PRZEGLĄD MODELI BERT DLA JĘZYKA POLSKIEGO

Piotr Rybak

allegro

Agenda

BERT: krótkie wprowadzenie

KLEJ: jak ewaluować modele BERT?

Przegląd: modeli BERT dla języka polskiego

HerBERT: wspólny model IPI PAN i Allegro

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

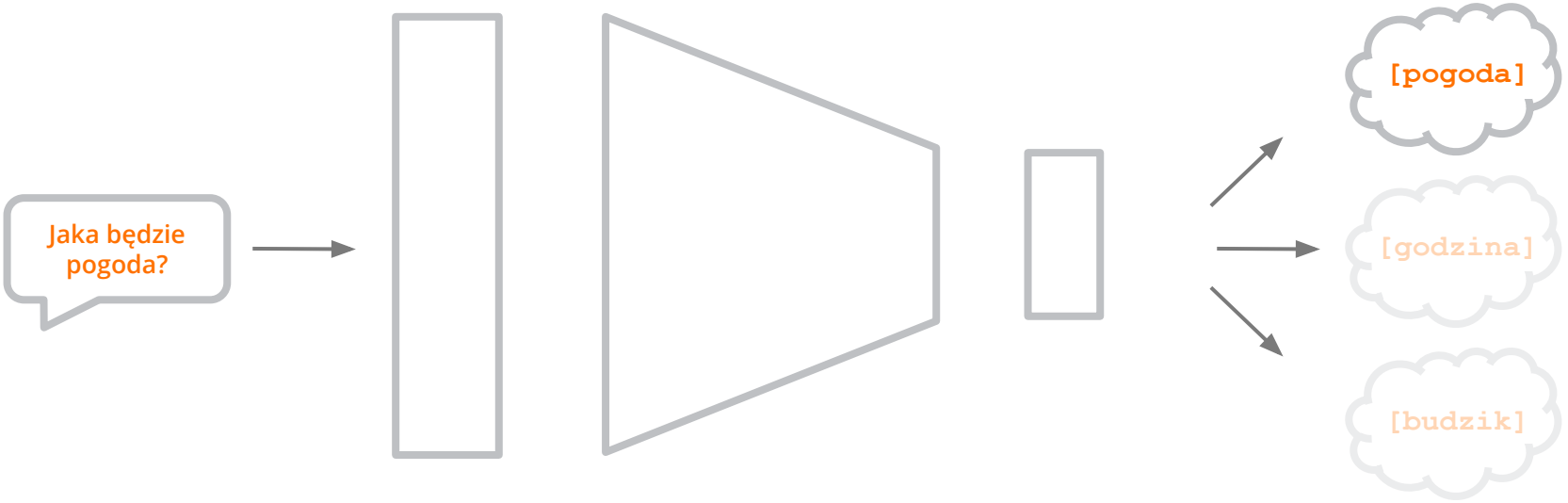
Abstract

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-

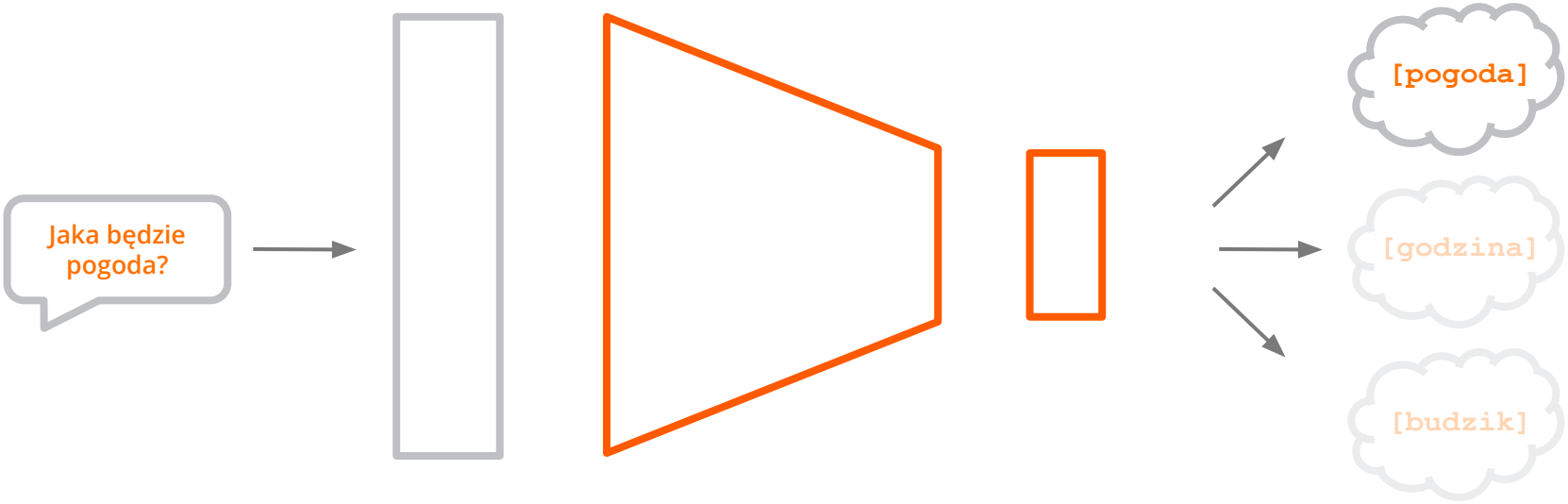

NLP SINCE 2019
BERT, BERT EVERYWHERE
imgflip.com


Jaka będzie pogoda?
[pogoda]
[godzina]
[budzik]
Wejście
Enkoder
Klasyfikator

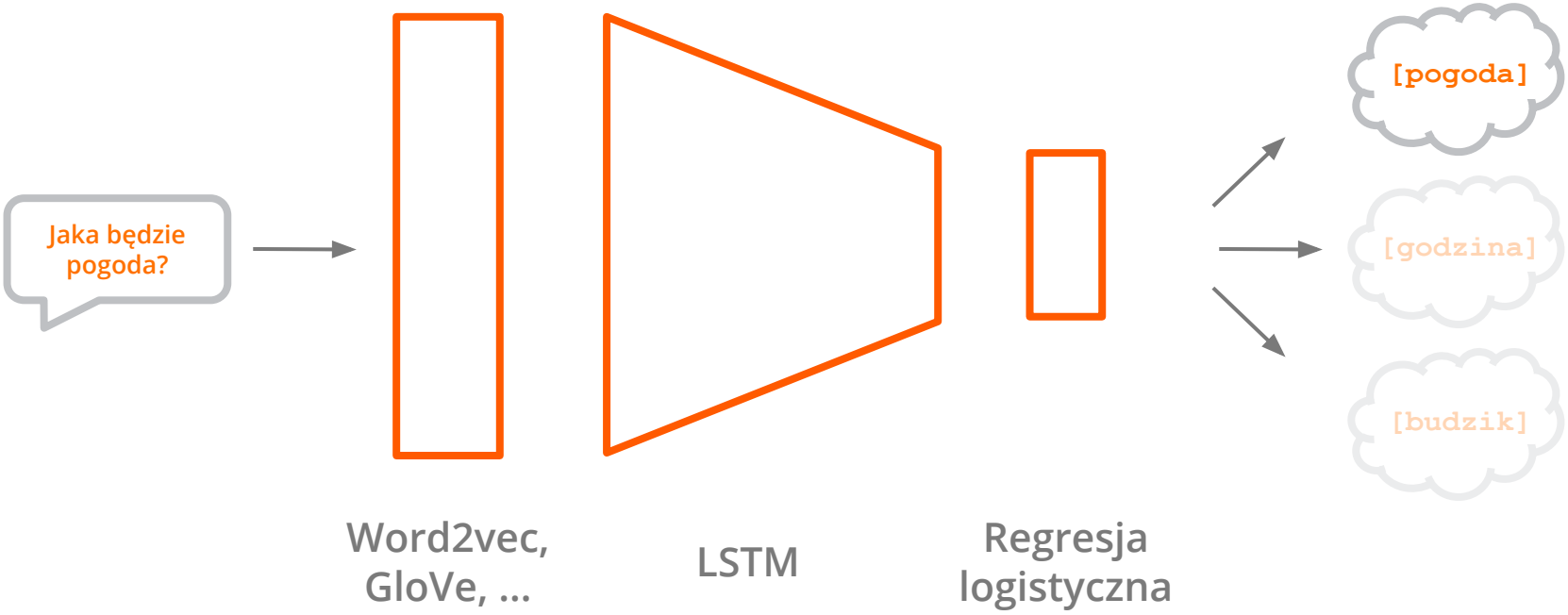
Bag-of-words


Jaka będzie pogoda?
[pogoda]
[godzina]
[budzik]
Bag-of-words
Regresja logistyczna, XGBoost, ...

Word Embeddings + LSTM



Fine-tuned Word Embeddings + LSTM



Fine-tuned Word Embeddings + LSTM

Dużo parametrów do wytrenowania

Fine-tuned Word Embeddings + LSTM

Dużo parametrów do wytrenowania

Łatwo przeuczyć model

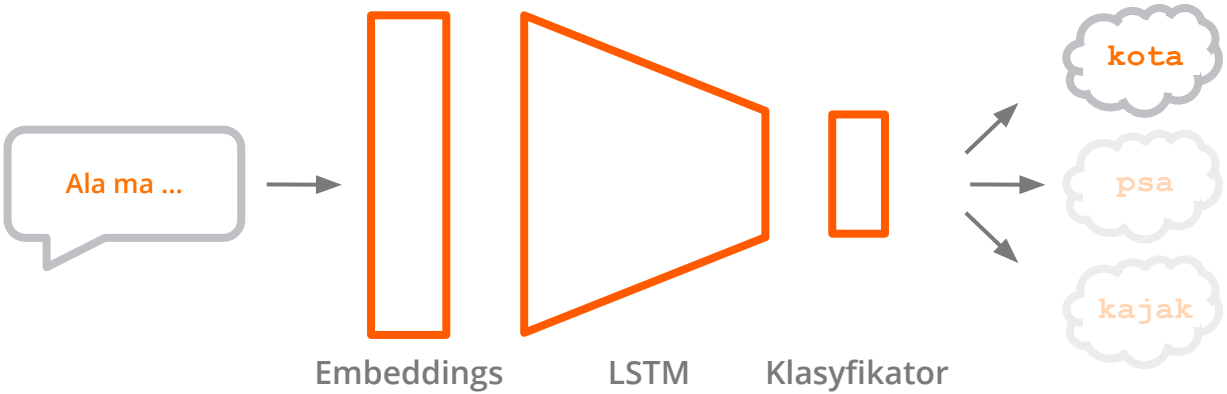
Fine-tuned Word Embeddings + LSTM

Dużo parametrów do wytrenowania

Łatwo przeuczyć model

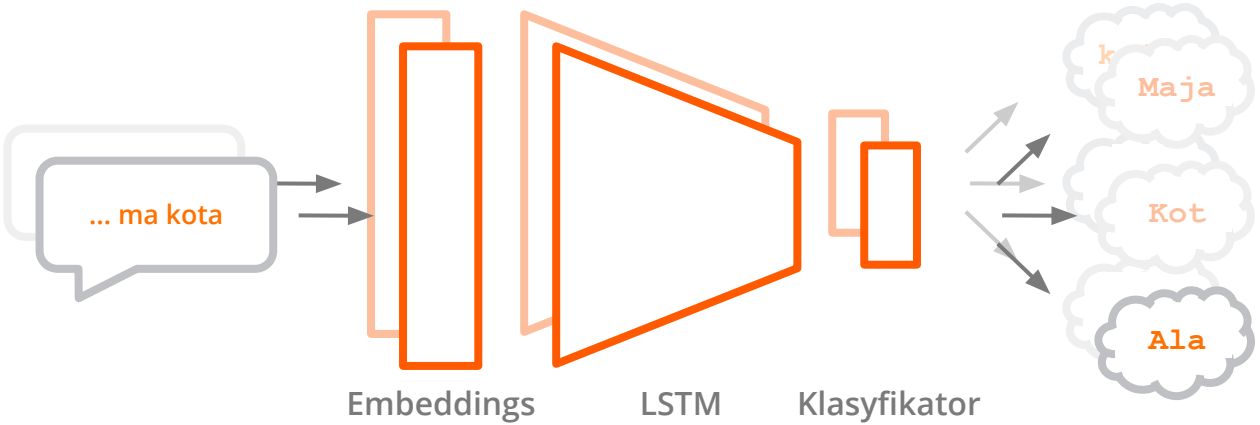
Niska reużywalność

ELMo: Forward Language Modeling



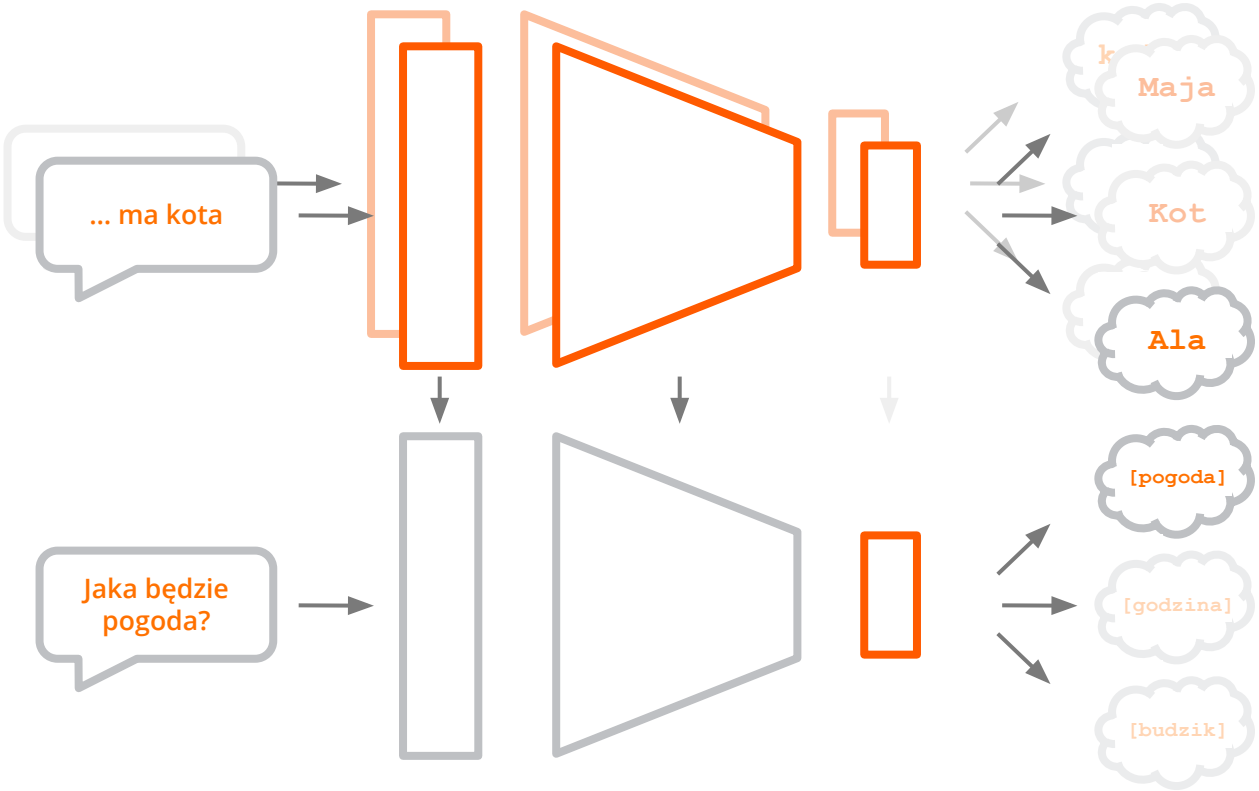
Matthew E. Peters et al., *Deep contextualized word representations*, 2018

