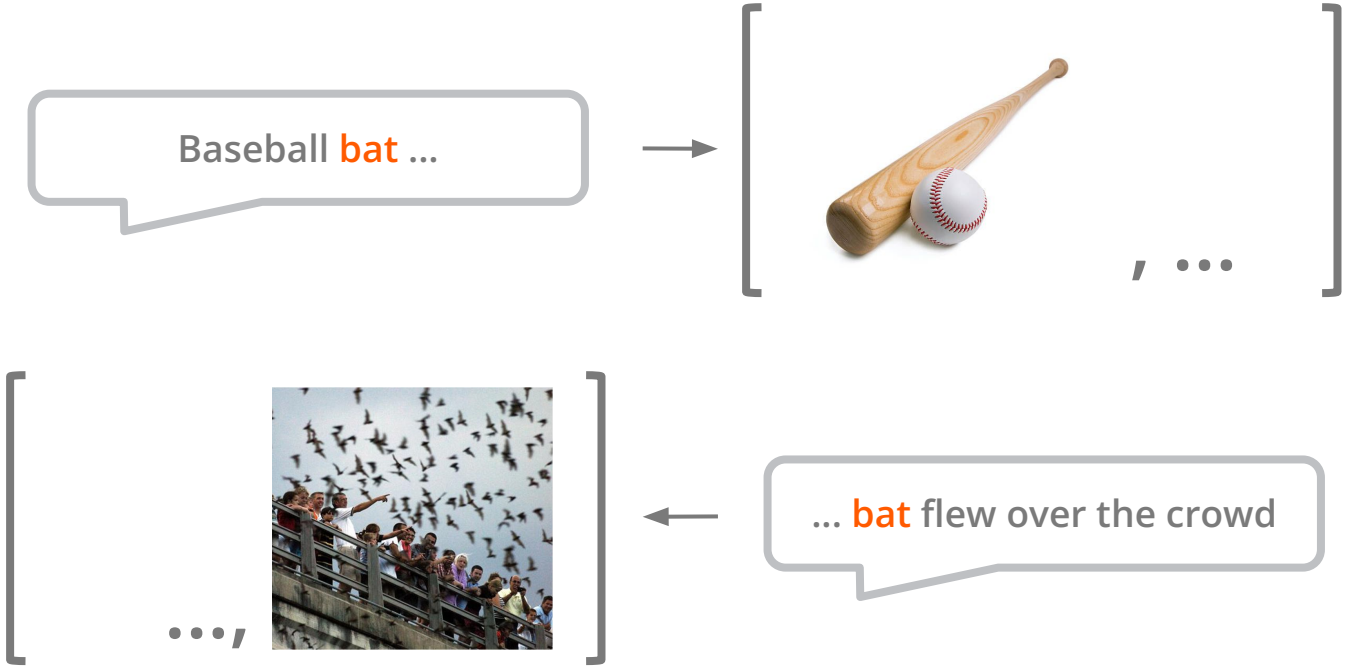


Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Forward + Backward Language Modeling



ELMo: Forward + Backward Language Modeling



ELMo: Forward + Backward Language Modeling

Baseball **bat** flew over the crowd

ELMo: Forward + Backward Language Modeling

Baseball **bat** flew over the crowd

BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

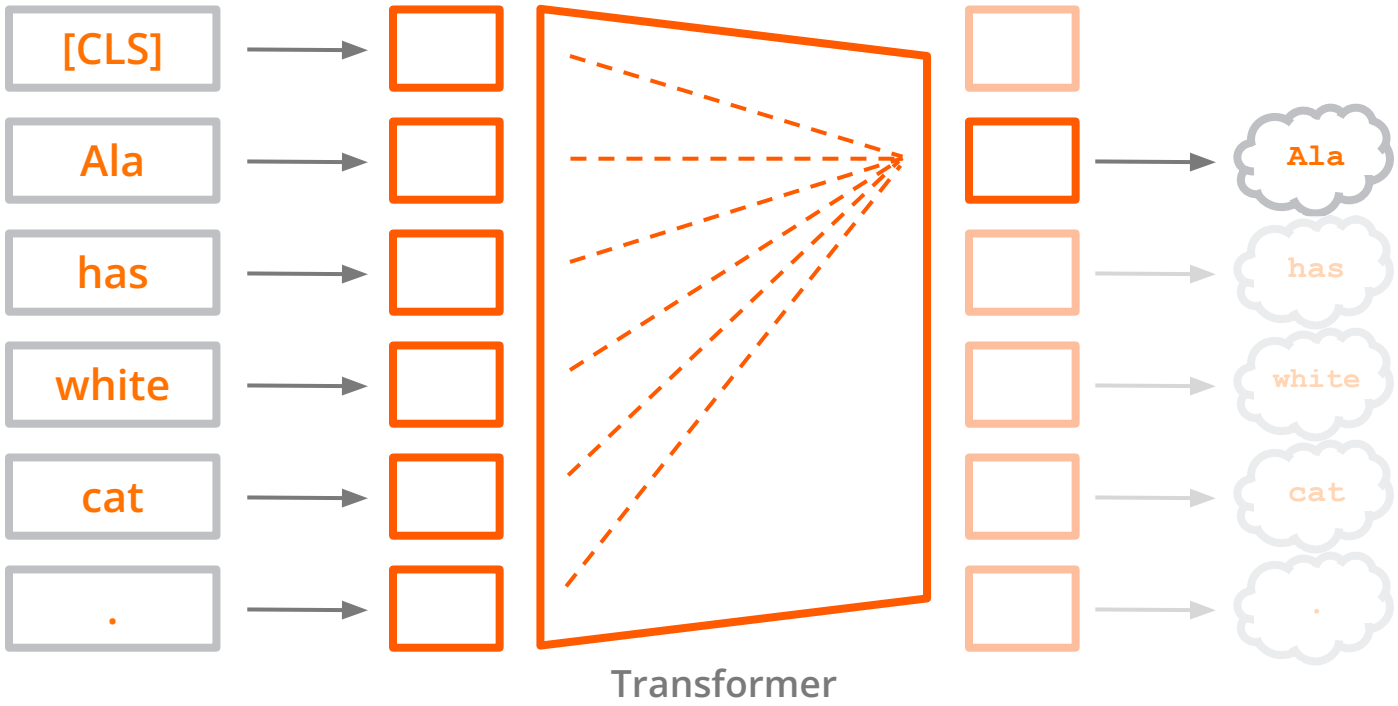
Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, *without* substantial task-specific architecture

models are required to produce fine-grained output at the token-level.

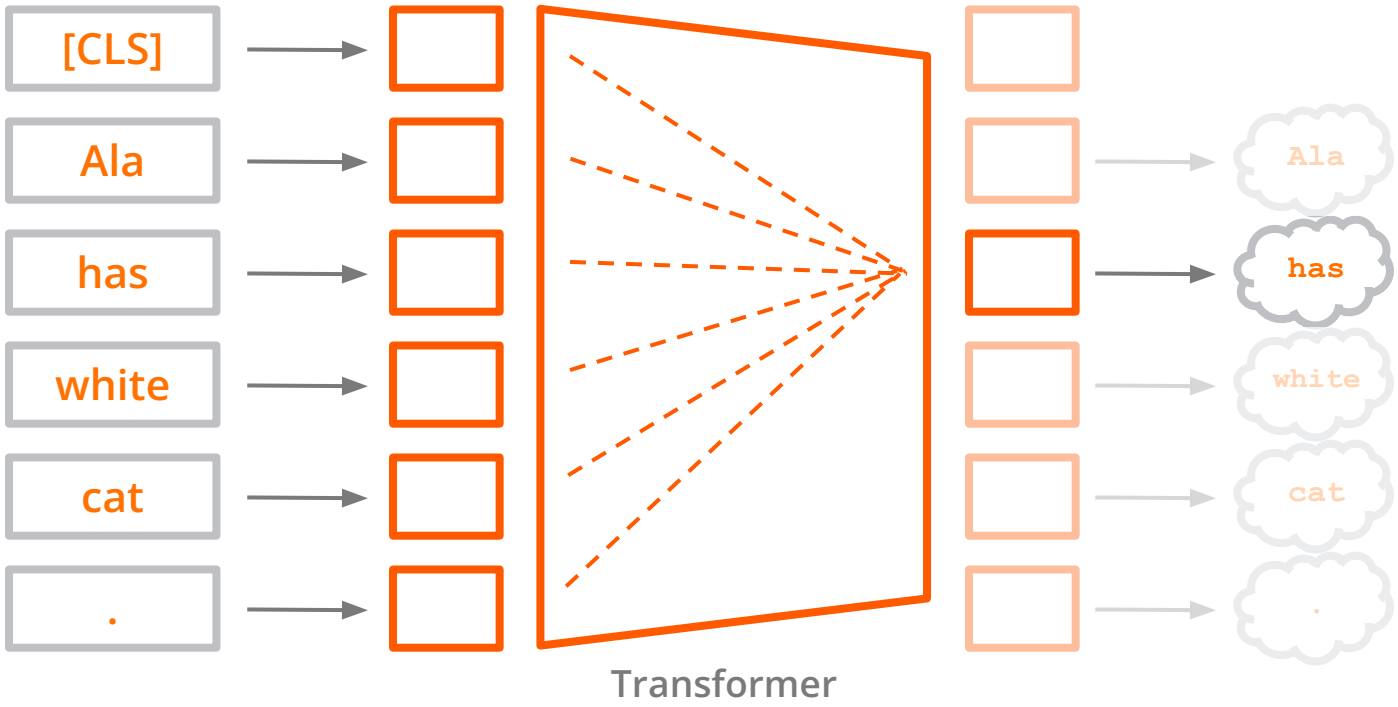
There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018), uses tasks-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning the pre-trained parameters. In previous work, both ap-

BERT: Masked Language Modeling



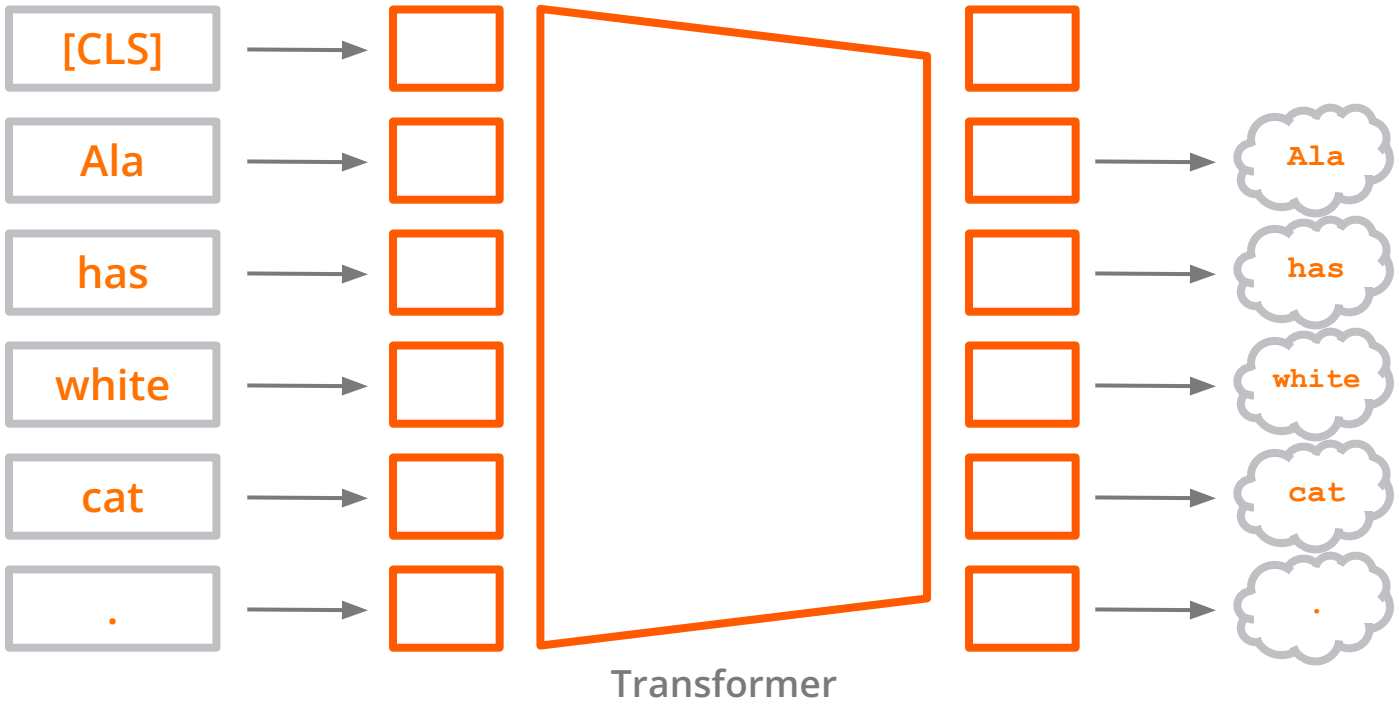
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



