WHAT WAS THE PROBLEM?

- Supervised machine learning-based approaches focus on **features**:
  - surface features,
  - lexical resources,
  - knowledge-based features,
  - linguistic features, or
  - user-based and platform-based metadata.

This requires:
- a **well-defined feature extraction approach**
- **enough labelled data**

SOLUTION!

Unsupervised language pre-trained model

+

Transfer Learning

+

Fine-Tuning

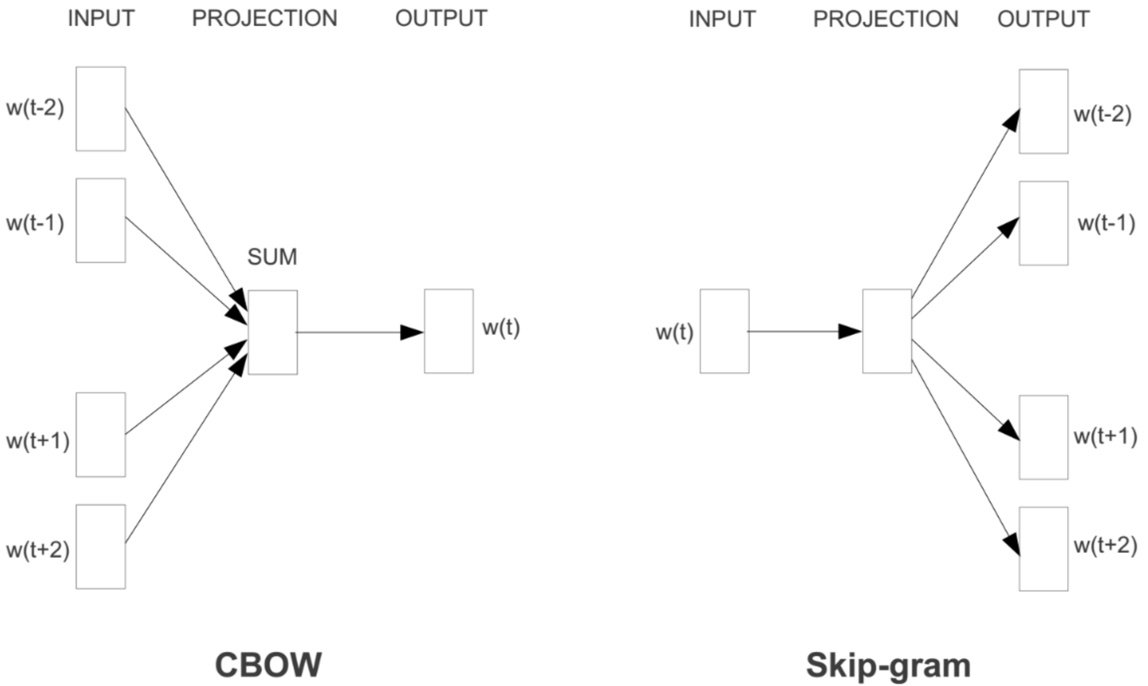
Word Embeddings - Distributional Hypothesis

"You shall know the meaning of a word by the company it keeps"

Firth (1957)

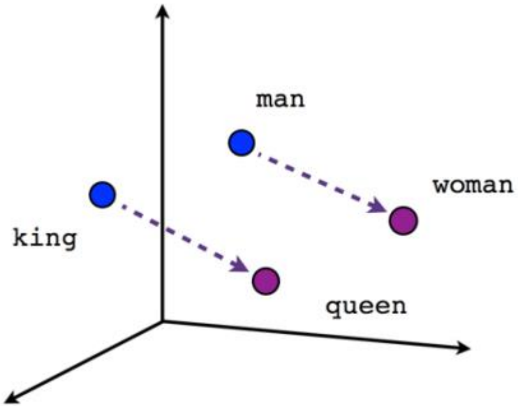
Associate a low-dimensional, dense vector w with each word w ∈ V so that similar words (in a distributional sense) share a similar vector representation.