ELMo: Backward Language Modeling

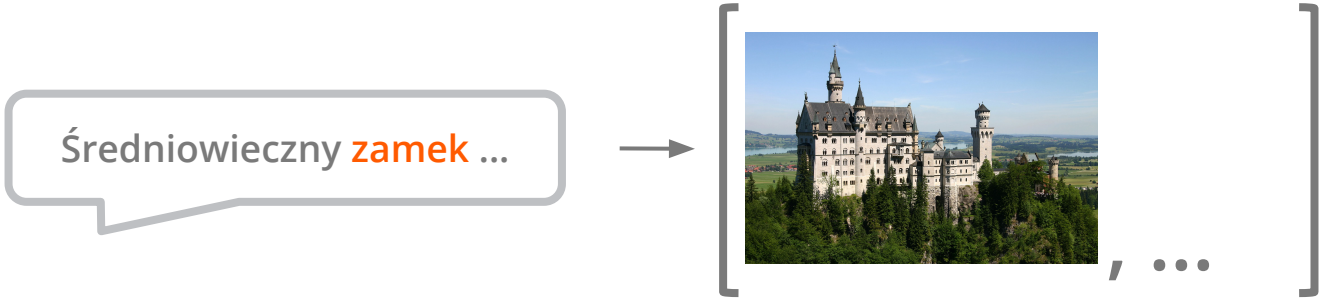


Matthew E. Peters et al., *Deep contextualized word representations*, 2018

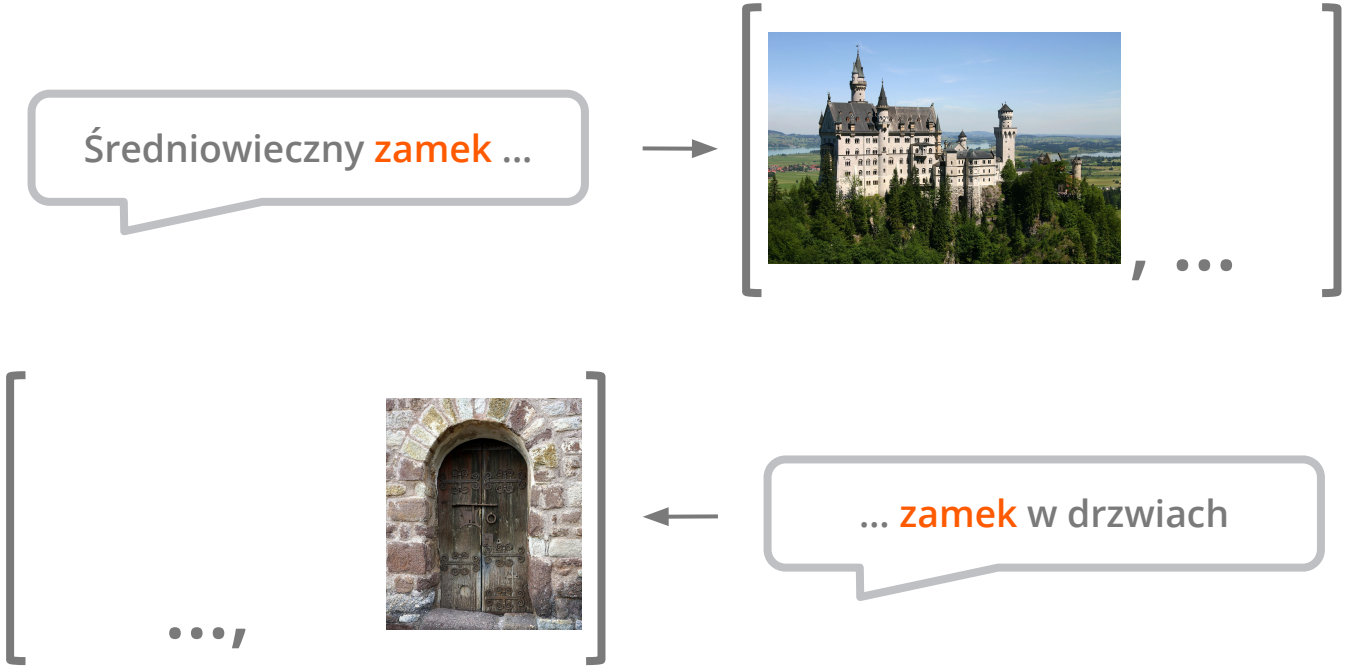
ELMo: Forward + Backward Language Modeling



ELMo: Forward + Backward Language Modeling



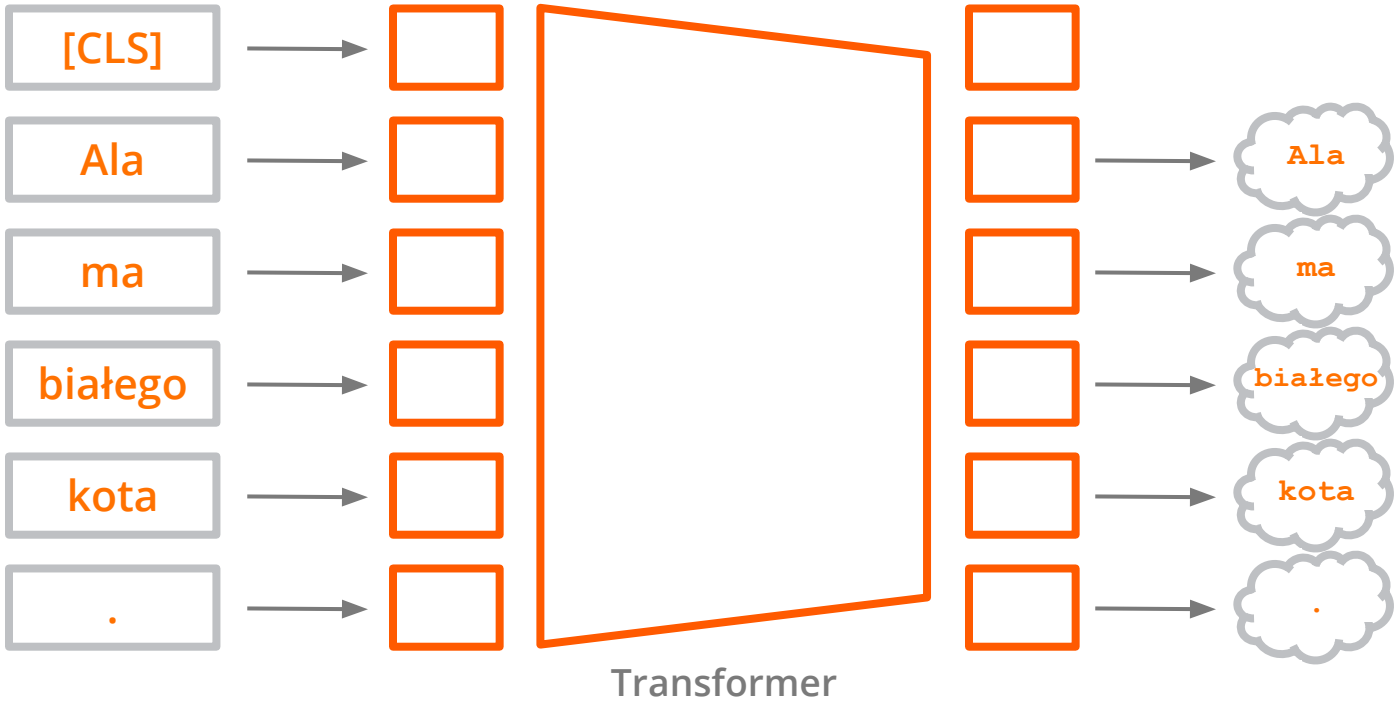
ELMo: Forward + Backward Language Modeling



ELMo: Forward + Backward Language Modeling

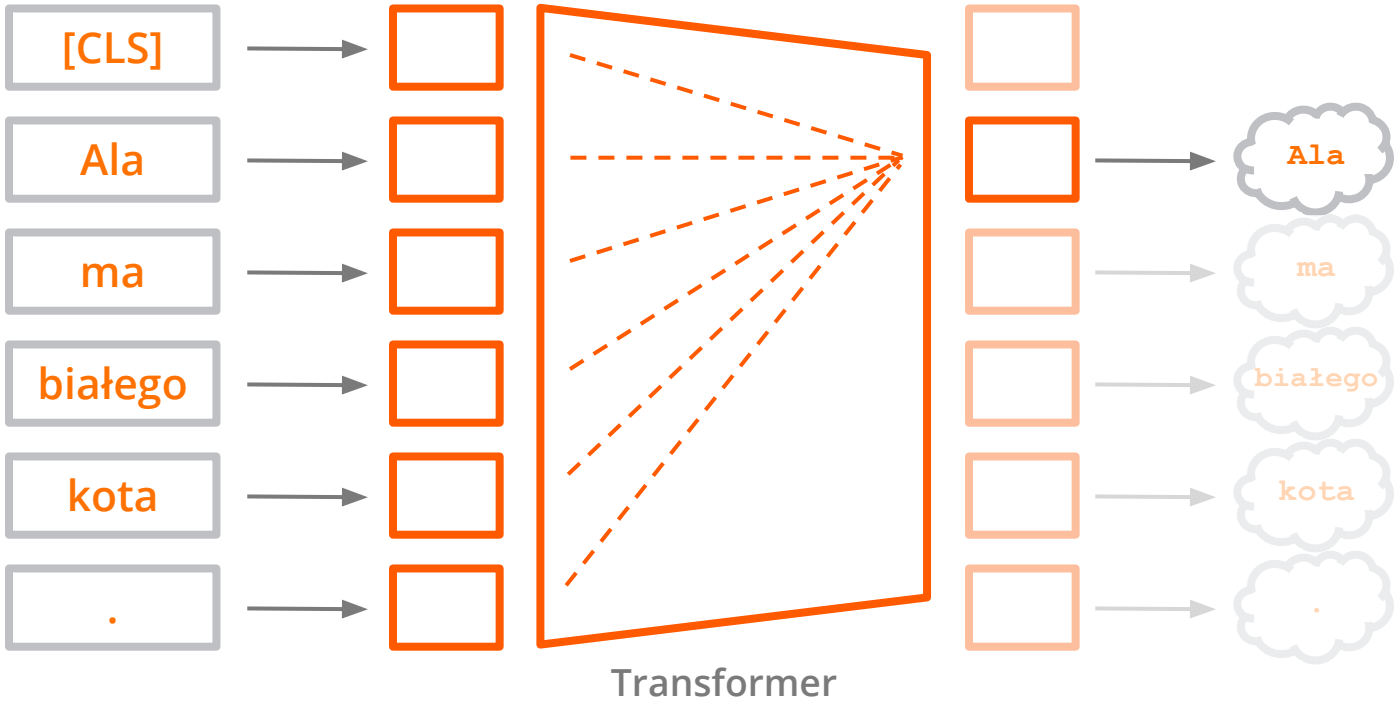
Średniowieczny **zamek** w drzwiach

BERT: Masked Language Modeling



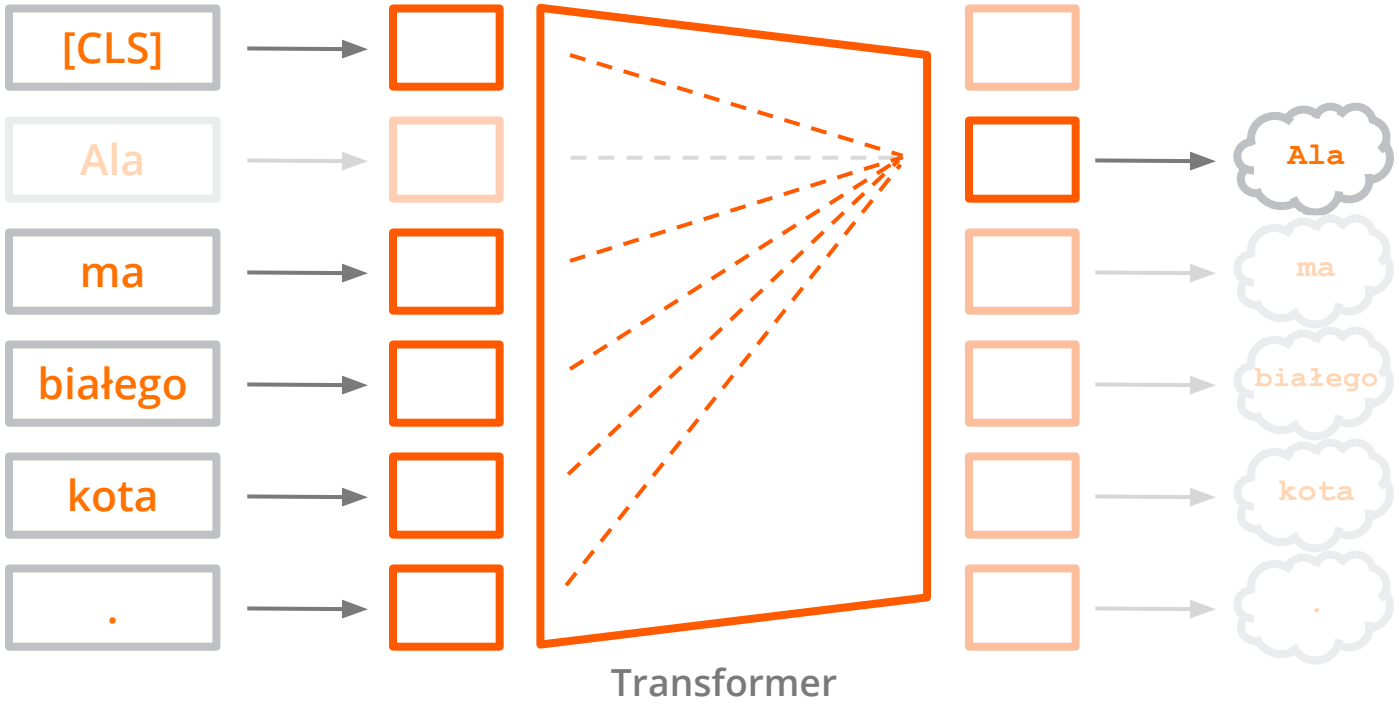
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