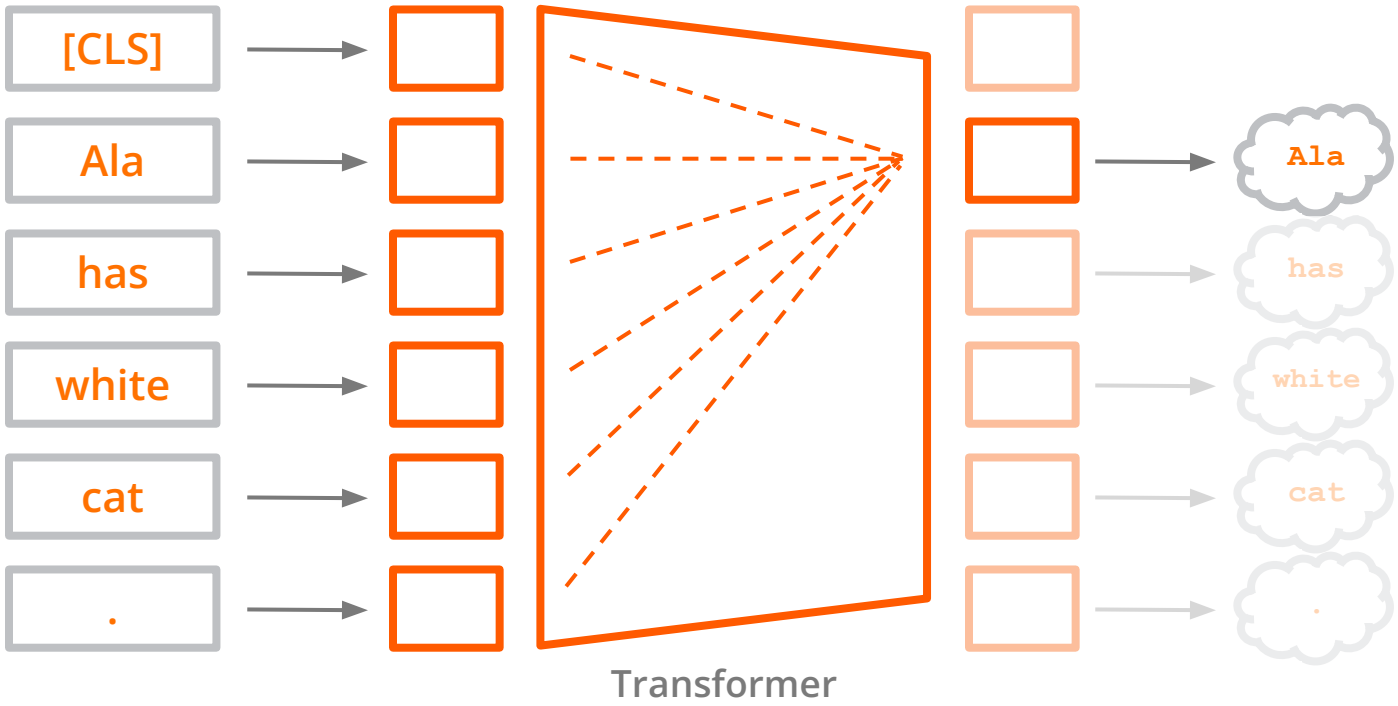
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



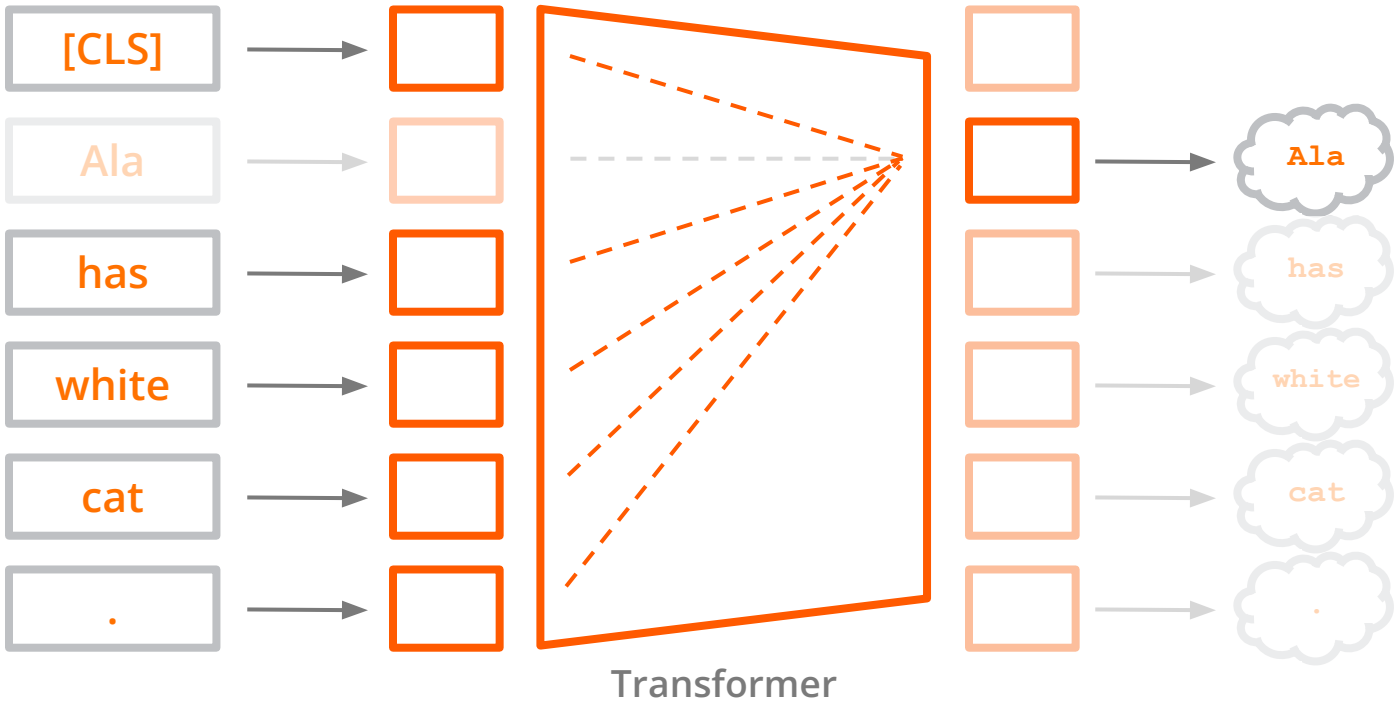
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



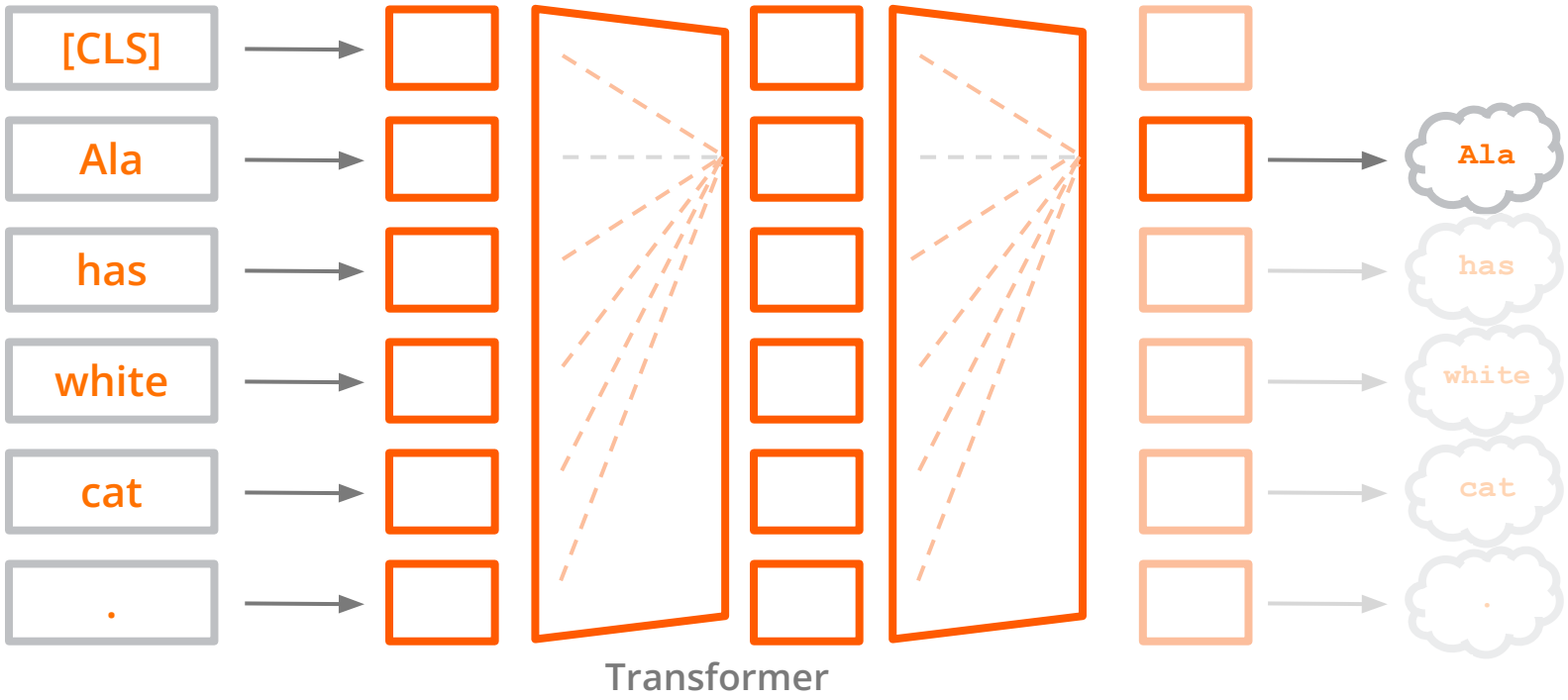
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