Word Embeddings – Word2Vec

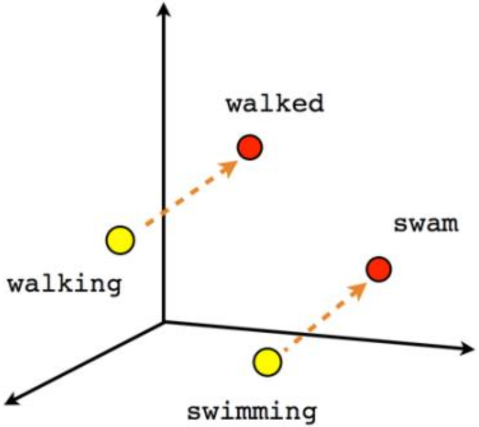


(Mikolov et al., 2013)

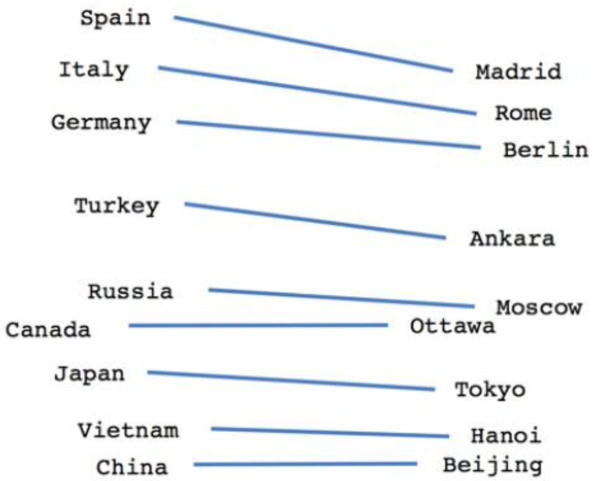
Word Embeddings – Properties



Male-Female



Verb tense



Country-Capital

https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space

Word Embeddings – Problem

I went to the bank

I went to the river bank

Word Embeddings – Problem

I went to the bank

I went to the river bank

[0.02, 0.04, … , 0.07]

Word Embeddings – Problem

The service was poor, but the food was ______________

yummy

delicious

poor

SOLUTION!

Train contextual representations on text corpus

I went to the bank

[0.01, 0.02, … , 0.07]

I went to the river bank

[0.06, 0.04, … , 0.09]

NLP's ImageNet moment

"If learning word vectors is like only learning edges, these approaches are like learning the full hierarchy of features, from edges to shapes to high-level semantic concepts."

Sebastian Ruder

Unsupervised language pre-trained model

- Unsupervised pre-training
- Supervised fine-tuning

BERT (Devlin et al., 2018)

Bidirectional

Encoder

Representations from

Transformers

Bidirectional

left-to-right

“NLP's ImageNet moment has [MASK]”

right-to-left

“[MASK] ImageNet moment has arrived ”

bidirectional

“NLP's ImageNet [MASK] has [MASK]”

Masking strategy

Unlabeled sentence: *my dog is hairy*

- 80% of the time: Replace the word with the [MASK] token, e.g., my dog is hairy ↠ my dog is [MASK]
- 10% of the time: Replace the word with a random word, e.g., my dog is hairy ↠ my dog is apple
- 10% of the time: Keep the word unchanged, e.g., my dog is hairy ↠ my dog is hairy.