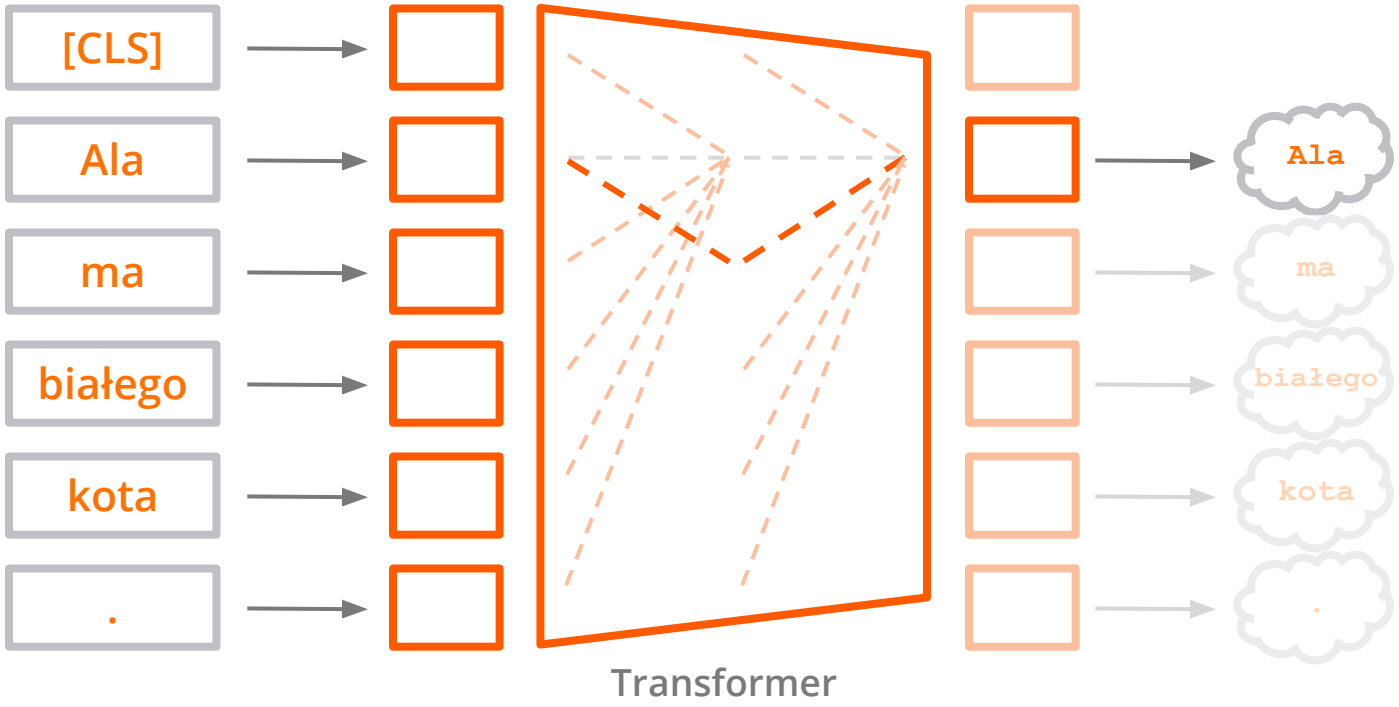
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

BERT: Masked Language Modeling



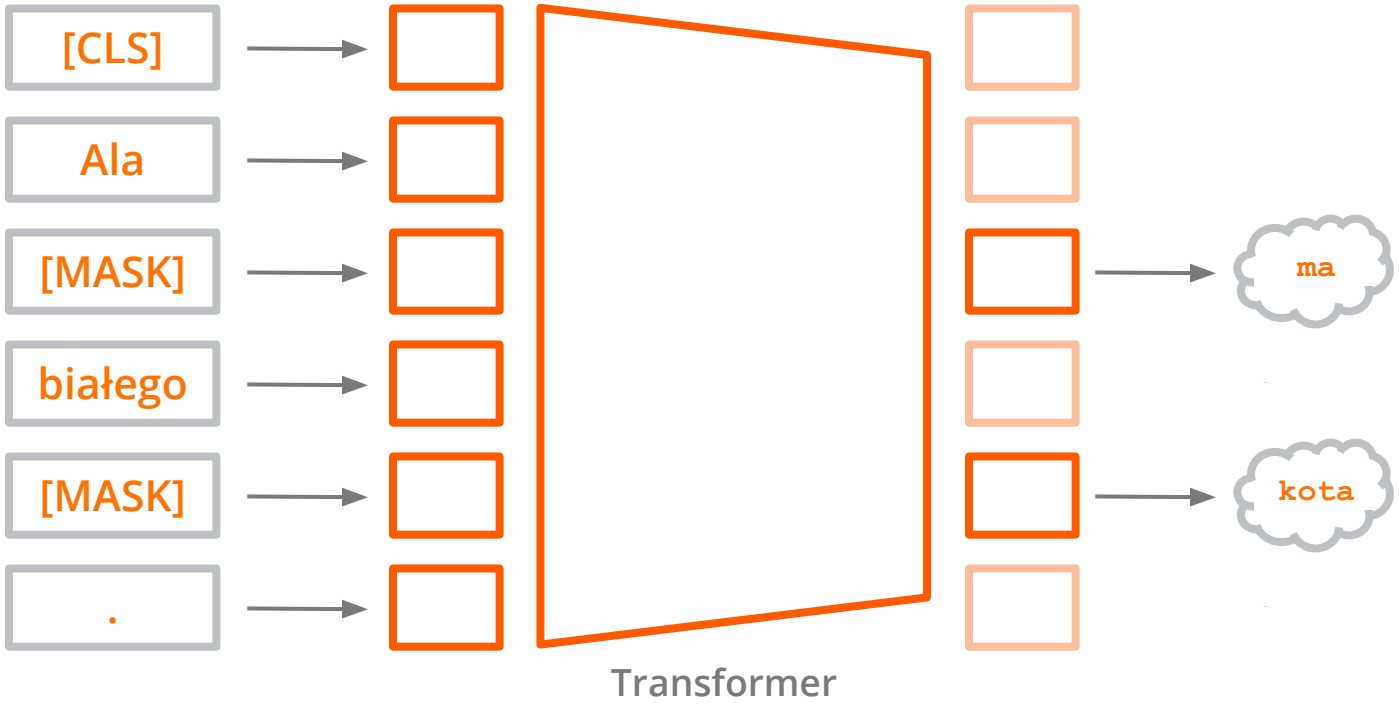
Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018

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% tokenów

15% tokenów

80% [MASK]

Żeby się czegoś nauczył

15% tokenów

80% [MASK]

Żeby się czegoś nauczył

10% ten sam

Żeby tworzył reprezentacje dla wszystkich słów, a nie tylko [MASK]

15% tokenów

80% [MASK]

Żeby się czegoś nauczył

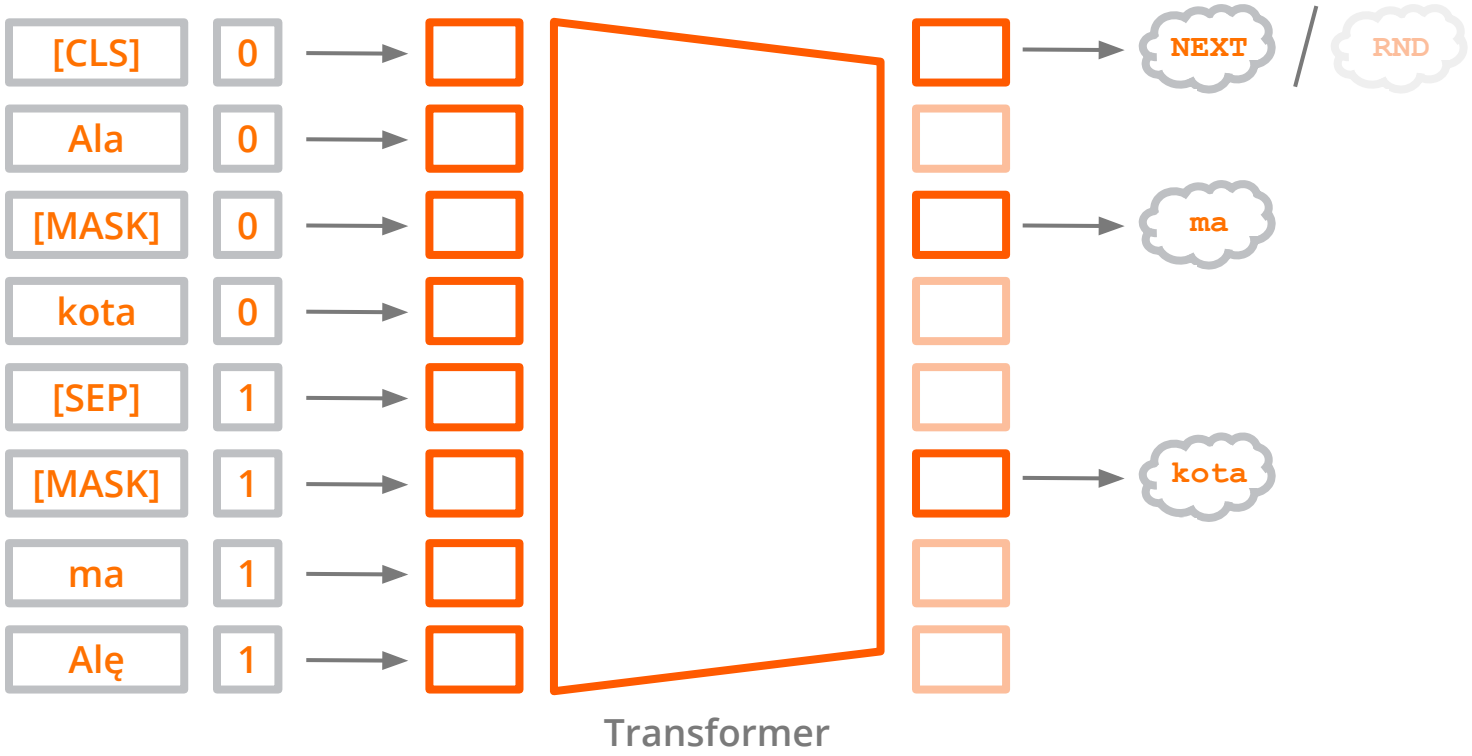
10% ten sam

Żeby tworzył reprezentacje dla wszystkich słów, a nie tylko [MASK]

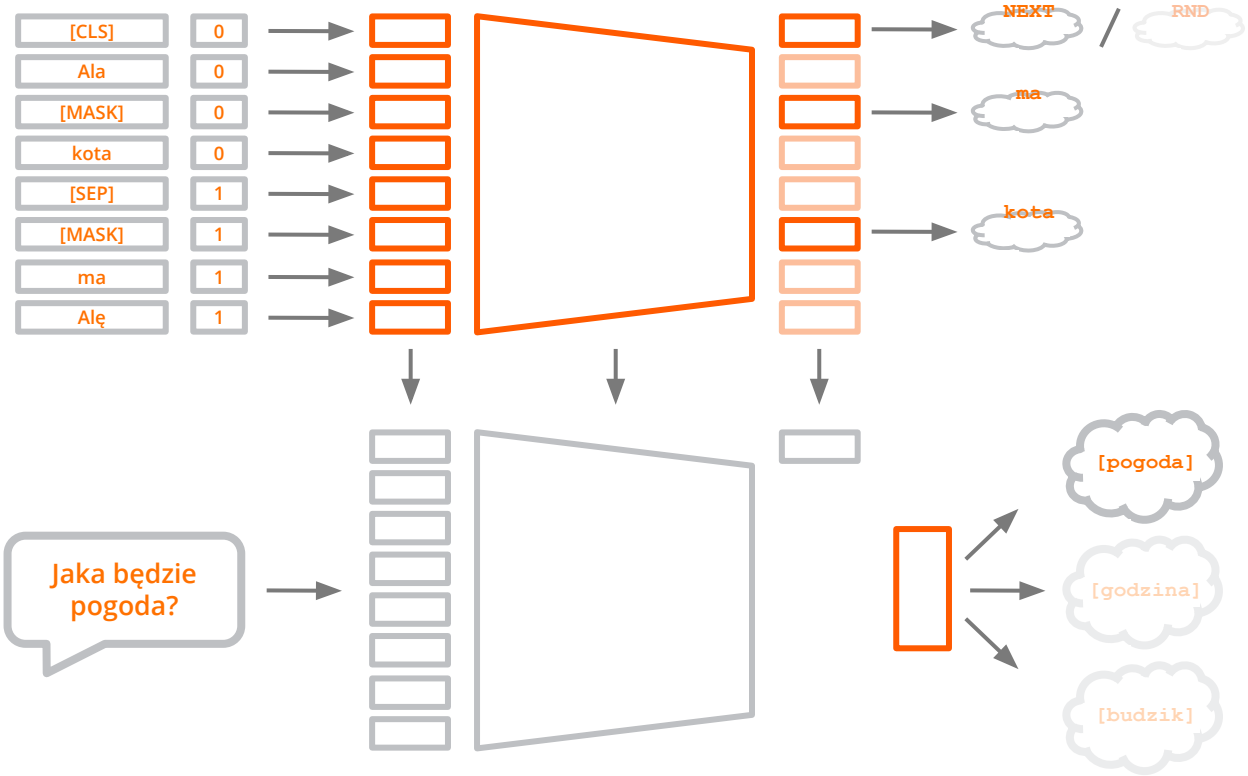
10% losowy

Żeby tworzył reprezentacje dla wszystkich słów, a nie kopiował słowo, które nie jest [MASK]