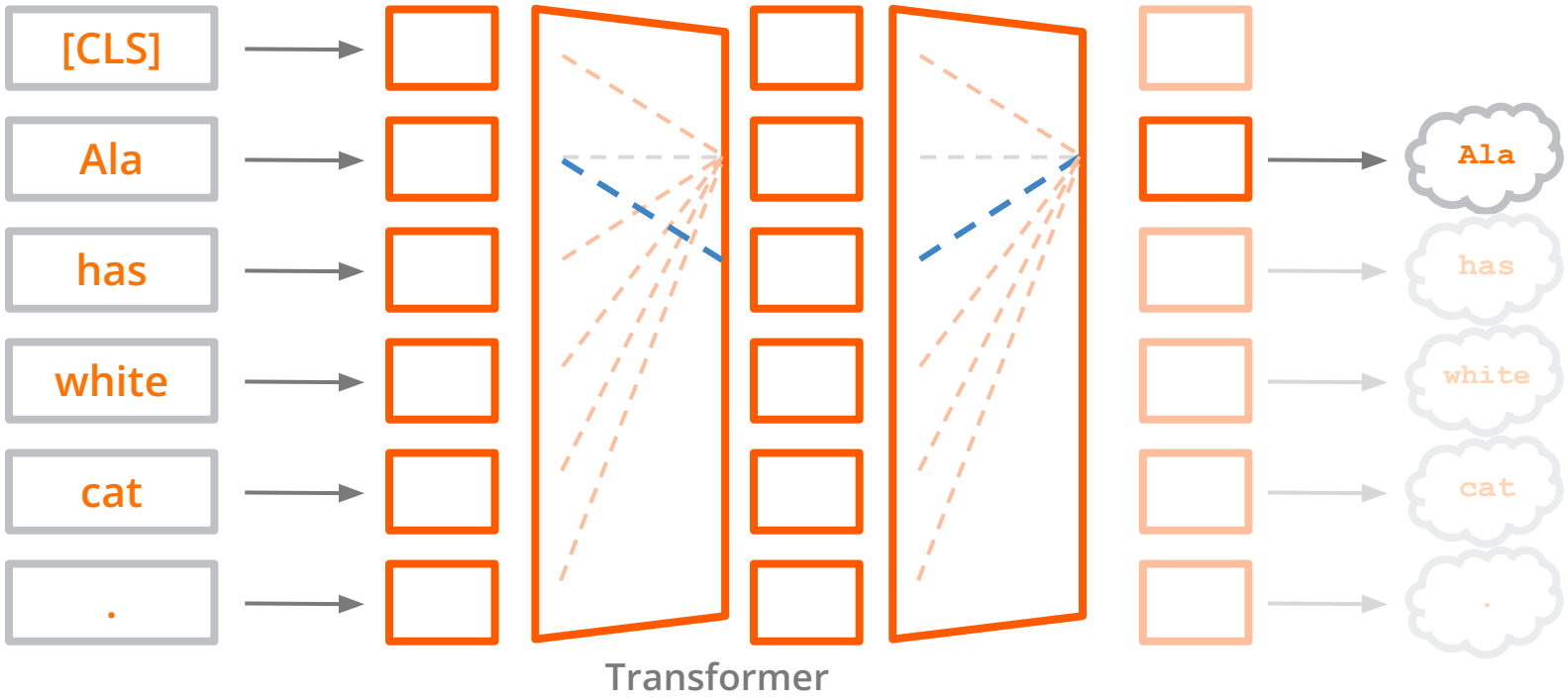
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



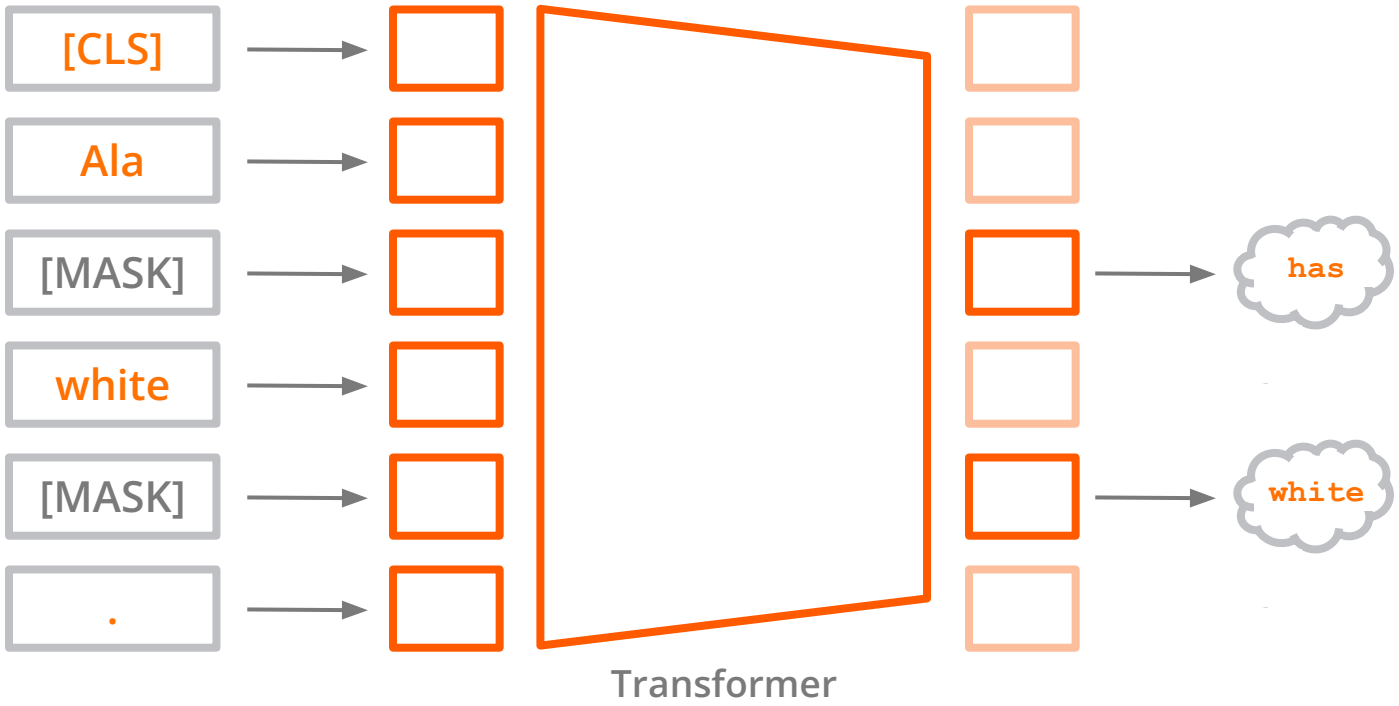
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

15% tokens

15% tokens

80% [MASK]

So it has to learn to use context

15% tokens

80% [MASK]

So it has to learn to use context

10% same

So it has to create meaningful embedding for non [MASK] tokens

15% tokens

80% [MASK]

So it has to learn to use context

10% same

So it has to create meaningful embedding for non [MASK] tokens

10% random

So it has to create meaningful embedding for all tokens instead of cheating

15% tokens

80% [MASK]

So it has to learn to use context

10% same

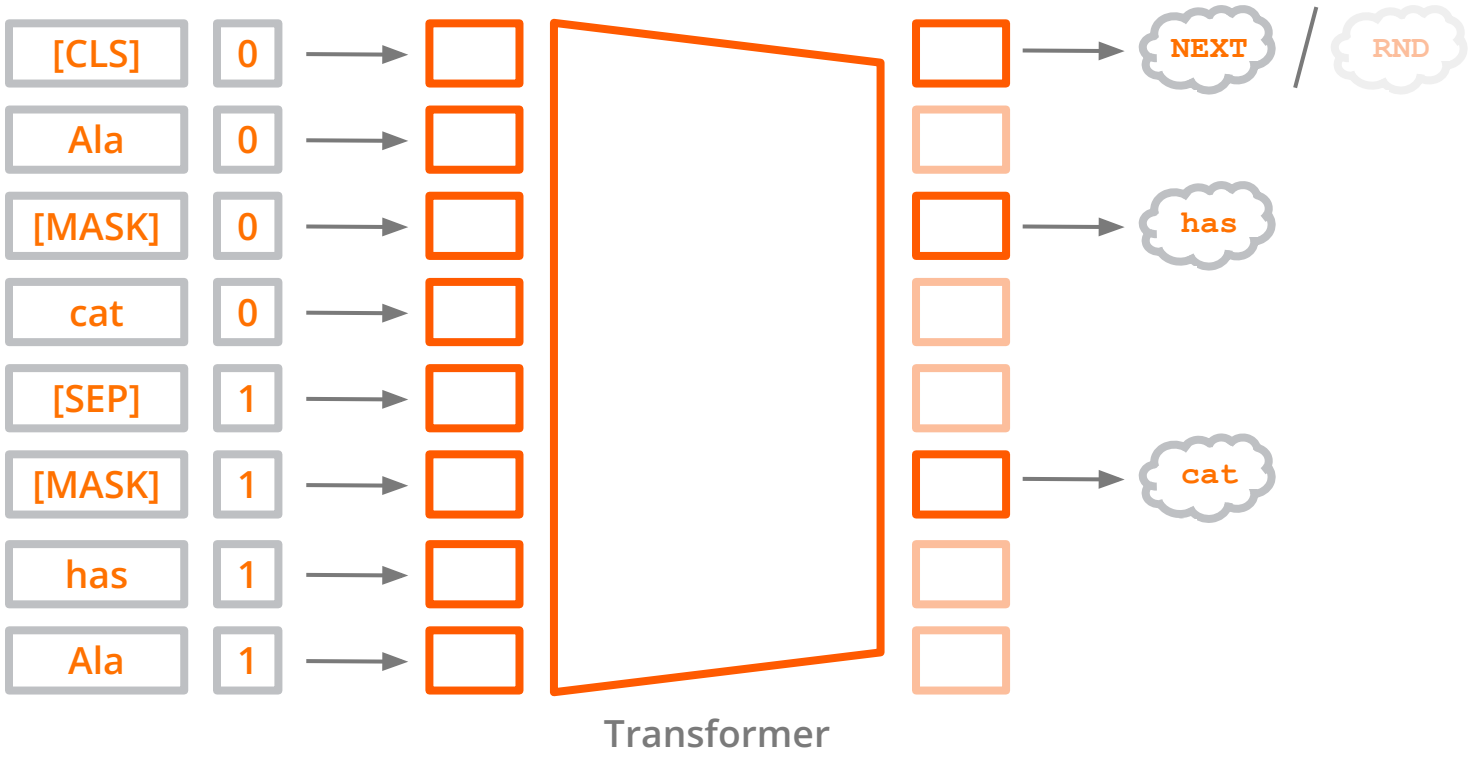
So it has to create meaningful embedding for non [MASK] tokens

10% random

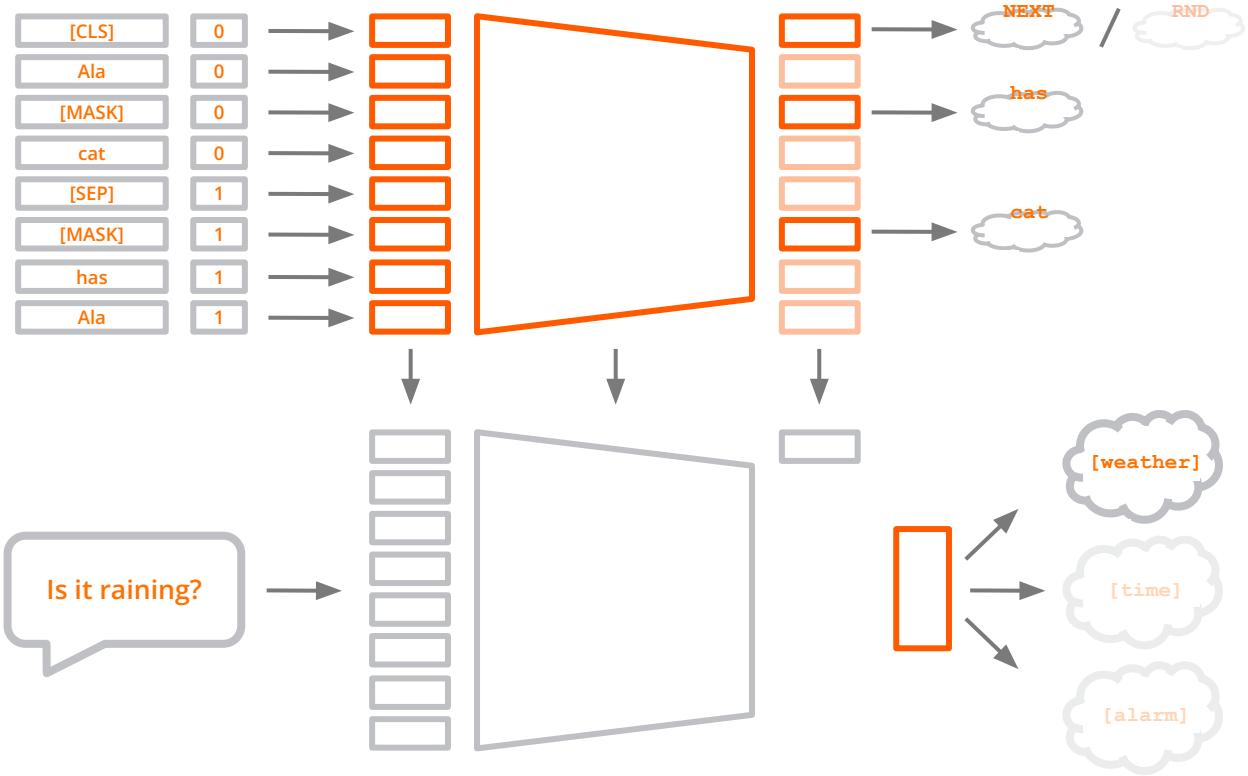
So it has to create meaningful embedding for all tokens instead of cheating