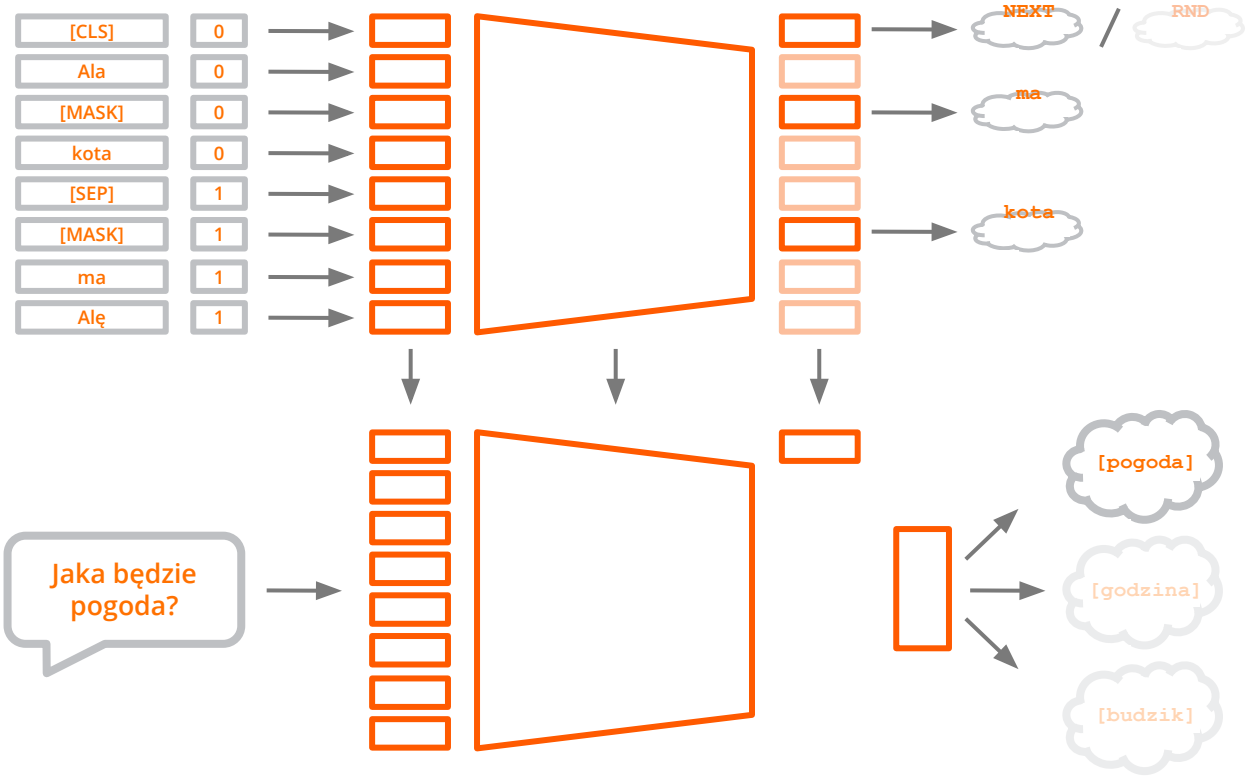
BERT: Next Sentence Prediction



BERT: feature extraction



BERT: fine-tuning



Wyrzucenie Next Sentence Prediction

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

RoBERTa

Wyrzucenie Next Sentence Prediction

Dynamiczne maskowanie tokenów

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

RoBERTa

Wyrzucenie Next Sentence Prediction

Dynamiczne maskowanie tokenów

Większy rozmiar batcha (2048 vs 256)

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

RoBERTa

Wyrzucenie Next Sentence Prediction

Dynamiczne maskowanie tokenów

Większy rozmiar batcha (2048 vs 256)

Więcej danych

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

Dodanie Sentence Order Prediction

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

ALBERT

Dodanie Sentence Order Prediction

Wspólne parametry dla wszystkich warstw

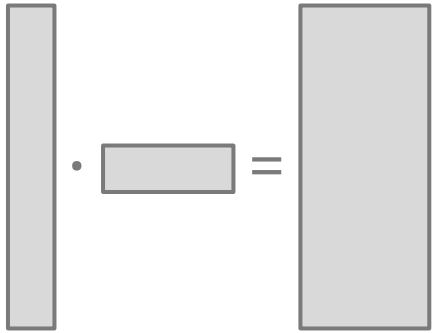
Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

ALBERT

Dodanie Sentence Order Prediction

Wspólne parametry dla wszystkich warstw

Faktoryzacja embeddingów



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

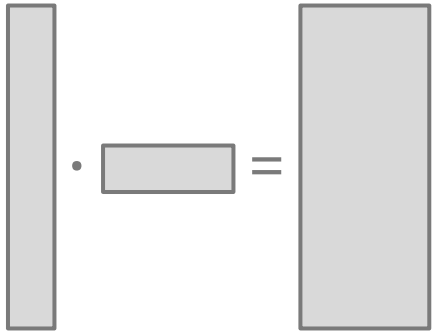
ALBERT

Dodanie Sentence Order Prediction

Wspólne parametry dla wszystkich warstw

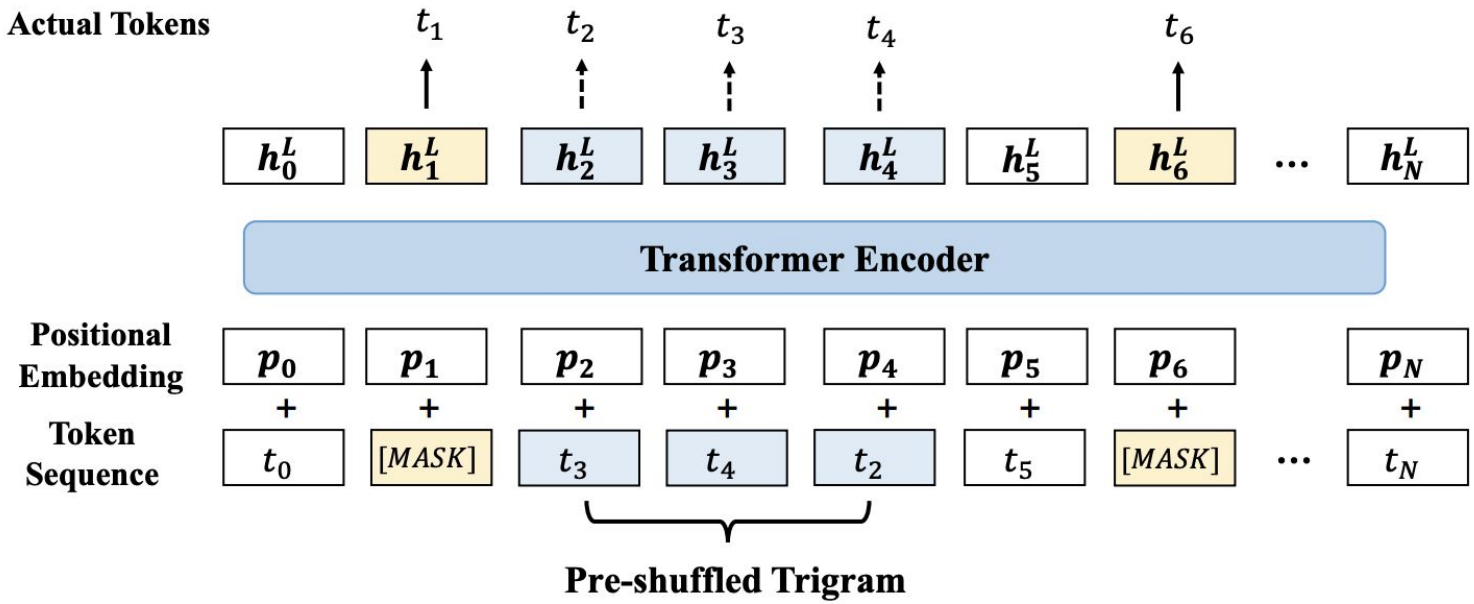
Faktoryzacja embeddingów

Więcej warstw



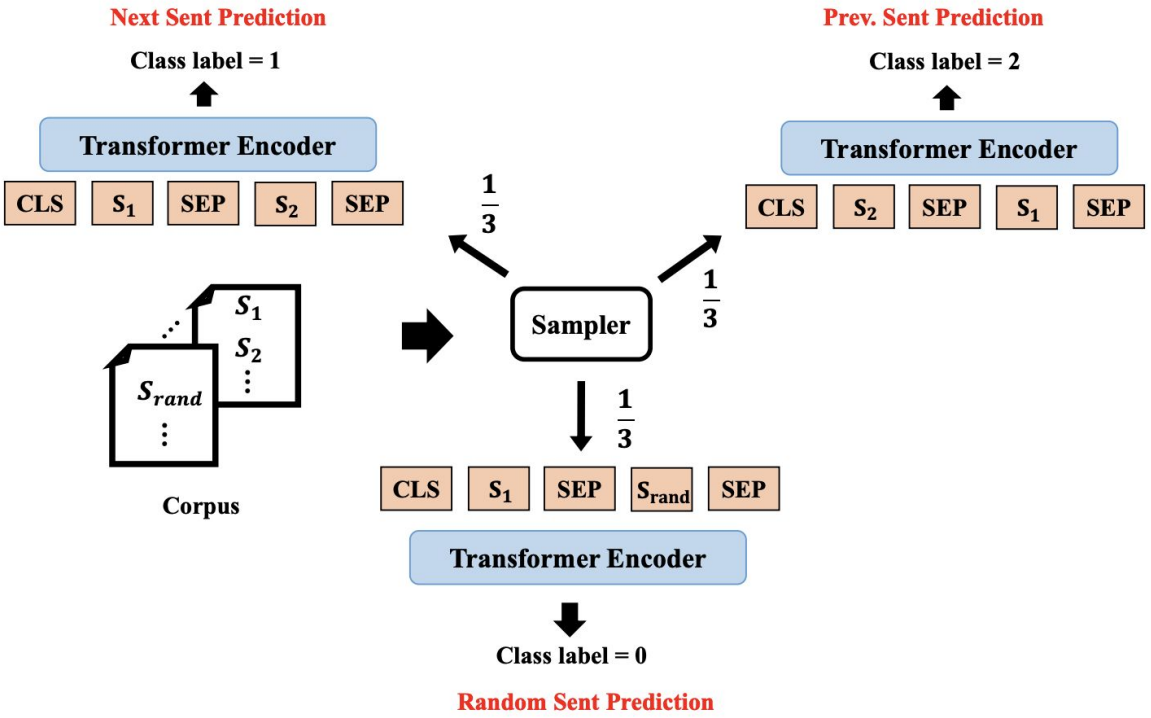
Zhenzhong Lan et al., *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*, 2019

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

BPE-Dropout

```
vocab = [“a”, …, “z”]
for i in range(size):
    c1, c2 = most_common(
        corpus,
        vocab,
    )

    vocab.append(c1 + c2)
```

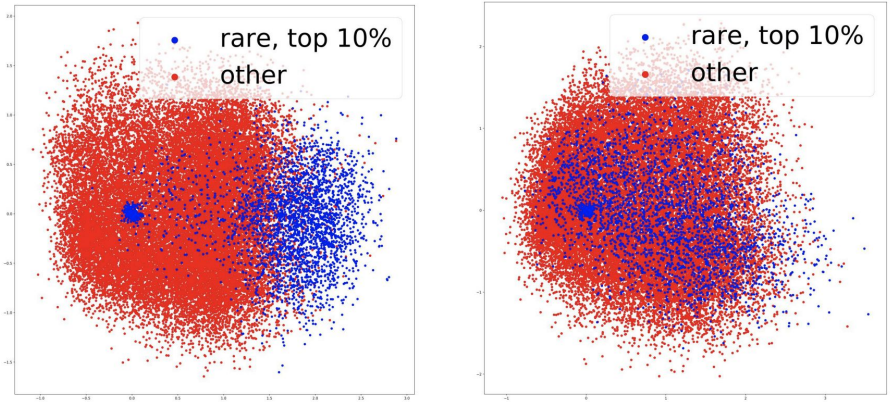
```
u-n-r-e-l-a-t-e-d
u-n re-l-a-t-e-d
u-n re-l-at-e-d
u-n re-l-at-ed
un re-l-at-ed
un re-l-ated
un rel-ated
un-related
unrelated
```

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

BPE-Dropout

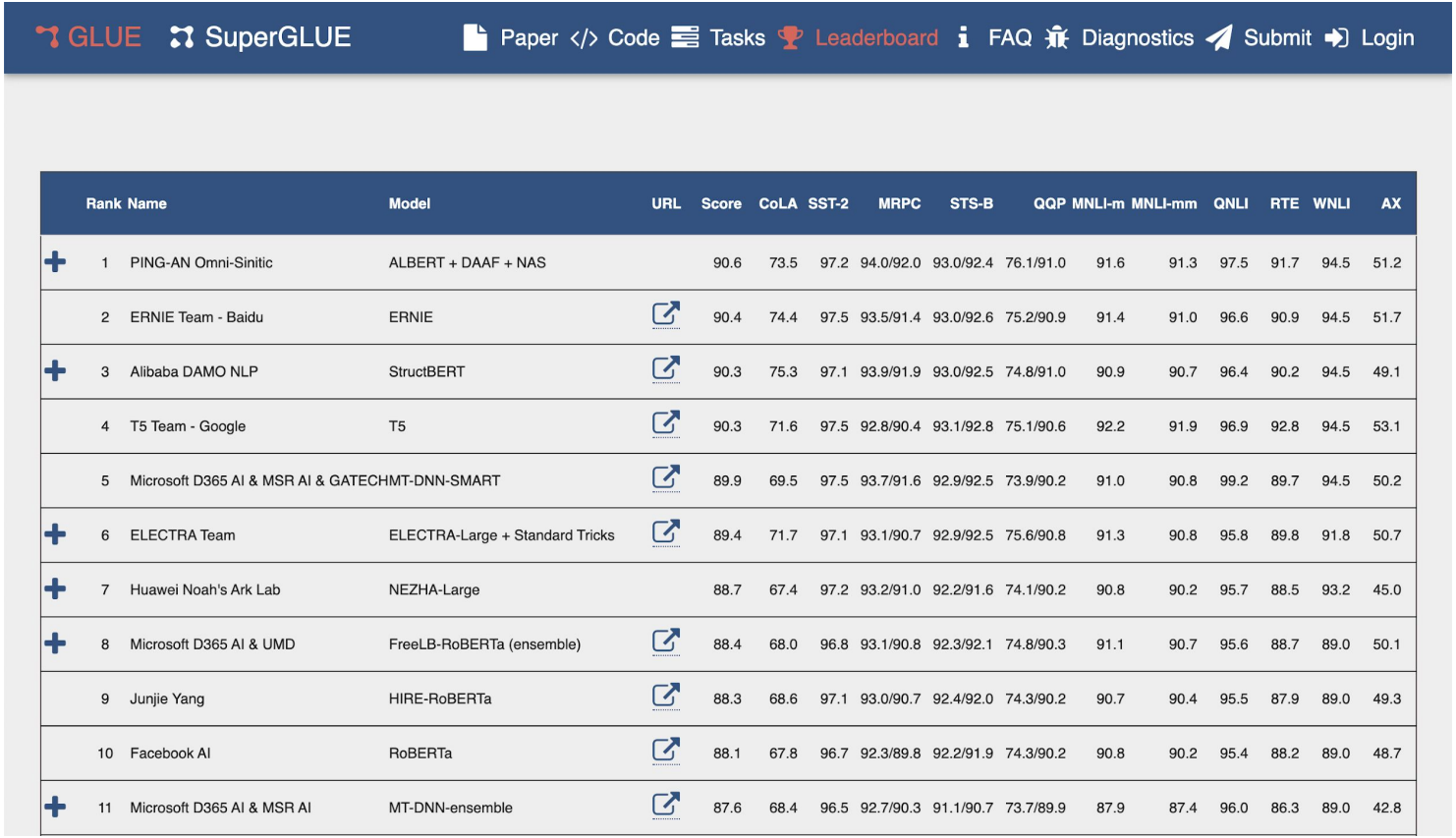
```
u-n_r-e-l-a_t-e_d
u-n re-l_a-t-e_d
u-n re_l-at-e_d
un re-l-at-e-d
un re_l-at-ed
un re-lat-ed
un relat_ed
```

```
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



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

BERT vs RoBERTa?

Cased vs Uncased?

Batch size?

Vocab size?

Base vs Large?

Which corpora?

https://www.flickr.com/photos/theresasthompson/7163227255

Kompleksowa
Lista
Ewaluacji
Językowych

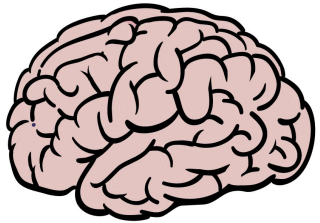
Piotr Rybak, Robert Mroczkowski, Janusz Tracz, Ireneusz Gawlik, *KLEJ: Comprehensive Benchmark for Polish Language Understanding*, 2020

9 Zadań

7 Zbiorów danych

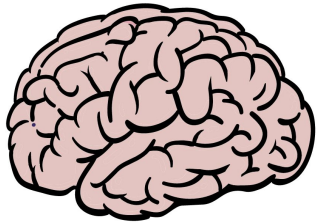
1 Interfejs

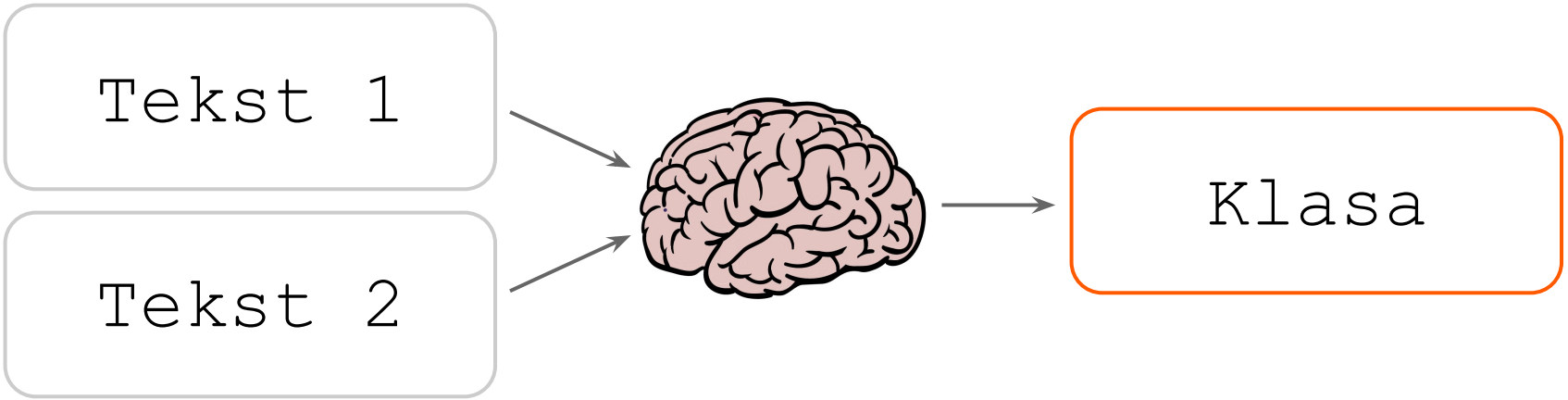
Tekst 1

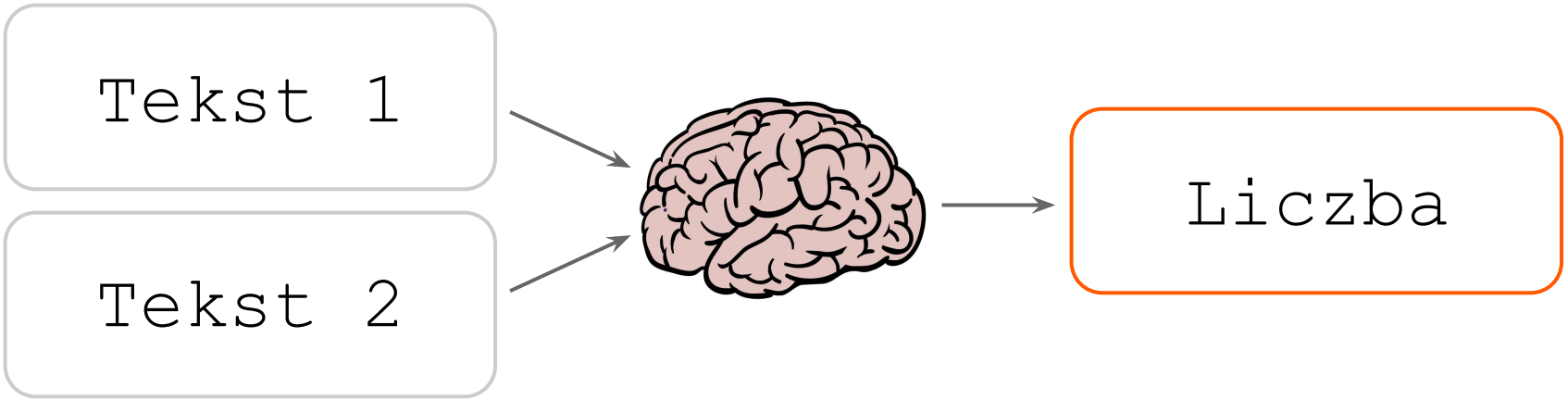


Tekst 1

Tekst 2



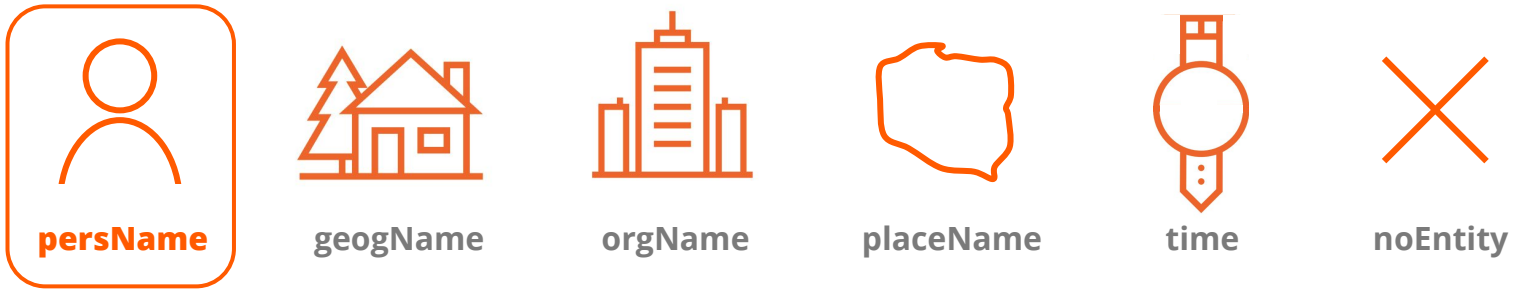

Tekst 1
Tekst 2
Klasa


Tekst 1
Tekst 2
Liczba

Spójrzmy na zadania!

NKJP: Named Entity Classification

```
– klej to się zepsuje panie Andrzeju, kupiłem worek kleju.
```



Adam Przepiórkowski, *Narodowy korpus języka polskiego*, 2012

CDSC: Relatedness

Dwie osoby przyglądają się biegnącemu mężczyźnie
z numerem, który jest przyklejony do klatki piersiowej.

Dwie osoby przyglądają się biegnącemu mężczyźnie
z przyklejonym numerem do klatki piersiowej.

Alina Wróblewska and Katarzyna Krasnowska-Kieraś, *Polish evaluation dataset for compositional distributional semantics models*, 2017

CDSC: Entailment

Dwie osoby przyglądają się biegnącemu mężczyźnie
z numerem, który jest przyklejony do klatki piersiowej.

Dwie osoby przyglądają się biegnącemu mężczyźnie
z przyklejonym numerem do klatki piersiowej.

Contradiction

Neutral

Entailment

Alina Wróblewska and Katarzyna Krasnowska-Kieraś, *Polish evaluation dataset for compositional distributional semantics models*, 2017

Cyberbullying Detection