If you use only random
it will only rely on context

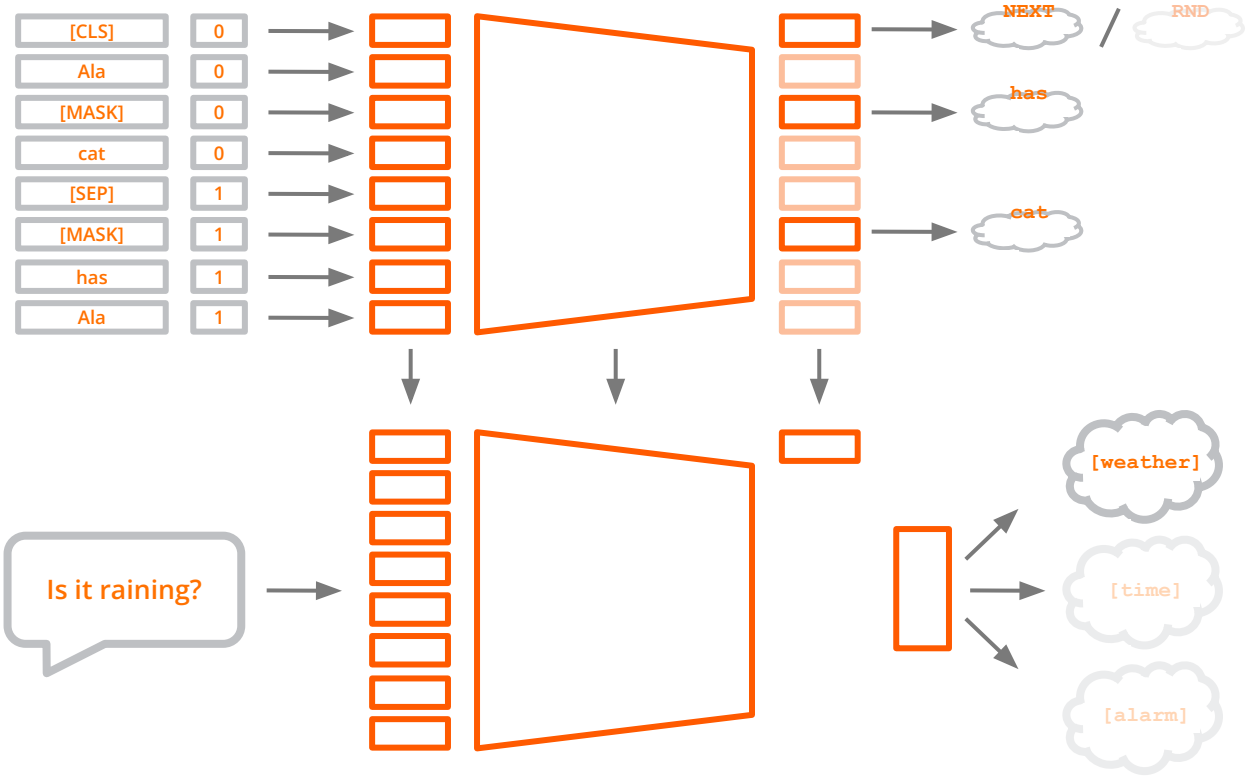
BERT: Next Sentence Prediction



BERT: feature extraction



BERT: fine-tuning



Evaluation

Perplexity

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Perplexity

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Vocabulary vs perplexity

Perplexity

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \ldots, w_N)}}$$

Vocabulary vs perplexity

What is the BERT vocabulary?

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”, “bac”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”, “bac”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”, “bac”, “baba”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = [“a”, “b”, “c”, ..., “z”, “ba”, “bac”, “baba”]

corpus = [
    “baba a baca”,
    “baba zbacza”,
    “zza bac”,
]
```

Byte Pair Encoding

```
vocab = ["a", "b", "c", ..., "z", "ba", "bac", "baba"]
```

b-a-b-a

ba-ba

baba

Byte Pair Encoding

```
vocab = ["a", "b", "c", ..., "z", "ba", "bac", "baba"]
```

b-a-b-a	**z-b-a-c-z-a**
ba-ba	**z-ba-c-z-a**
baba	**z-bac-z-a**

General Language Understanding Evaluation Benchmark

Evaluate fine-tuned model on downstream tasks

Alex Wang, *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, 2019

General Language Understanding Evaluation Benchmark

Evaluate fine-tuned model on downstream tasks

Nine diverse tasks

- semantic similarity, NLI, sentiment analysis

Alex Wang, *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, 2019

General Language Understanding Evaluation Benchmark

Evaluate fine-tuned model on downstream tasks

Nine diverse tasks

- semantic similarity, NLI, sentiment analysis

Common format

- [CLS] sentence 1 [SEP] sentence 2

Alex Wang, *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, 2019

General Language Understanding Evaluation Benchmark

Evaluate fine-tuned model on downstream tasks

Nine diverse tasks

- semantic similarity, NLI, sentiment analysis

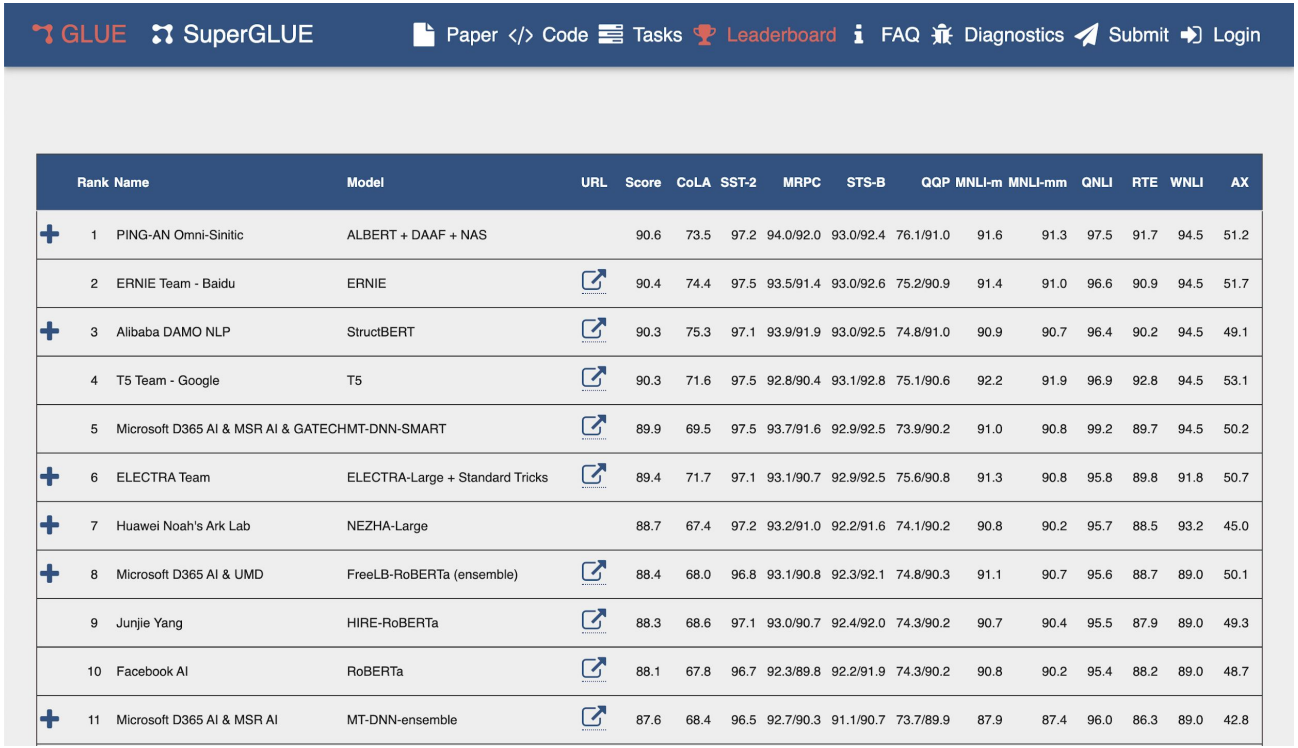
Common format

- [CLS] sentence 1 [SEP] sentence 2

Public leaderboard

Alex Wang, *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, 2019

General Language Understanding Evaluation Benchmark



Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
2	ERNIE Team - Baidu	ERNIE		90.4	74.4	97.5	93.5/91.4	93.0/92.6	75.2/90.9	91.4	91.0	96.6	90.9	94.5	51.7
3	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5	74.8/91.0	90.9	90.7	96.4	90.2	94.5	49.1
4	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
5	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
6	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
7	Huawei Noah's Ark Lab	NEZHA-Large		88.7	67.4	97.2	93.2/91.0	92.2/91.6	74.1/90.2	90.8	90.2	95.7	88.5	93.2	45.0
8	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
9	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
10	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
11	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8

Alex Wang, *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, 2019

What come after BERT?

RoBERTa

Remove Next Sentence Prediction

Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019

RoBERTa

Remove Next Sentence Prediction

Dynamic token masking

Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019

RoBERTa

Remove Next Sentence Prediction

Dynamic token masking

Larger batch size (2048 vs 256)

Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019

RoBERTa

Remove Next Sentence Prediction

Dynamic token masking

Larger batch size (2048 vs 256)

More training data

Yinhan Liu et al., *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, 2019

RoBERTa architecture

Alexis Conneau et al., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020

XLM-RoBERTa

RoBERTa architecture

Common Crawl for 100 languages

Alexis Conneau et al., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020

XLM-RoBERTa

RoBERTa architecture

Common Crawl for 100 languages

Larger vocab size (250k vs 50k)

Alexis Conneau et al., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020

Add Sentence Order Prediction

Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

ALBERT

Add Sentence Order Prediction

Shared weights for all layers

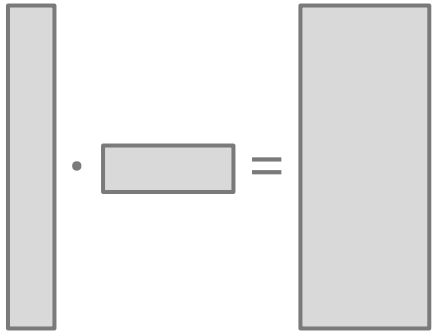
Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

ALBERT

Add Sentence Order Prediction

Shared weights for all layers

Factorization of word embeddings



Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

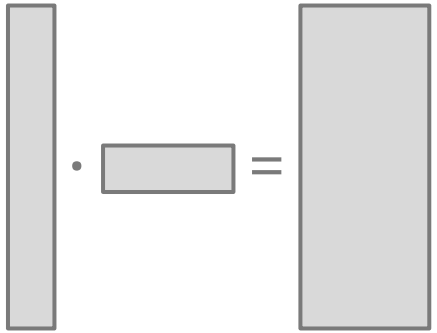
ALBERT

Add Sentence Order Prediction

Shared weights for all layers

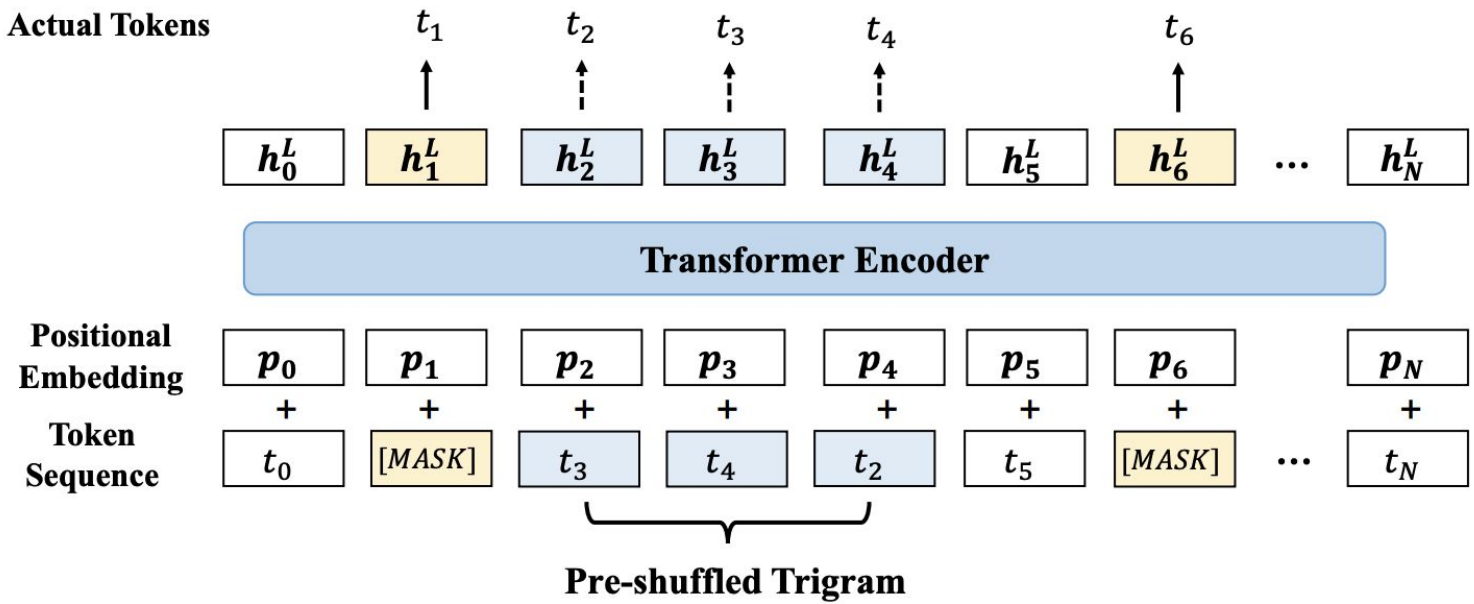
Factorization of word embeddings

More layers



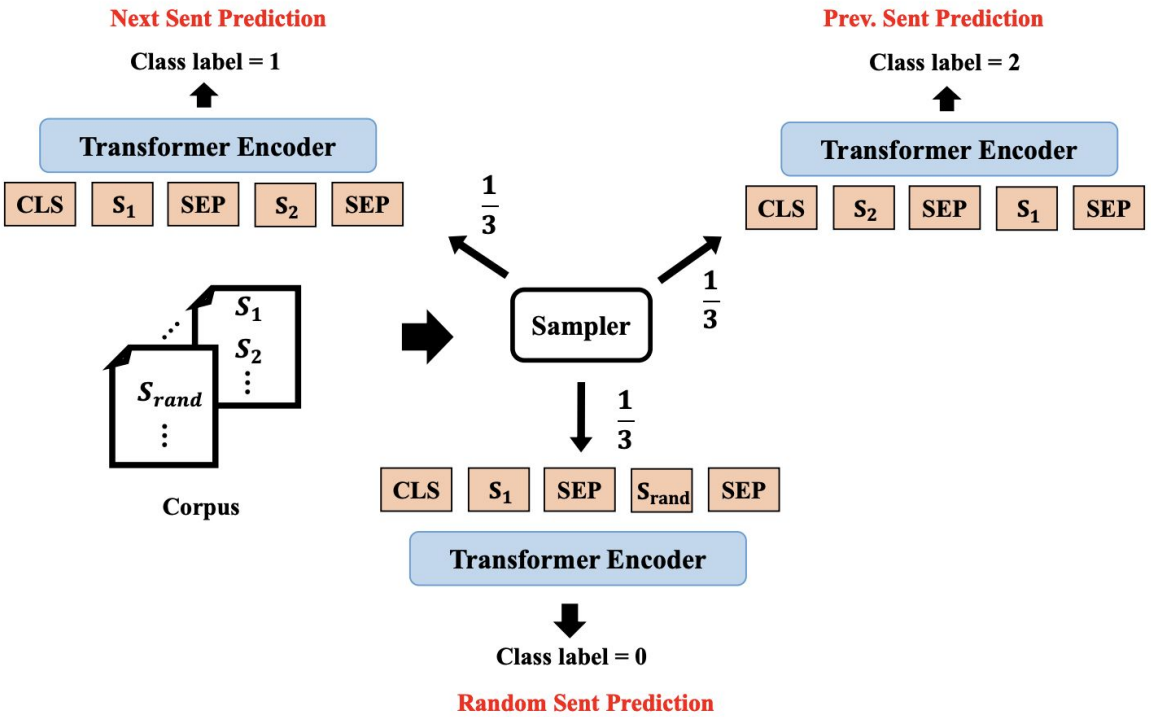
Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

StructBERT: Word Structural Objective



Wei Wang et al., *StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding*, 2019

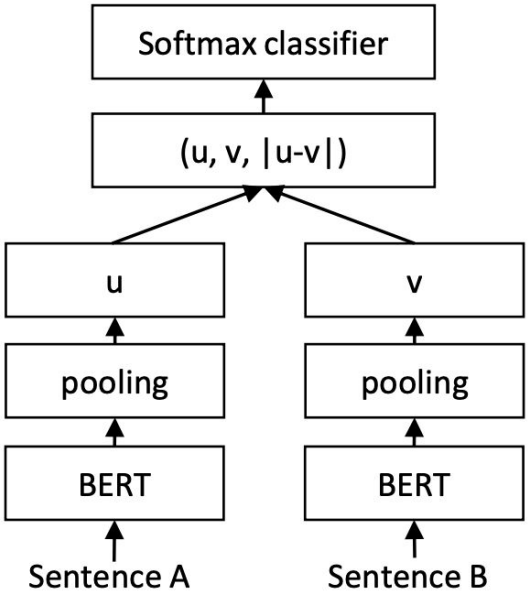
StructBERT: Sentence Structural Objective



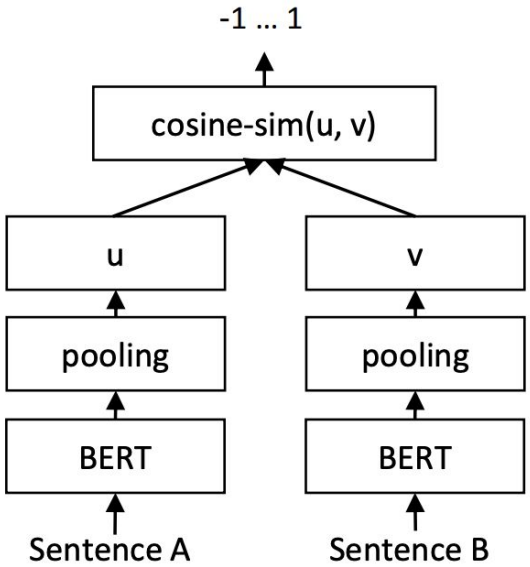
Wei Wang et al., *StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding*, 2019

Sentence BERT

Training



Prediction

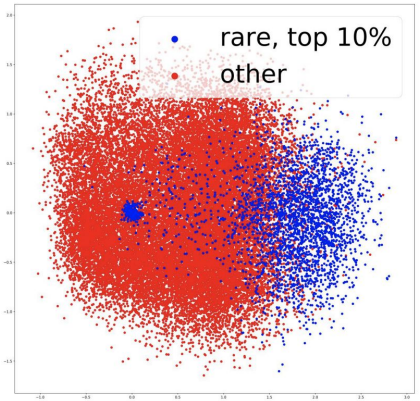


Nils Reimers and Iryna Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, 2019

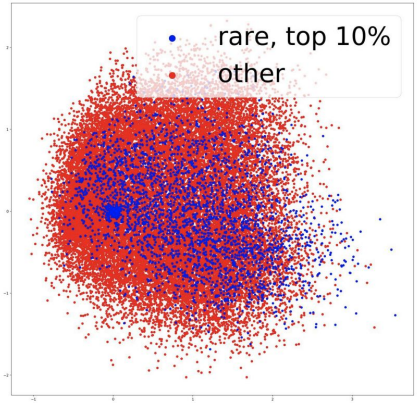
BPE-Dropout

```
u-n_r-e-l-a_t-e_d
u-n re-l_a-t-e_d
u-n re_l-at-e_d
un re-l-at-e-d
un re_l-at-ed
un re-lat-ed
un relat_ed
```

```
u-n-r-e-l-a_t-e-d
u_n re_l-a-t-e-d
u_n re-l-at-e-d
u_n re-l-ate_d
u_n rel-ate-d
u_n relate_d
```



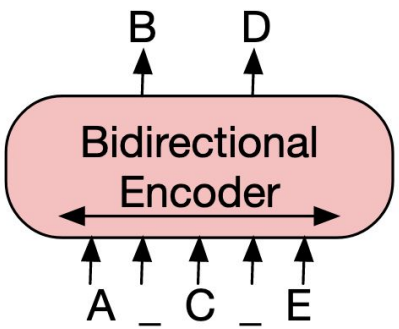
(a) BPE



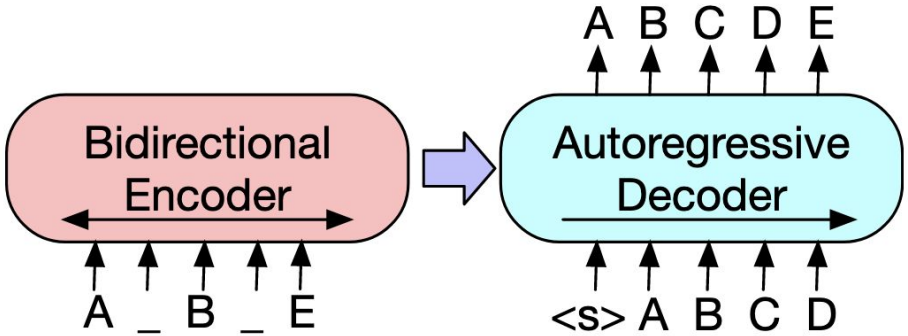
(b) *BPE-dropout*

Ivan Provilkov et al., *BPE-Dropout: Simple and Effective Subword Regularization*, 2020

BART



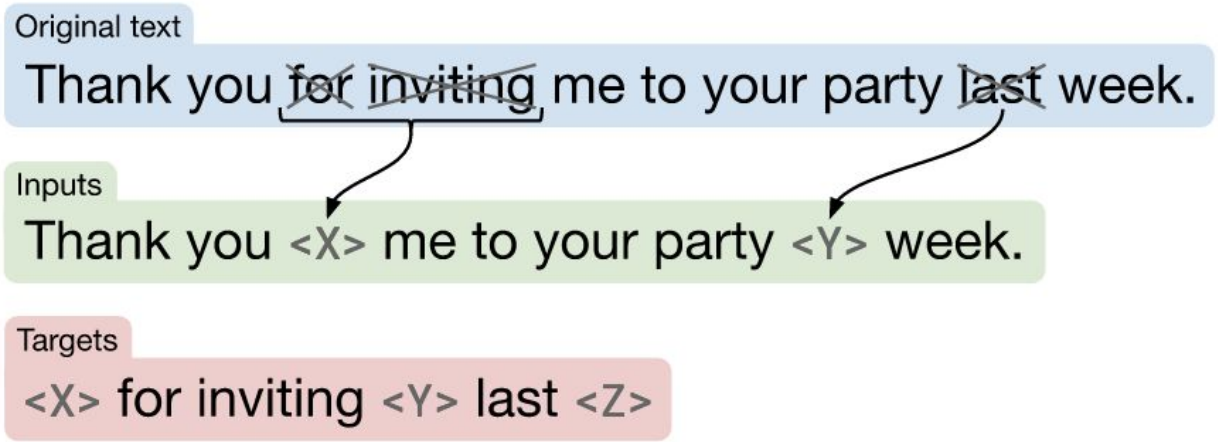
BERT



BART

Mike Lewis et al., *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019

T5

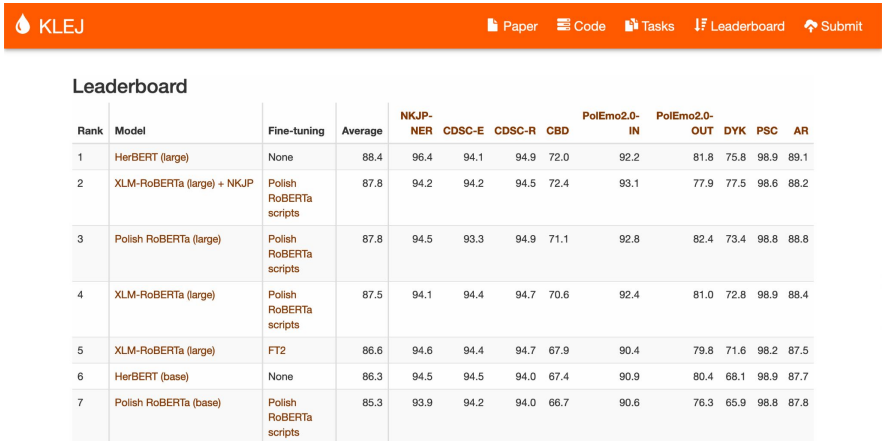


Colin Raffel et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, 2019

Resources for Polish language

KLEJ Benchmark

Kompleksowa
Lista
Ewaluacji
Językowych



Leaderboard

Rank	Model	Fine-tuning	Average	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	DYK	PSC	AR
1	HerBERT (large)	None	88.4	96.4	94.1	94.9	72.0	92.2	81.8	75.8	98.9	89.1
2	XLM-RoBERTa (large) + NKJP	Polish RoBERTa scripts	87.8	94.2	94.2	94.5	72.4	93.1	77.9	77.5	98.6	88.2
3	Polish RoBERTa (large)	Polish RoBERTa scripts	87.8	94.5	93.3	94.9	71.1	92.8	82.4	73.4	98.8	88.8
4	XLM-RoBERTa (large)	Polish RoBERTa scripts	87.5	94.1	94.4	94.7	70.6	92.4	81.0	72.8	98.9	88.4
5	XLM-RoBERTa (large)	FT2	86.6	94.6	94.4	94.7	67.9	90.4	79.8	71.6	98.2	87.5
6	HerBERT (base)	None	86.3	94.5	94.5	94.0	67.4	90.9	80.4	68.1	98.9	87.7
7	Polish RoBERTa (base)	Polish RoBERTa scripts	85.3	93.9	94.2	94.0	66.7	90.6	76.3	65.9	98.8	87.8

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, Ireneusz Gawlik, *KLEJ: Comprehensive Benchmark for Polish Language Understanding*, 2020

Name	Train	Dev	Test	Domain	Metrics	Objective
			Single-Sentence Tasks			
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
			Multi-Sentence Tasks			
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

Name	Train	Dev	Test	Domain	Metrics	Objective
				Single-Sentence Tasks		
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
				Multi-Sentence Tasks		
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

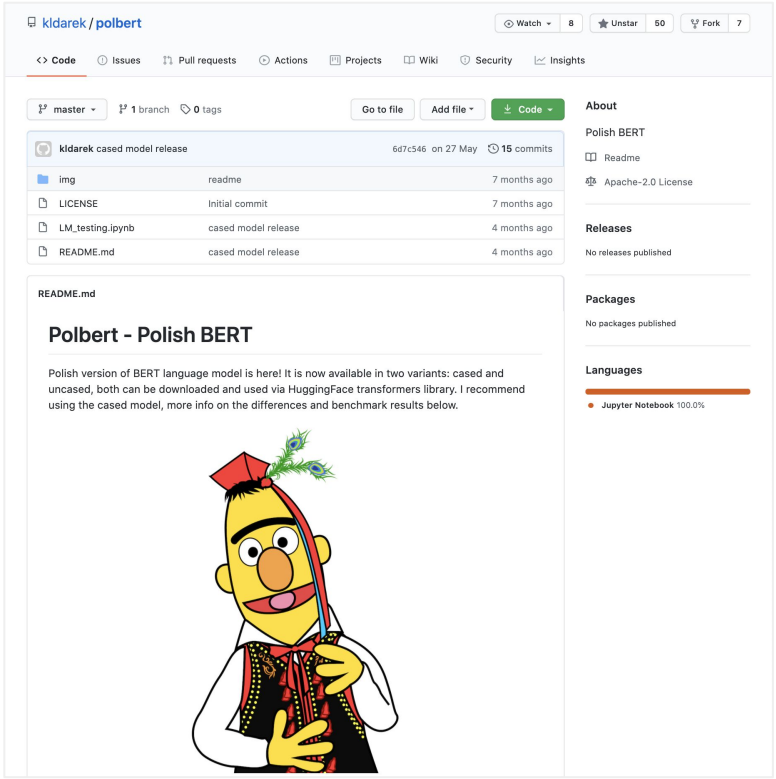
Name	Train	Dev	Test	Domain	Metrics	Objective
				Single-Sentence Tasks		
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
				Multi-Sentence Tasks		
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

Name	Train	Dev	Test	Domain	Metrics	Objective
				Single-Sentence Tasks		
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
				Multi-Sentence Tasks		
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

Name	Train	Dev	Test	Domain	Metrics	Objective
Single-Sentence Tasks						
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
Multi-Sentence Tasks						
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

Name	Train	Dev	Test	Domain	Metrics	Objective
Single-Sentence Tasks						
NKJP-NER	16k	2k	2k	Balanced corpus	Accuracy	NER classification
CDSC-R	8k	1k	1k	Image captions	Spearman corr.	Semantic relatedness
CDSC-E	8k	1k	1k	Image captions	Accuracy	Textual entailment
Multi-Sentence Tasks						
CBD	10k	-	1k	Social Media	F1-Score	Cyberbullying detection
PolEmo2.0-IN	6k	0.7k	0.7k	Online reviews	Accuracy	Sentiment analysis
PolEmo2.0-OUT	6k	0.5k	0.5k	Online reviews	Accuracy	Sentiment analysis
Czy wiesz?	5k	-	1k	Wikipedia	F1-Score	Question answering
PSC	4k	-	1k	News articles	F1-Score	Paraphrase
AR	10k	1k	1k	Online reviews	1 − wMAE	Sentiment analysis

Polbert



Model

- BERT Base

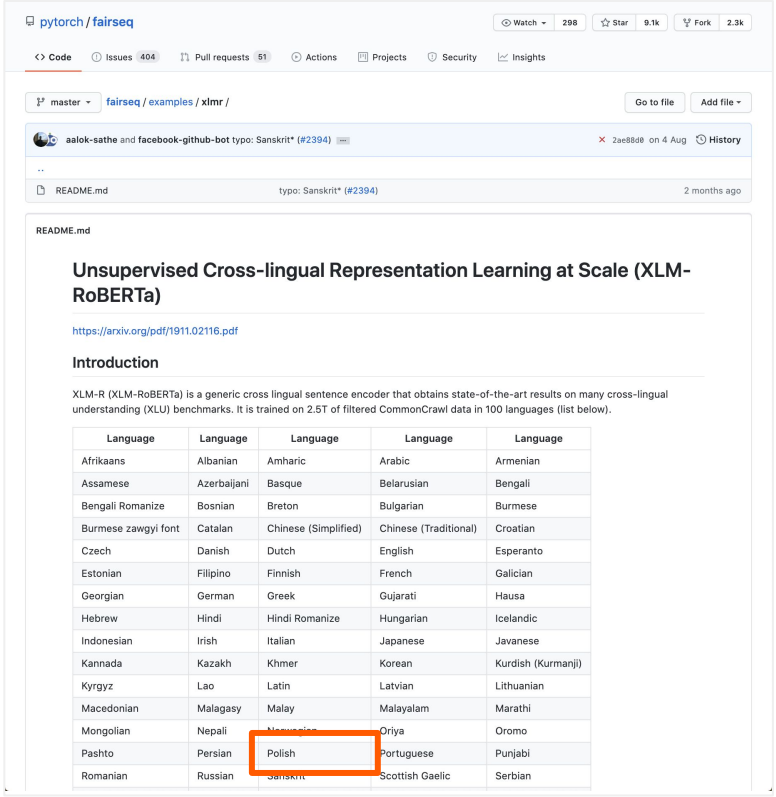
Corpora

- Open Subtitles
- ParaCrawl
- Korpus Parlamentarny
- Wikipedia

Polbert

Model	Size	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
MultiBERT	Base	79.5	91.4	**93.8**	92.9	40.0	85.0	66.6	**64.2**	97.9	83.3
Polbert	Base	**81.7**	**93.6**	93.4	**93.8**	**52.7**	**87.4**	**71.1**	59.1	**98.6**	**85.2**

XLM-RoBERTa



pytorch / fairseq

Watch 298 · Star 9.1k · Fork 2.3k

Code · Issues 404 · Pull requests 51 · Actions · Projects · Security · Insights

master · fairseq / examples / xlmr /

Go to file · Add file

aalok-sathe and facebook-github-bot typo: Sanskrit* (#2394) · 2ae88d0 on 4 Aug · History

README.md — typo: Sanskrit* (#2394) — 2 months ago