```
Gdzie moi koledzy @anonymized_account @anonymized_account
pewnie odurzają się gdzieś klejem na dworcu
```

Michal Ptaszynski et al., *Results of the PolEval 2019 Shared Task 6: First Dataset and Open Shared Task for Automatic Cyberbullying Detection in Polish Twitter*, 2019

PolEmo 2.0: In-domain [Hotel & Medicine]

Przejrzeliśmy klejącą się kartę menu i zrezygnowaliśmy z dalszego posiłku. Wszędzie kurz, może to taki styl - nie wiem. Dla mnie spore rozczarowanie, gdyż spodziewałam się pięknej restauracji a nie wszechobecnego kurzu i zapachu jedzenia jak w podrzędnej jadłodajni.

Jan Kocoń et al., *Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews*, 2019

PolEmo 2.0: Out-of-domain [Product & University]

Z zewnątrz nawet nieźle wyglądają i to chyba ich największa zaleta. Nie najlepiej wykonane, niewygodne, kabel za krótki, dźwięk może być. Wszystko co mogło zostać wykonane z marnego plastiku, zostało z niego wykonane i połączone tanim klejem oby się jakoś trzymało.

Jan Kocoń et al., *Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews*, 2019

Did You Know?

Czy guma arabska nadaje się do spożycia?

Na przykład śluz z drzewa wiśniowego (klej wiśniowy) lub z drzewa akacji senegalskiej (tzw. guma arabska) są stosowane do wyrobu klejów.

Correct

Incorrect

Michał Marcinczuk et al., *Open dataset for development of Polish Question Answering systems*, 2013

Polish Summaries Corpus

Kierowcy, którzy po pierwszym maja będą chcieli zarejestrować samochód, dostaną białą tablicę, naklejki legalizacyjne i nowy dowód rejestracyjny. latem 1997 roku na światło dzienne wypłynął projekt rozporządzenia w sprawie tablic. Według projektu na produkcję tablic miały być tylko trzy koncesje. taki projekt wywołał przerażenie wśród drobnych wytwórców. Ostatecznie - nie ma koncesji, wystarczy certyfikat Instytutu Transportu Samochodowego, który dostało blisko sześćdziesiąt firm.

W Niemczech jest ponad 10 tysięcy tanich sklepów, w Polsce na razie - kilkaset. Choć tanie tzw. dyskontowe sklepy nie budzą na razie takich emocji wśród rodzimych kupców, jak hipermarkety, to zdobywają coraz większy udział w polskim handlu detalicznym i to udział zdominowany przez zachodnie sieci handlowe. Ocenia się, że tanie sklepy będą - obok hipermarketów - jednym z dwóch głównych kanałów dystrybucji towarów. Ich główną zaletą, poza niskimi cenami, jest to, że są położone blisko.

Similar

Not Similar

Maciej Ogrodniczuk and Mateusz Kopeć, *The Polish Summaries Corpus*, 2014

Allegro Reviews

Bardzo fajny uchwyt ale klej ktory znajduje sie na czesci ktora przyklejamy do telefonu jest bardzo slaby i odkleja sie ciagle mimo ze odtłuscilem telefon przed naklejeniem. Oprócz tego - bez zastrzezen, uchwyt godny polecenia, dyskretny, w aucie prezentuje sie super, magnes utrzymuje bez problemu 6"" telefon, nie spada, trzyma sie ok.

Piotr Rybak et al., *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

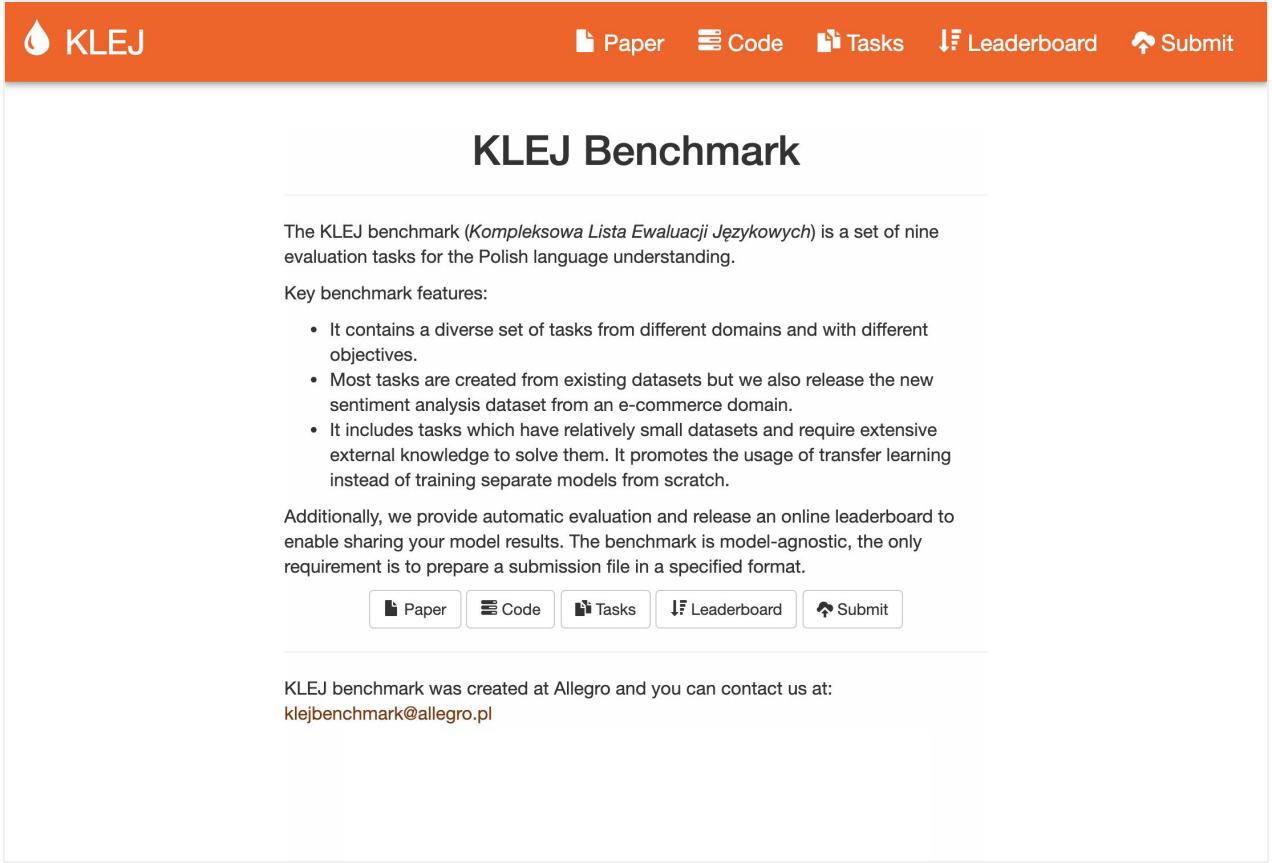
Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Random	28.3	20.7	59.2	0.9	11.2	27.8	28.5	18.9	30.4	56.9

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Random	28.3	20.7	59.2	0.9	11.2	27.8	28.5	18.9	30.4	56.9
LSTM	63.0	45.0	87.5	84.7	20.7	79.6	60.7	22.3	84.4	81.8
LSTM + fastText	67.7	67.3	87.8	81.6	32.8	83.2	61.1	27.5	86.5	81.4
LSTM + ELMo	76.6	93.0	88.9	90.4	50.2	88.5	72.1	28.8	92.7	85.1
LSTM + ELMo + fine-tune	76.7	**93.4**	89.3	91.1	47.9	87.4	70.6	30.9	93.7	**86.2**
LSTM + ELMo + attention	75.8	93.0	90.0	90.3	46.8	88.8	70.2	26.0	92.1	85.4

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Random	28.3	20.7	59.2	0.9	11.2	27.8	28.5	18.9	30.4	56.9
LSTM	63.0	45.0	87.5	84.7	20.7	79.6	60.7	22.3	84.4	81.8
LSTM + fastText	67.7	67.3	87.8	81.6	32.8	83.2	61.1	27.5	86.5	81.4
LSTM + ELMo	76.6	93.0	88.9	90.4	50.2	88.5	72.1	28.8	92.7	85.1
LSTM + ELMo + fine-tune	76.7	**93.4**	89.3	91.1	47.9	87.4	70.6	30.9	93.7	**86.2**
LSTM + ELMo + attention	75.8	93.0	90.0	90.3	46.8	88.8	70.2	26.0	92.1	85.4
Multi-BERT	79.5	91.4	**93.8**	92.9	40.0	85.0	66.6	**64.2**	97.9	83.3
Slavic-BERT	79.8	93.3	93.7	**93.3**	43.1	87.1	67.6	57.4	**98.3**	84.3
XLM-17	80.2	91.9	93.7	92.0	44.8	86.3	70.6	61.8	96.3	84.5

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Random	28.3	20.7	59.2	0.9	11.2	27.8	28.5	18.9	30.4	56.9
LSTM	63.0	45.0	87.5	84.7	20.7	79.6	60.7	22.3	84.4	81.8
LSTM + fastText	67.7	67.3	87.8	81.6	32.8	83.2	61.1	27.5	86.5	81.4
LSTM + ELMo	76.6	93.0	88.9	90.4	50.2	88.5	72.1	28.8	92.7	85.1
LSTM + ELMo + fine-tune	76.7	**93.4**	89.3	91.1	47.9	87.4	70.6	30.9	93.7	**86.2**
LSTM + ELMo + attention	75.8	93.0	90.0	90.3	46.8	88.8	70.2	26.0	92.1	85.4
Multi-BERT	79.5	91.4	**93.8**	92.9	40.0	85.0	66.6	**64.2**	97.9	83.3
Slavic-BERT	79.8	93.3	93.7	**93.3**	43.1	87.1	67.6	57.4	**98.3**	84.3
XLM-17	80.2	91.9	93.7	92.0	44.8	86.3	70.6	61.8	96.3	84.5
HerBERT	**80.5**	92.7	92.5	91.9	**50.3**	**89.2**	**76.3**	52.1	95.3	84.5

Table 3: Baseline evaluation on KLEJ benchmark. *AVG* is the average score across all tasks.


KLEJ
Paper
Code
Tasks
Leaderboard
Submit
KLEJ Benchmark
The KLEJ benchmark (*Kompleksowa Lista Ewaluacji Językowych*) is a set of nine evaluation tasks for the Polish language understanding.
Key benchmark features:
- It contains a diverse set of tasks from different domains and with different objectives.
- Most tasks are created from existing datasets but we also release the new sentiment analysis dataset from an e-commerce domain.
- It includes tasks which have relatively small datasets and require extensive external knowledge to solve them. It promotes the usage of transfer learning instead of training separate models from scratch.
Additionally, we provide automatic evaluation and release an online leaderboard to enable sharing your model results. The benchmark is model-agnostic, the only requirement is to prepare a submission file in a specified format.
Paper
Code
Tasks
Leaderboard
Submit
KLEJ benchmark was created at Allegro and you can contact us at:
klejbenchmark@allegro.pl

NKJP-NER

The *NKJP-NER* is based on a human-annotated part of NKJP. We extracted sentences with named entities of exactly one type. The task is to predict the type of the named entity.

More info Download

License:

```
GNU GPL v.3
```

Original citation:

```
@book{przepiorkowski2012narodowy,
title={Narodowy korpus j{\k{e}}zyka polskiego},
author={Przepi{\'o}rkowski, Adam},
year={2012},
publisher={Naukowe PWN}
}
```

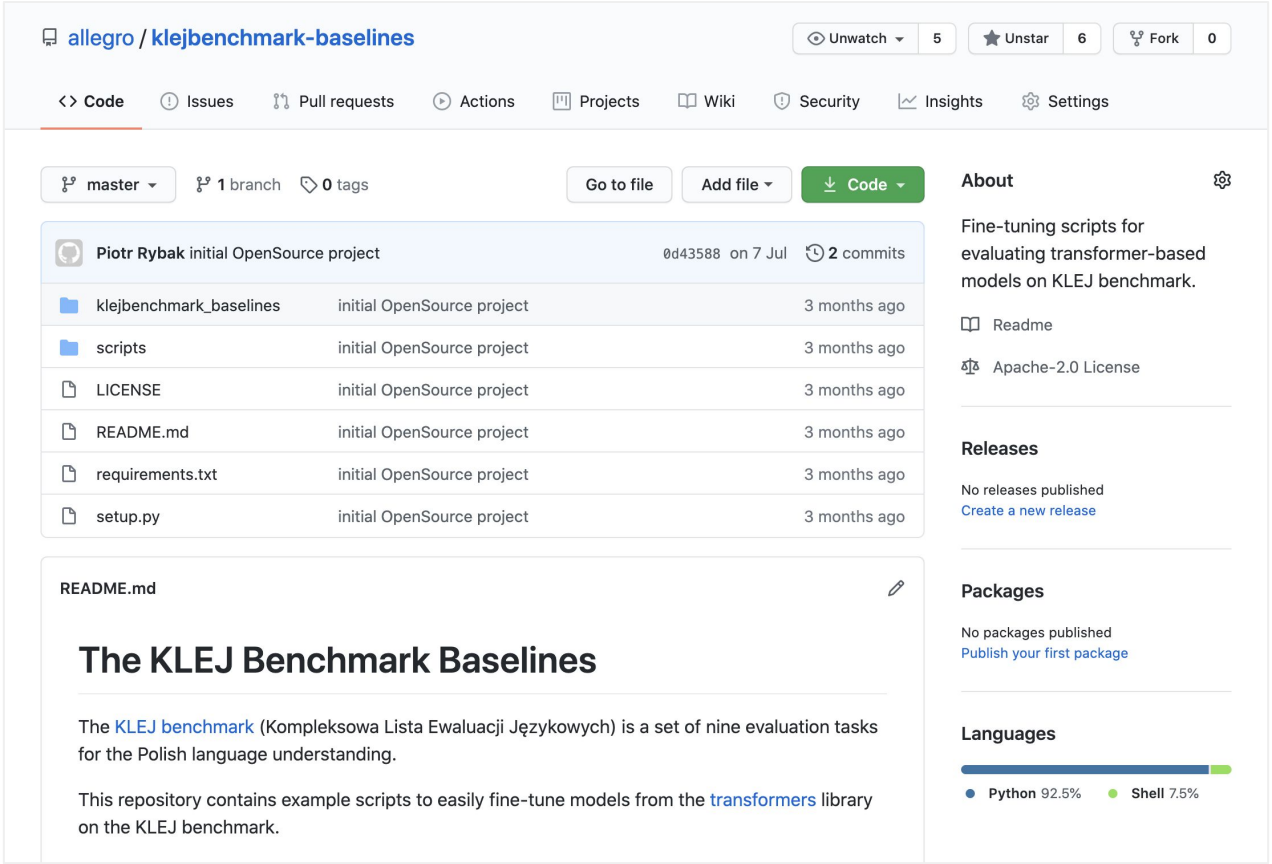
CDSC-E

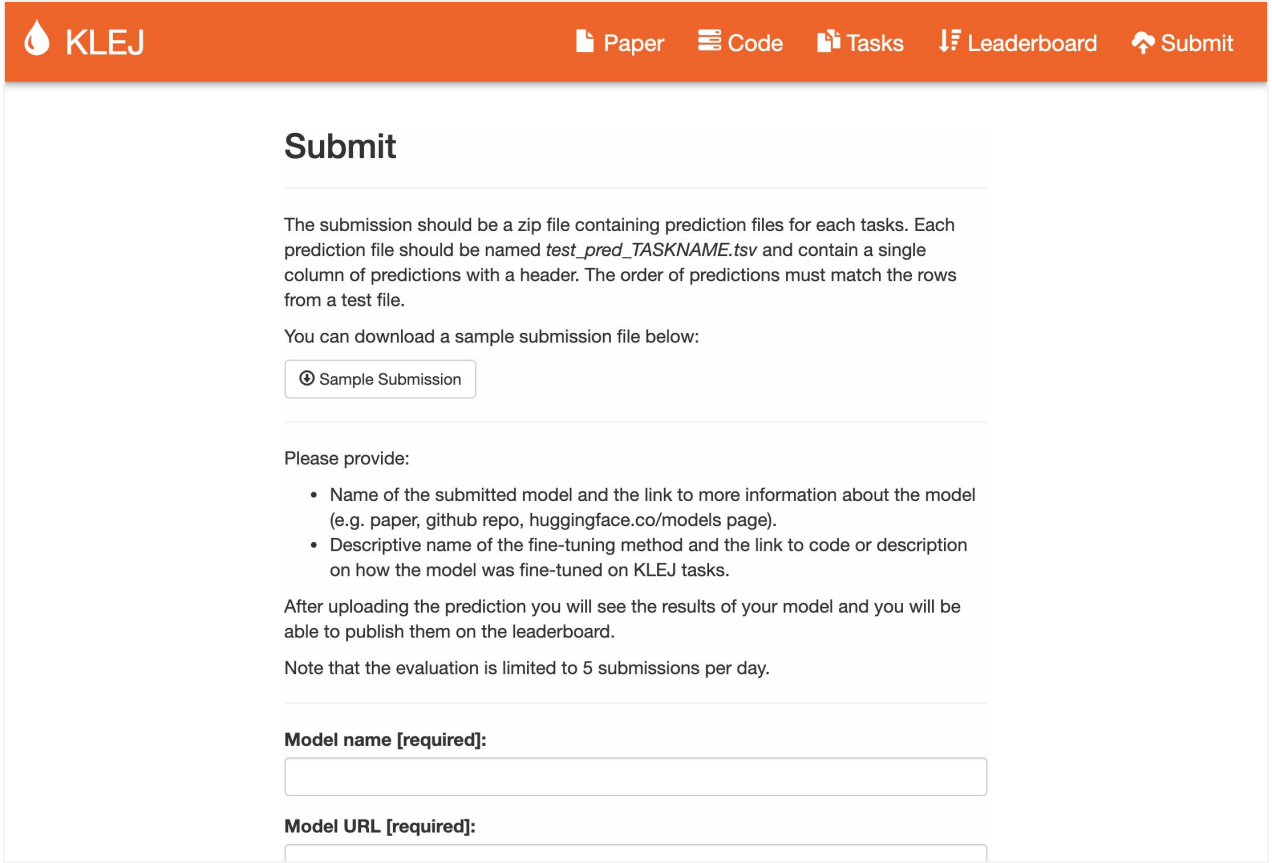
The *Compositional Distributional Semantics Corpus* consists of pairs of sentences which are human-annotated for their entailment.

More info Download

License:

```
CC BY-NC-SA 4.0
```


allegro / klejbenchmark-baselines
Unwatch 5
Unstar 6
Fork 0
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
master
1 branch
0 tags
Go to file
Add file
Code
Piotr Rybak initial OpenSource project
0d43588 on 7 Jul
2 commits
klejbenchmark_baselines initial OpenSource project 3 months ago
scripts initial OpenSource project 3 months ago
LICENSE initial OpenSource project 3 months ago
README.md initial OpenSource project 3 months ago
requirements.txt initial OpenSource project 3 months ago
setup.py initial OpenSource project 3 months ago
README.md
The KLEJ Benchmark Baselines
The KLEJ benchmark (Kompleksowa Lista Ewaluacji Językowych) is a set of nine evaluation tasks for the Polish language understanding.
This repository contains example scripts to easily fine-tune models from the transformers library on the KLEJ benchmark.
About
Fine-tuning scripts for evaluating transformer-based models on KLEJ benchmark.
Readme
Apache-2.0 License
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Languages
Python 92.5%
Shell 7.5%


KLEJ
Paper
Code
Tasks
Leaderboard
Submit
Submit
The submission should be a zip file containing prediction files for each tasks. Each prediction file should be named test_pred_TASKNAME.tsv and contain a single column of predictions with a header. The order of predictions must match the rows from a test file.
You can download a sample submission file below:
Sample Submission
Please provide:
Name of the submitted model and the link to more information about the model (e.g. paper, github repo, huggingface.co/models page).
Descriptive name of the fine-tuning method and the link to code or description on how the model was fine-tuned on KLEJ tasks.
After uploading the prediction you will see the results of your model and you will be able to publish them on the leaderboard.
Note that the evaluation is limited to 5 submissions per day.
Model name [required]:
Model URL [required]:

KLEJ

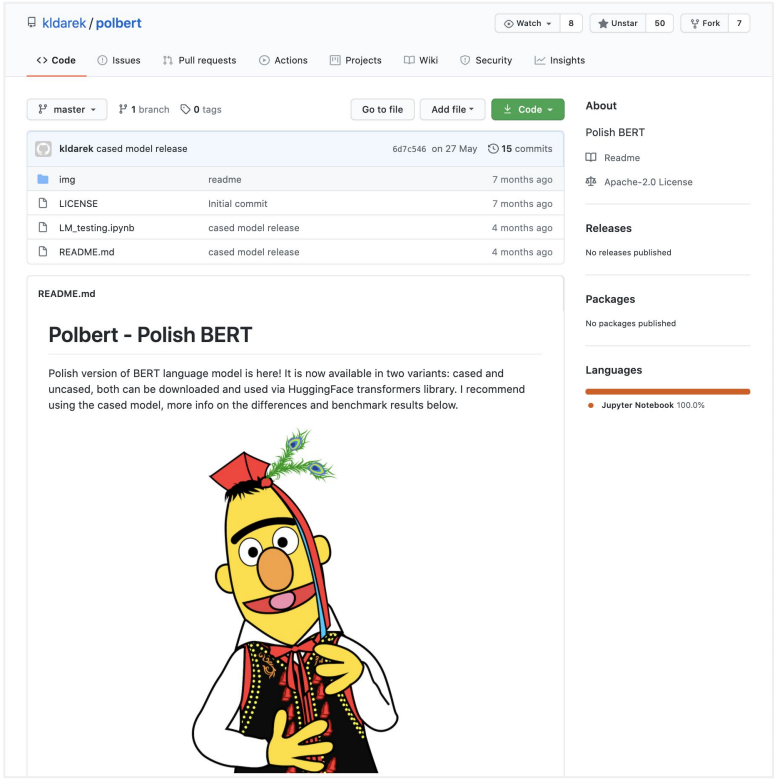
Paper Code Tasks Leaderboard Submit

Leaderboard

Rank	Model	Fine-tuning	Average	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	DYK	PSC	AR
1	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
2	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
3	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
4	XLM-RoBERTa (large)	FT2	86.6	94.6	94.4	94.7	67.9	90.4	79.8	71.6	98.2	87.5
5	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
6	XLM-RoBERTa (large)	v0.1.0	84.7	94.6	94.4	94.7	50.7	90.4	79.8	71.6	98.2	87.5
7	Polbert	v0.1.0	81.7	93.6	93.4	93.8	52.7	87.4	71.1	59.1	98.6	85.2

allegro

Polbert



Model

- BERT Base

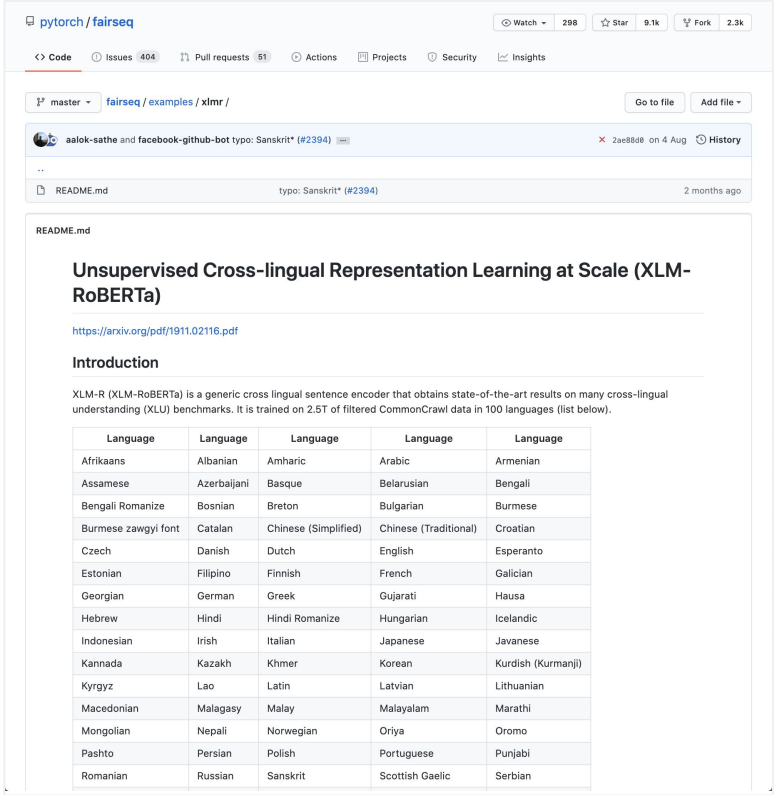
Korpus

- Open Subtitles
- ParaCrawl
- Korpus Parlamentarny
- Wikipedia

KLEJ: 81.7

Dariusz Kłeczek, *https://github.com/kldarek/polbert*

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	Norwegian	Oriya	Oromo
Pashto	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	Sanskrit	Scottish Gaelic	Serbian

Model

- RoBERTa Base & Large

Korpus

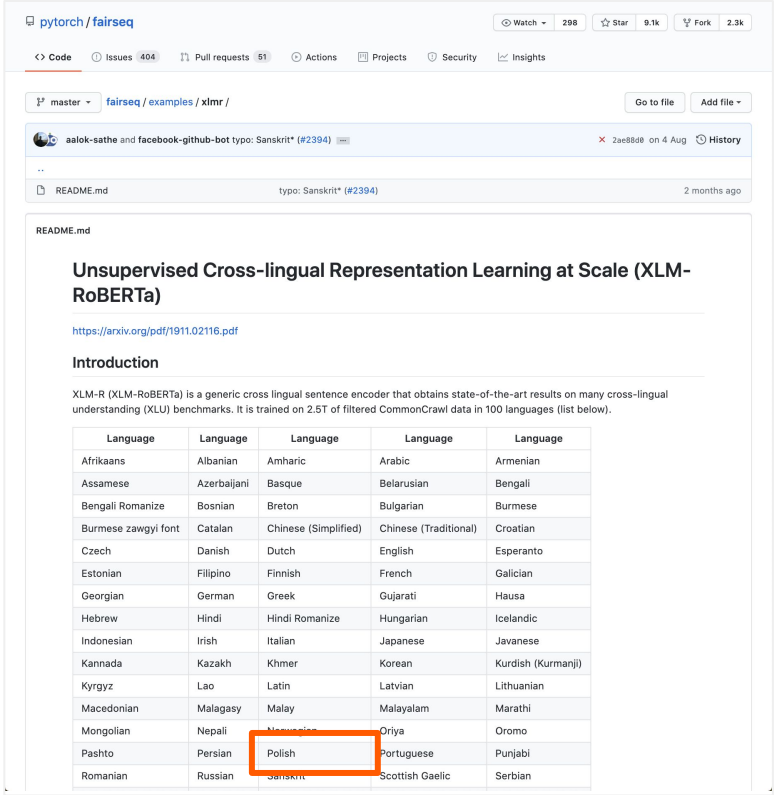
- Common Crawl

KLEJ

- Base: 81.5
- Large: 84.7

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

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	Norwegian	Oriya	Oromo
Pashto	Persian	Polish	Portuguese	Punjabi
Romanian	Russian	Sanskrit	Scottish Gaelic	Serbian

Model

- RoBERTa Base & Large

Korpus

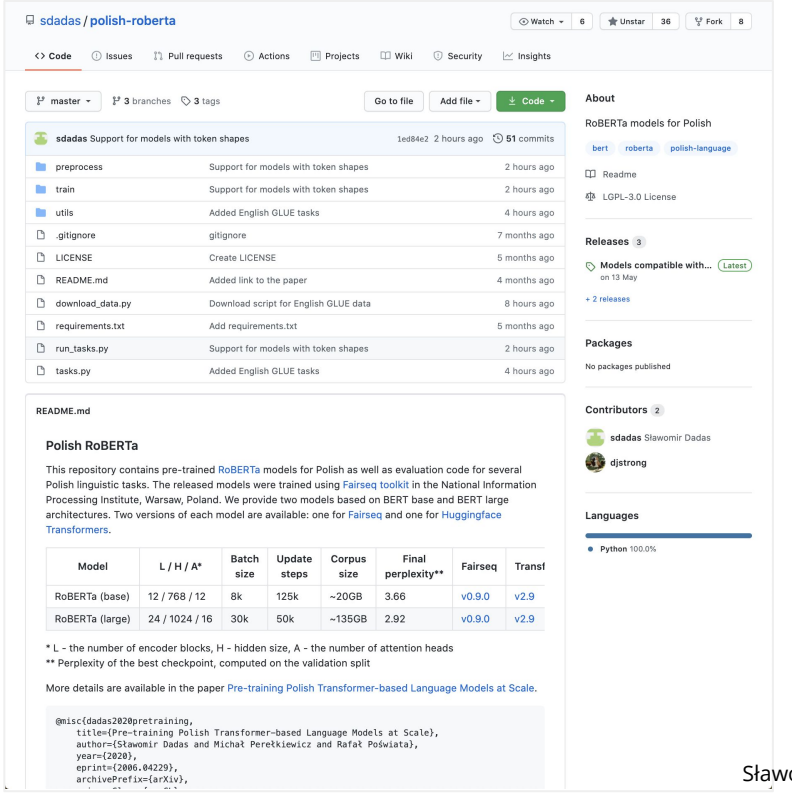
- Common Crawl

KLEJ

- Base: 81.5
- Large: 84.7

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

Polish RoBERTa



Model

- RoBERTa Base & Large

Korpus

- Common Crawl
- Korpus Parlamentarny
- Wikipedia
- Inne

KLEJ: 85.3 & 87.8

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

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**

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

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**

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

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

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

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

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

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

Krzysztof Wróbel, *https://github.com/k-nlp/klej-submission#xlmrnkjp*

HerBERT 2.0: Korpus

Small

Corpus	Tokens	Documents	Avg len
Source Corpora			
NKJP	1357M	3.9M	347
Wikipedia	260M	1.4M	190
Wolne Lektury	41M	5.5k	7447
Final Corpora			
Small	1658M	5.3M	313

HerBERT 2.0: Korpus

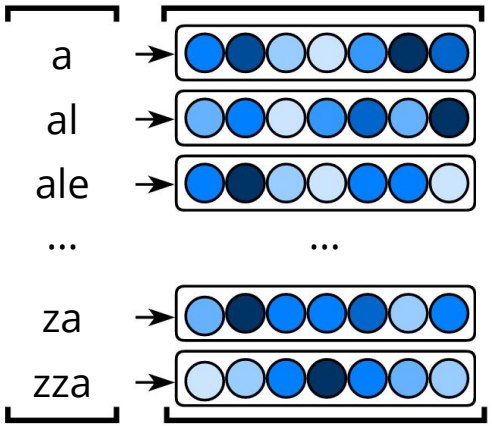
Large

Corpus	Tokens	Documents	Avg len
Source Corpora			
NKJP	1357M	3.9M	347
Wikipedia	260M	1.4M	190
Wolne Lektury	41M	5.5k	7447
CCNet Head	2641M	7.0M	379
CCNet Middle	3243M	7.9M	409
Open Subtitles	1056M	1.1M	961
Final Corpora			
Small	1658M	5.3M	313
Large	8599M	21.3M	404

HerBERT 2.0: Korpus

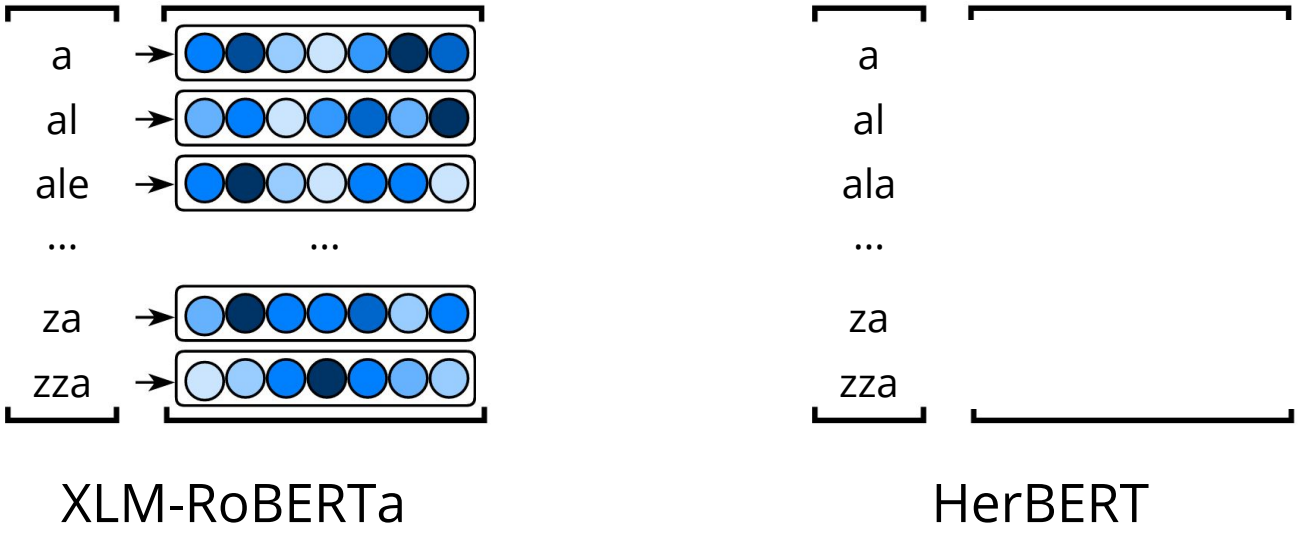
Corpus	SSO	BPE	Score
Small	1.0	Yes	84.69 ± 0.38
Large	1.0	Yes	84.38 ± 0.32
Small	0.1	Yes	84.94 ± 0.14
Large	0.1	Yes	85.24 ± 0.40
Small	0.0	Yes	85.38 ± 0.19
Large	0.0	Yes	**85.44 ± 0.32**
Small	0.0	No	85.18 ± 0.43
Large	0.0	No	85.41 ± 0.30

Inicjalizacja modelu wagami XLM-RoBERTa

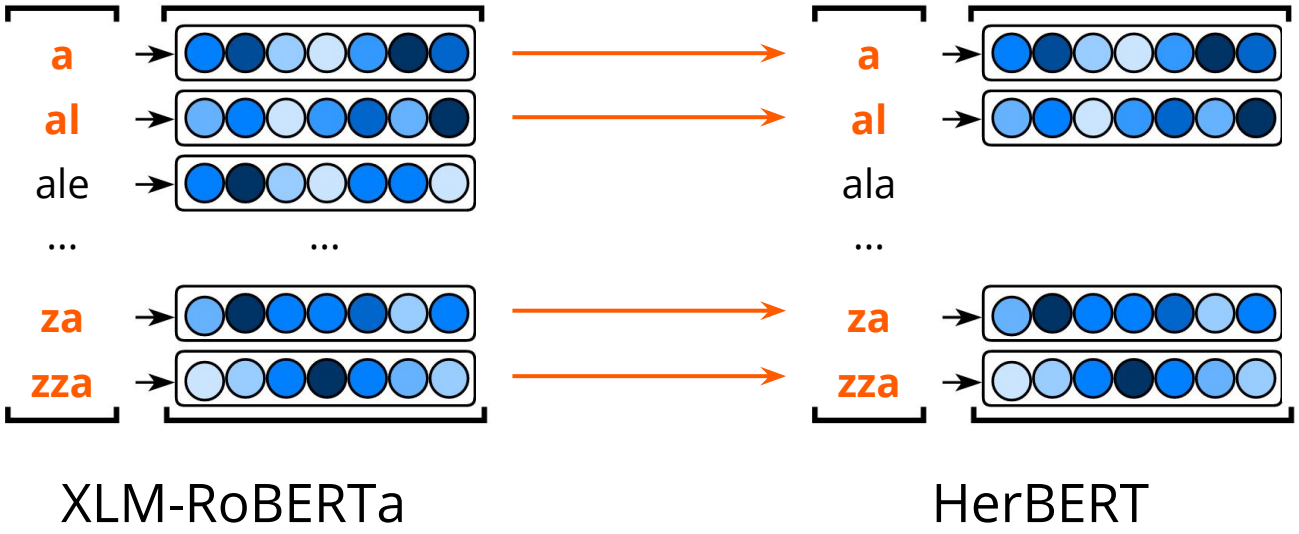


XLM-RoBERTa

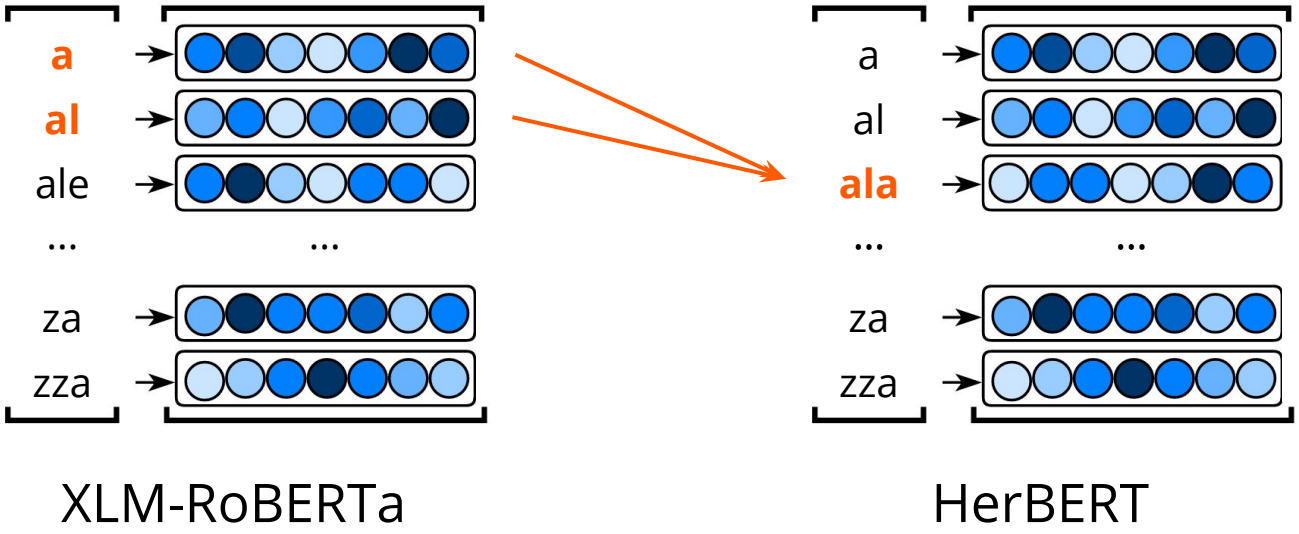
Inicjalizacja modelu wagami XLM-RoBERTa



Inicjalizacja modelu wagami XLM-RoBERTa



Inicjalizacja modelu wagami XLM-RoBERTa



HerBERT 2.0: Inicjalizacja

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

HerBERT 2.0: Inicjalizacja

Init	Pretraining	Score
Random	No	44.88 ± 0.20
XLM-R	No	$\underline{75.34 \pm 2.16}$
XLM-R	-	84.70 ± 0.29

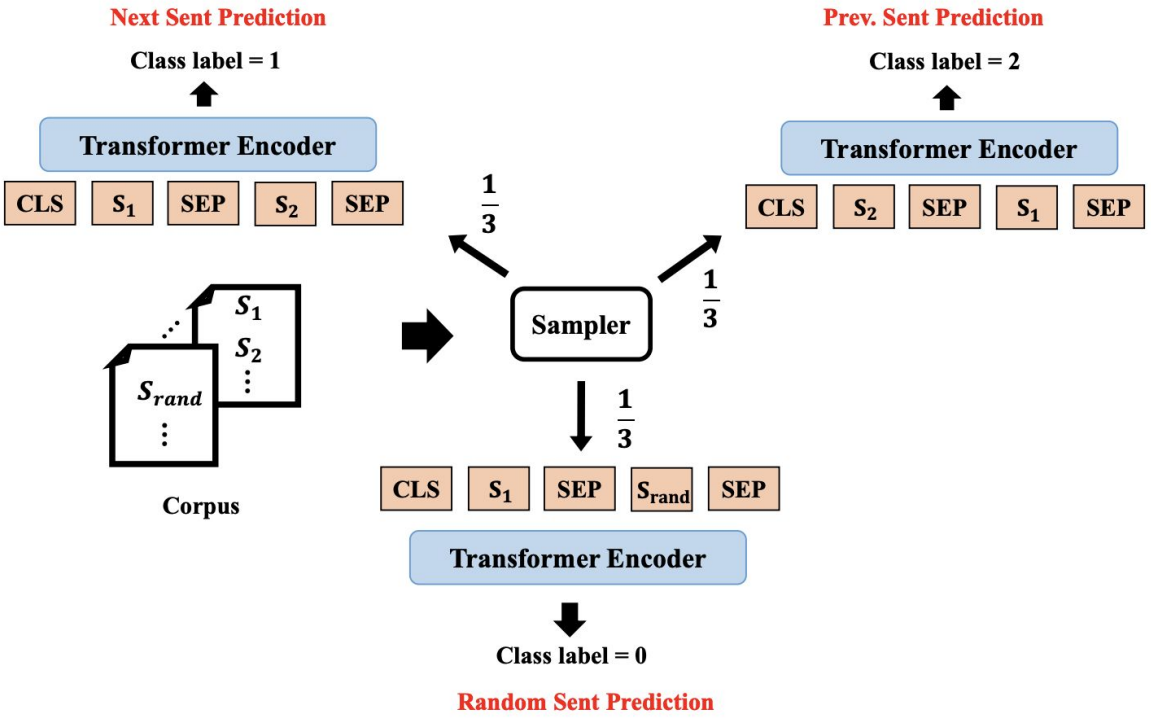
HerBERT 2.0: Inicjalizacja

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 2.0: Inicjalizacja

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 2.0: Sentence Structural Objective



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

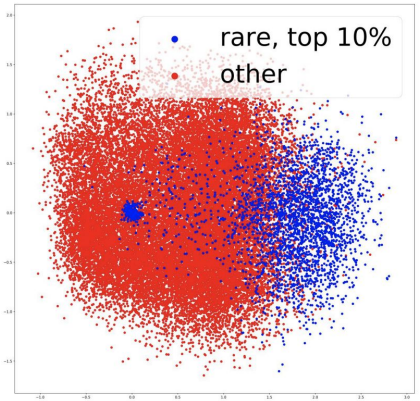
HerBERT 2.0: Sentence Structural Objective

SSO	Corpus	BPE	Score
0.0	Small	Yes	85.38 ± 0.19
0.1	Small	Yes	84.94 ± 0.14
1.0	Small	Yes	84.69 ± 0.38
0.0	Large	Yes	**85.44 ± 0.32**
0.1	Large	Yes	85.24 ± 0.40
1.0	Large	Yes	84.38 ± 0.32
0.0	Large	No	85.41 ± 0.30
0.1	Large	No	85.03 ± 0.27
1.0	Large	No	85.00 ± 0.28

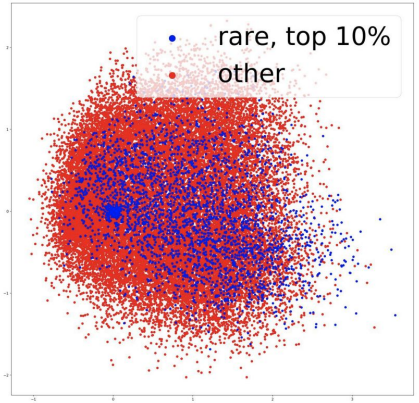
HerBERT 2.0: BPE-Dropout