README.md

Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa)

https://arxiv.org/pdf/1911.02116.pdf

Introduction

XLM-R (XLM-RoBERTa) is a generic cross lingual sentence encoder that obtains state-of-the-art results on many cross-lingual understanding (XLU) benchmarks. It is trained on 2.5T of filtered CommonCrawl data in 100 languages (list below).

Language	Language	Language	Language	Language
Afrikaans	Albanian	Amharic	Arabic	Armenian
Assamese	Azerbaijani	Basque	Belarusian	Bengali
Bengali Romanize	Bosnian	Breton	Bulgarian	Burmese
Burmese zawgyi font	Catalan	Chinese (Simplified)	Chinese (Traditional)	Croatian
Czech	Danish	Dutch	English	Esperanto
Estonian	Filipino	Finnish	French	Galician
Georgian	German	Greek	Gujarati	Hausa
Hebrew	Hindi	Hindi Romanize	Hungarian	Icelandic
Indonesian	Irish	Italian	Japanese	Javanese
Kannada	Kazakh	Khmer	Korean	Kurdish (Kurmanji)
Kyrgyz	Lao	Latin	Latvian	Lithuanian
Macedonian	Malagasy	Malay	Malayalam	Marathi
Mongolian	Nepali	[illegible]	Oriya	Oromo
Pashto	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	[illegible]	Scottish Gaelic	Serbian

Model

- RoBERTa Base & Large

Corpora

- Common Crawl for 100 languages

Alexis Conneau et al., *Unsupervised Cross-lingual Representation Learning at Scale*, 2020

XLM-RoBERTa

Model	Size	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
MultiBERT	Base	79.5	91.4	93.8	92.9	40.0	85.0	66.6	64.2	97.9	83.3
Polbert	Base	**81.7**	**93.6**	93.4	**93.8**	**52.7**	87.4	71.1	**59.1**	**98.6**	**85.2**
XLM-RoBERTa	Base	81.5	92.1	**94.1**	93.3	51.0	**89.5**	**74.7**	55.8	98.2	**85.2**

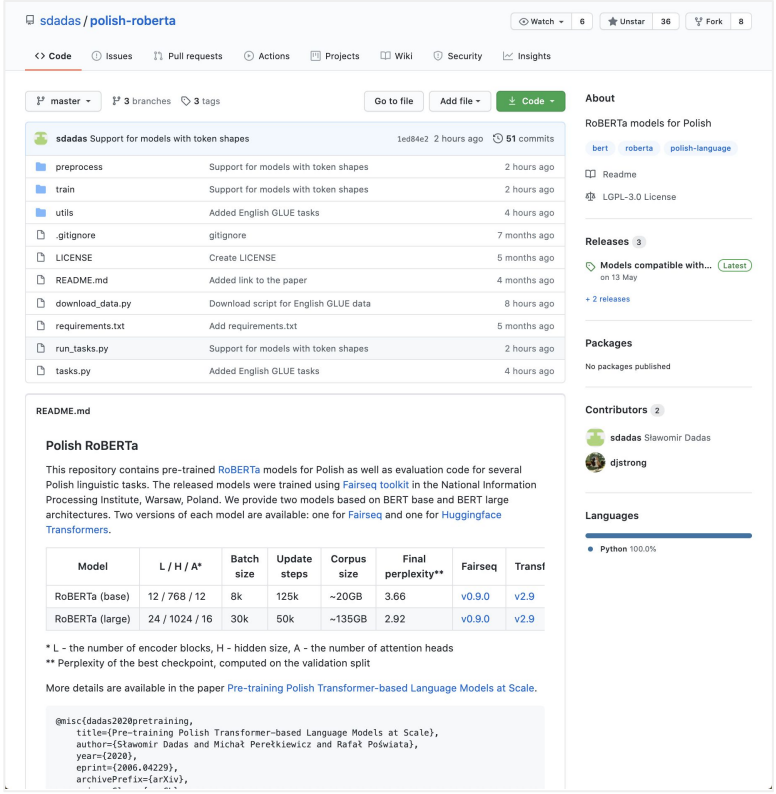
XLM-RoBERTa

Model	Size	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
MultiBERT	Base	79.5	91.4	93.8	92.9	40.0	85.0	66.6	64.2	97.9	83.3
Polbert	Base	81.7	93.6	93.4	93.8	**52.7**	87.4	71.1	59.1	**98.6**	85.2
XLM-RoBERTa	Base	81.5	92.1	94.1	93.3	51.0	89.5	74.7	55.8	98.2	85.2
XLM-RoBERTa	Large	**84.7**	**94.6**	**94.4**	**94.7**	50.7	**90.4**	**79.8**	**71.6**	98.2	**87.5**

XLM-RoBERTa

Model	Size	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
MultiBERT	Base	79.5	91.4	93.8	92.9	40.0	85.0	66.6	64.2	97.9	83.3
Polbert	Base	81.7	93.6	93.4	93.8	**52.7**	87.4	71.1	59.1	**98.6**	85.2
XLM-RoBERTa	Base	81.5	92.1	94.1	93.3	51.0	89.5	74.7	55.8	98.2	85.2
XLM-RoBERTa	Large	**84.7**	**94.6**	**94.4**	**94.7**	50.7	**90.4**	**79.8**	**71.6**	98.2	**87.5**

Polish RoBERTa



sdadas / polish-roberta

Watch 6 · Unstar 36 · Fork 8

Code · Issues · Pull requests · Actions · Projects · Wiki · Security · Insights

master · 3 branches · 3 tags · Go to file · Add file · Code

sdadas Support for models with token shapes — 1ed84e2 2 hours ago · 51 commits

preprocess	Support for models with token shapes	2 hours ago
train	Support for models with token shapes	2 hours ago
utils	Added English GLUE tasks	4 hours ago
.gitignore	gitignore	7 months ago
LICENSE	Create LICENSE	5 months ago
README.md	Added link to the paper	4 months ago
download_data.py	Download script for English GLUE data	8 hours ago
requirements.txt	Add requirements.txt	5 months ago
run_tasks.py	Support for models with token shapes	2 hours ago
tasks.py	Added English GLUE tasks	4 hours ago

README.md

Polish RoBERTa

This repository contains pre-trained RoBERTa models for Polish as well as evaluation code for several Polish linguistic tasks. The released models were trained using Fairseq toolkit in the National Information Processing Institute, Warsaw, Poland. We provide two models based on BERT base and BERT large architectures. Two versions of each model are available: one for Fairseq and one for Huggingface Transformers.

Model	L / H / A*	Batch size	Update steps	Corpus size	Final perplexity**	Fairseq	Transf
RoBERTa (base)	12 / 768 / 12	8k	125k	~20GB	3.66	v0.9.0	v2.9
RoBERTa (large)	24 / 1024 / 16	30k	50k	~135GB	2.92	v0.9.0	v2.9

* L - the number of encoder blocks, H - hidden size, A - the number of attention heads
** Perplexity of the best checkpoint, computed on the validation split

More details are available in the paper Pre-training Polish Transformer-based Language Models at Scale.

```
@misc{dadas2020pretraining,
    title={Pre-training Polish Transformer-based Language Models at Scale},
    author={Sławomir Dadas and Michał Perełkiewicz and Rafał Poświata},
    year={2020},
    eprint={2006.04229},
    archivePrefix={arXiv},
```

About: RoBERTa models for Polish

bert · roberta · polish-language

Readme

LGPL-3.0 License

Releases 3

Models compatible with... Latest on 13 May

+ 2 releases

Packages

No packages published

Contributors 2

sdadas Sławomir Dadas

djstrong

Languages

Python 100.0%

Model

- RoBERTa Base & Large

Corpora

- Common Crawl
- Korpus Parlamentarny
- Wikipedia
- Other

Sławomir Dadas et al., *Pre-training Polish Transformer-based Language Models at Scale*, 2020

Polish RoBERTa

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	**87.8**	94.5	93.3	**94.9**	**71.1**	**92.8**	**82.4**	**73.4**	**98.8**	**88.8**

Polish RoBERTa

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	**87.8**	94.5	93.3	**94.9**	**71.1**	**92.8**	**82.4**	**73.4**	**98.8**	**88.8**

Polish RoBERTa

Model	Fine-tuning	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	Old	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	New	**87.8**	94.5	93.3	**94.9**	**71.1**	**92.8**	**82.4**	**73.4**	**98.8**	**88.8**

Polish RoBERTa

Model	Fine-tuning	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	Old	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	New	**87.8**	94.5	93.3	**94.9**	**71.1**	**92.8**	**82.4**	**73.4**	98.8	**88.8**
XLM-RoBERTa	New	87.5	94.1	**94.4**	94.7	70.6	92.4	81.0	72.8	**98.9**	88.4

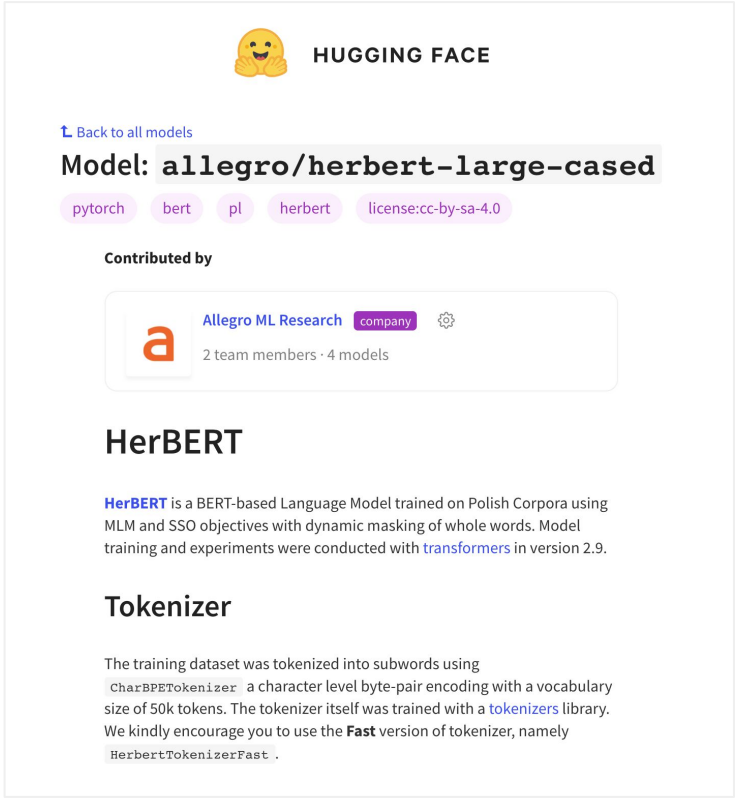
Polish RoBERTa

Model	Fine-tuning	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	Old	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	New	**87.8**	94.5	93.3	**94.9**	**71.1**	**92.8**	**82.4**	**73.4**	98.8	**88.8**
XLM-RoBERTa	New	87.5	94.1	**94.4**	94.7	70.6	92.4	81.0	72.8	**98.9**	88.4

XLM-RoBERTa + NKJP

Model	Fine-tuning	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	Old	84.7	**94.6**	**94.4**	94.7	50.7	90.4	79.8	71.6	98.2	87.5
Polish RoBERTa	New	87.8	94.5	93.3	**94.9**	71.1	92.8	**82.4**	73.4	98.8	**88.8**
XLM-RoBERTa	New	87.5	94.1	**94.4**	94.7	70.6	92.4	81.0	72.8	**98.9**	88.4
XLM-RoBERTa + NKJP	New	**87.8**	94.2	94.2	94.5	**72.4**	**93.1**	77.9	**77.5**	98.6	88.2

HerBERT



Model

- RoBERTa Base & Large

Corpora

- Common Crawl
- Wikipedia
- Open Subtitles
- Wolne Lektury
- NKJP

Robert Mroczkowski et al., *HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish*, 2021

Initialization with XLM-RoBERTa weights

Initialization with XLM-RoBERTa weights

XLM-RoBERTa [250k]

HerBERT [50k]

Initialization with XLM-RoBERTa weights

XLM-RoBERTa [250k]

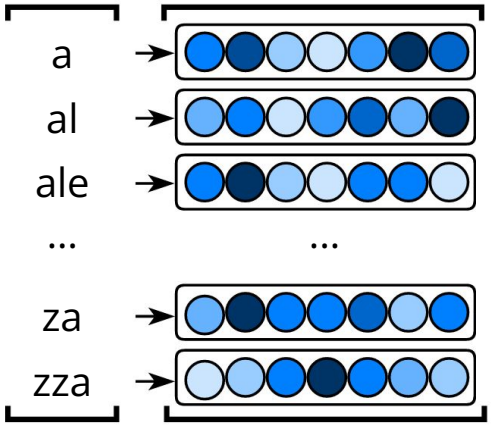
```
['Sło', 'wa', 'cki', 'wielki', 'm', 'poet', 'ą', 'był']
```

HerBERT [50k]