```
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

HerBERT 2.0: BPE-Dropout

BPE	Init	SSO	Corpus	Score
No	Random	1.0	Large	82.38 ± 0.33
Yes	Random	1.0	Large	81.92 ± 0.26
No	XLM-R	1.0	Large	85.00 ± 0.28
Yes	XLM-R	1.0	Large	84.38 ± 0.32
No	XLM-R	0.0	Large	85.41 ± 0.30
Yes	XLM-R	0.0	Large	**85.44 ± 0.32**
No	XLM-R	0.0	Small	85.18 ± 0.43
Yes	XLM-R	0.0	Small	85.38 ± 0.19

HerBERT 2.0

Inicjalizacja modelu: wagami XLM-RoBERTa

Korpus: Large

SSO: 0.1

BPE-Dropout: Tak*

HerBERT 2.0: KLEJ

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Base Models										
XLM-RoBERTa	84.7 ± 0.29	91.7	93.3	93.4	66.4	90.9	77.1	64.3	97.6	87.3
Polish RoBERTa	85.6 ± 0.29	94.0	94.2	94.2	63.6	90.3	76.9	71.6	98.6	87.4
HerBERT	86.3 ± 0.36	94.5	94.5	94.0	67.4	90.9	80.4	68.1	98.9	87.7

HerBERT 2.0: KLEJ

Model	AVG	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	Czy wiesz?	PSC	AR
Base Models										
XLM-RoBERTa	84.7 ± 0.29	91.7	93.3	93.4	66.4	90.9	77.1	64.3	97.6	87.3
Polish RoBERTa	85.6 ± 0.29	94.0	94.2	94.2	63.6	90.3	76.9	71.6	98.6	87.4
HerBERT	86.3 ± 0.36	94.5	94.5	94.0	67.4	90.9	80.4	68.1	98.9	87.7
Large Models										
XLM-RoBERTa	86.8 ± 0.30	94.2	**94.7**	93.9	67.6	92.1	81.6	70.0	98.3	88.5
Polish RoBERTa	87.5 ± 0.29	94.9	93.4	94.7	69.3	**92.2**	81.4	74.1	**99.1**	88.6
HerBERT	**88.4 ± 0.19**	**96.4**	94.1	**94.9**	**72.0**	**92.2**	**81.8**	**75.8**	98.9	**89.1**

HerBERT 2.0: Tagowanie NKJP

Model	Accuracy	F1-Score
Base Models		
XLM-RoBERTa	95.97 ± 0.04	95.79 ± 0.05
Polish RoBERTa	96.13 ± 0.03	95.92 ± 0.03
HerBERT	**96.46 ± 0.04**	**96.27 ± 0.04**

HerBERT 2.0: Tagowanie NKJP

Model	Accuracy	F1-Score
Base Models		
XLM-RoBERTa	95.97 ± 0.04	95.79 ± 0.05
Polish RoBERTa	96.13 ± 0.03	95.92 ± 0.03
HerBERT	**96.46 ± 0.04**	**96.27 ± 0.04**
Large Models		
XLM-RoBERTa	97.07 ± 0.05	96.93 ± 0.05
Polish RoBERTa	97.21 ± 0.02	97.05 ± 0.03
HerBERT	**97.30 ± 0.02**	**97.17 ± 0.02**

HerBERT 2.0: Parsowanie zależnościowe PDB

Model	LAS	MLAS	BLEX
Static Embeddings			
No embeddings	87.46	73.40	79.98
FastText	89.78	79.56	83.56

HerBERT 2.0: Parsowanie zależnościowe PDB

Model	LAS	MLAS	BLEX
Static Embeddings			
No embeddings	87.46	73.40	79.98
FastText	89.78	79.56	83.56
Base Models			
XLM-RoBERTa	93.15	85.67	88.35
Polish RoBERTa	**93.91**	**87.63**	**89.20**
HerBERT	93.60	86.81	88.69

HerBERT 2.0: Parsowanie zależnościowe PDB

Model	LAS	MLAS	BLEX
Static Embeddings			
No embeddings	87.46	73.40	79.98
FastText	89.78	79.56	83.56
Base Models			
XLM-RoBERTa	93.15	85.67	88.35
Polish RoBERTa	**93.91**	**87.63**	**89.20**
HerBERT	93.60	86.81	88.69
Large Models			
XLM-RoBERTa	93.67	87.01	88.82
Polish RoBERTa	93.71	87.31	88.65
HerBERT	93.29	86.46	88.10

Wkrótce na:

https://huggingface.co/allegro

DZIĘKUJĘ!

piotr.rybak@allegro.pl

allegro