```
['Słowa', 'cki', 'wielkim', 'poetą', 'był']
```

Initialization with XLM-RoBERTa weights

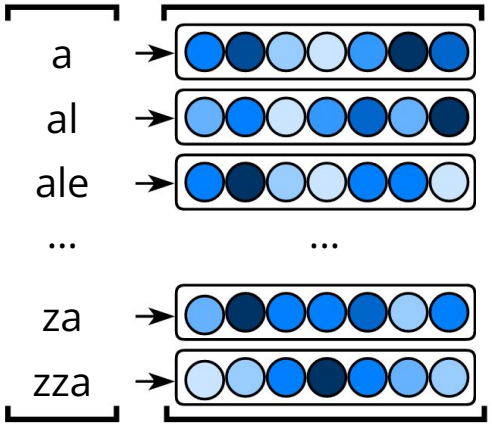
[250k]



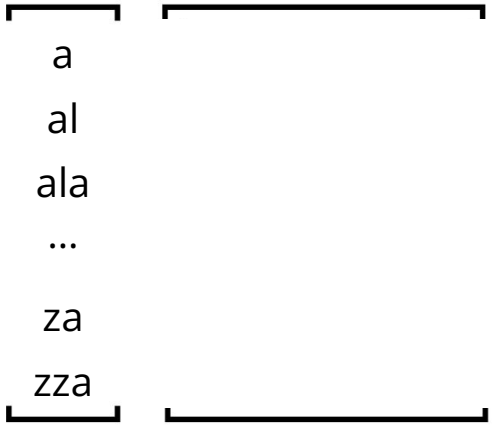
XLM-RoBERTa

Initialization with XLM-RoBERTa weights

[250k]



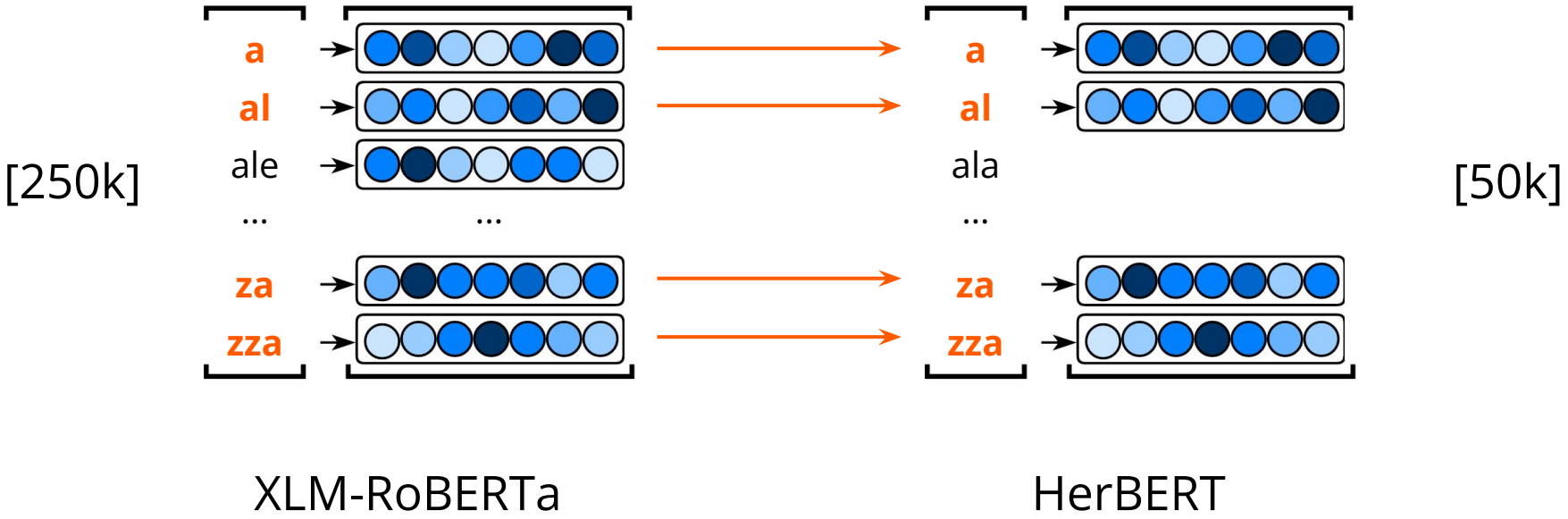
XLM-RoBERTa



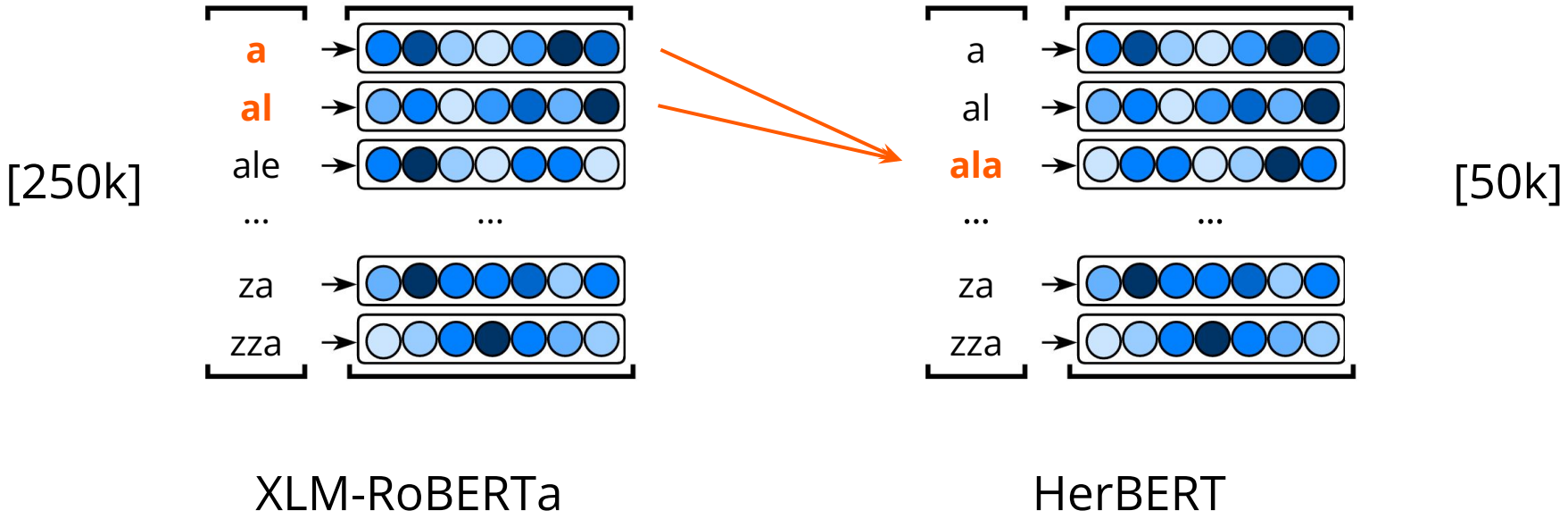
HerBERT

[50k]

Initialization with XLM-RoBERTa weights



Initialization with XLM-RoBERTa weights



HerBERT: Initialization

Init	Pretraining	Score
Random	No	44.88 ± 0.20
XLM-R	-	84.70 ± 0.29

HerBERT: Initialization

Init	Pretraining	Score
Random	No	44.88 ± 0.20
XLM-R	No	75.34 ± 2.16
XLM-R	-	84.70 ± 0.29

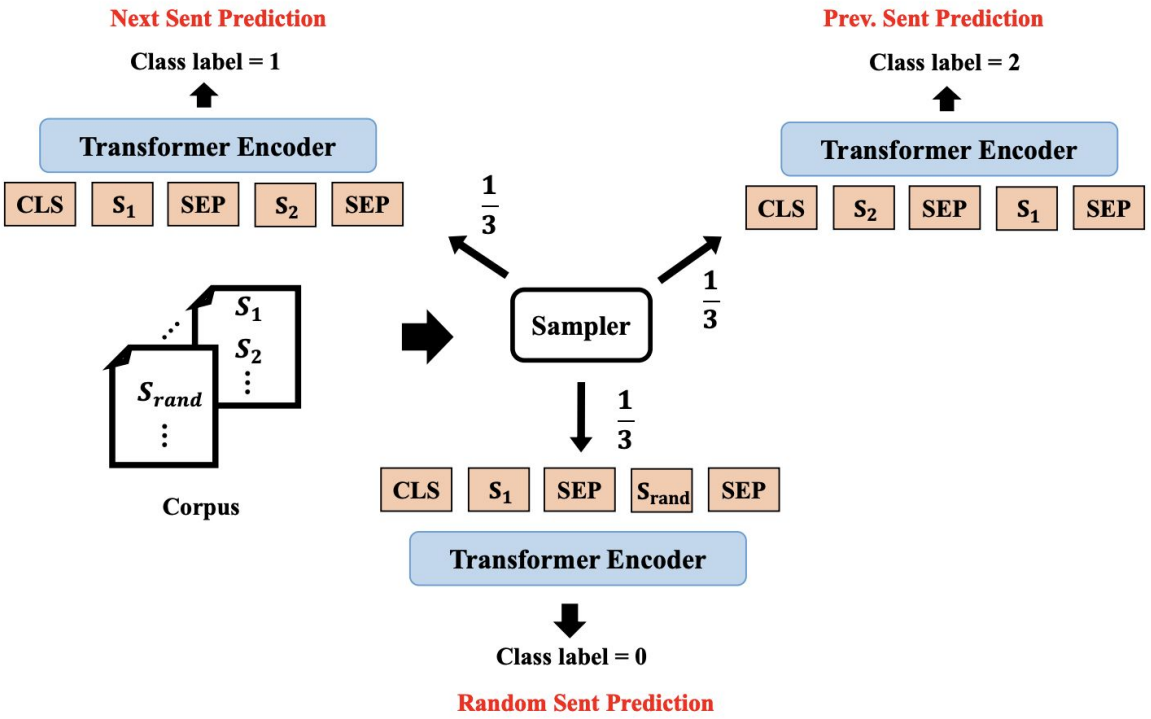
HerBERT: Initialization

Init	Pretraining	Score
Random	No	44.88 ± 0.20
XLM-R	No	75.34 ± 2.16
Random	Yes	82.38 ± 0.33
XLM-R	-	84.70 ± 0.29

HerBERT: Initialization

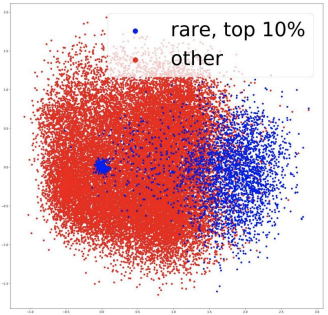
Init	Pretraining	Score
Random	No	44.88 ± 0.20
XLM-R	No	75.34 ± 2.16
Random	Yes	82.38 ± 0.33
XLM-R	Yes	**85.00 ± 0.28**
XLM-R	-	84.70 ± 0.29

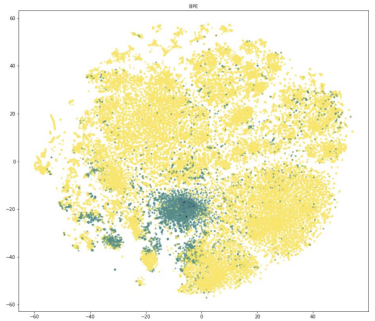
HerBERT: Sentence Structural Objective

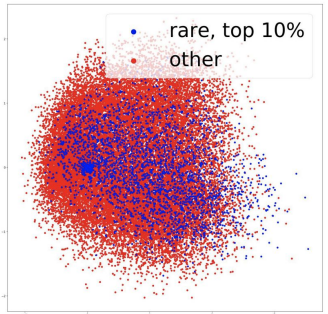


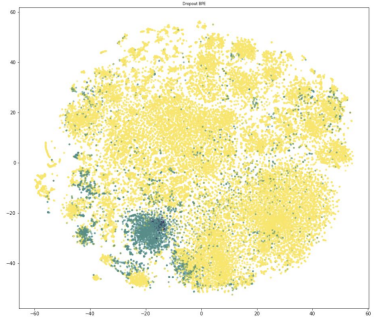
BPE

BPE-Dropout









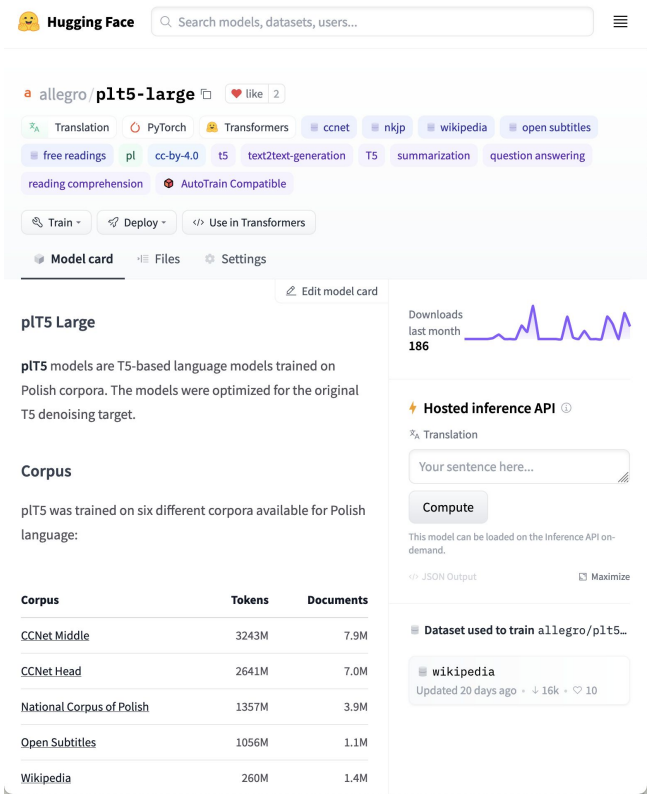
Original

HerBERT

BPE	Init	Score
No	Random	82.38 ± 0.33
Yes	Random	81.92 ± 0.26
No	XLM-R	85.00 ± 0.28
Yes	XLM-R	84.38 ± 0.32

Model	Size	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PE2.0-IN	PE2.0-OUT	DYK	PSC	AR
XLM-RoBERTa	Large	87.5	94.1	**94.4**	94.7	70.6	92.4	81.0	72.8	**98.9**	88.4
Polish RoBERTa	Large	87.8	94.5	93.3	**94.9**	71.1	**92.8**	**82.4**	73.4	98.8	88.8
HerBERT	Large	**88.4**	**96.4**	94.1	**94.9**	**72.0**	92.2	81.8	**75.8**	**98.9**	**89.1**

plT5



Model

- T5 Small, Base & Large

Corpora

- Common Crawl
- Wikipedia
- Open Subtitles
- Wolne Lektury
- NKJP

https://huggingface.co/allegro/plt5-large

Thank you!

piotr.cezary.rybak@gmail.com