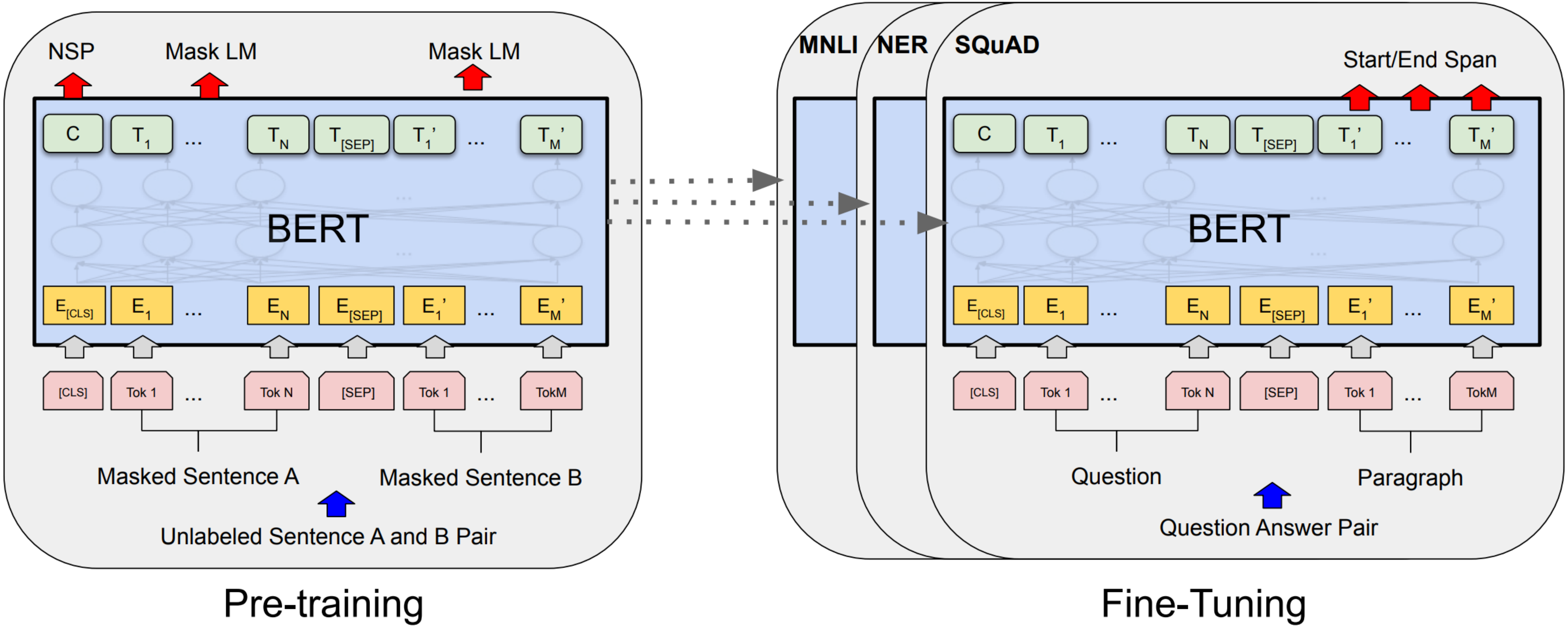
Next Sentence Prediction

To learn relationships between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence.

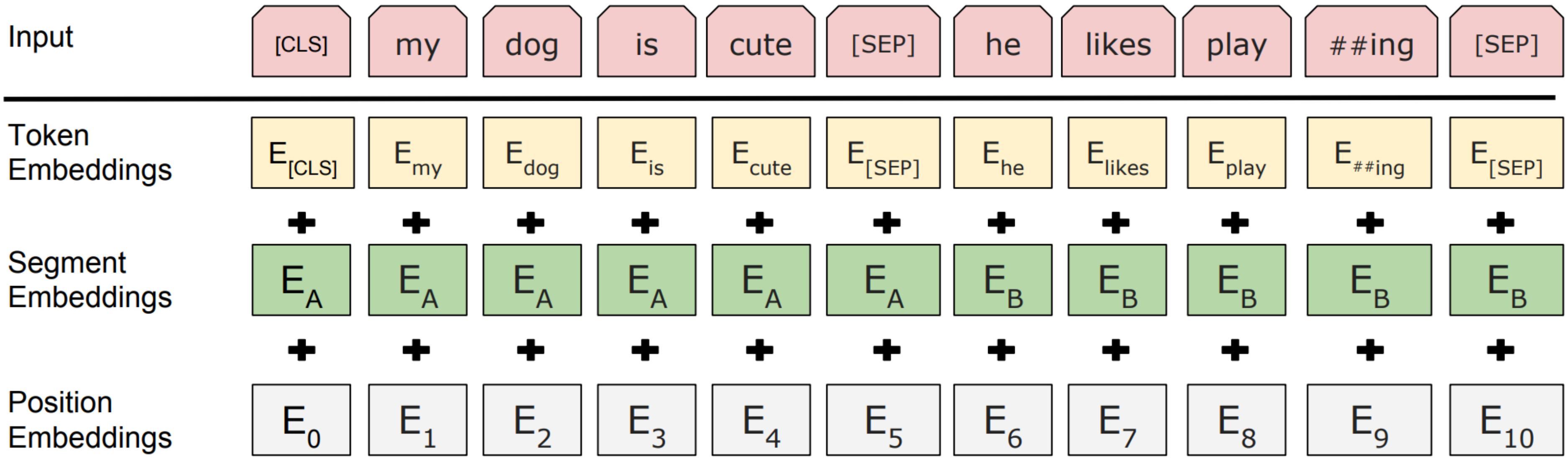
Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

BERT – Pre-training



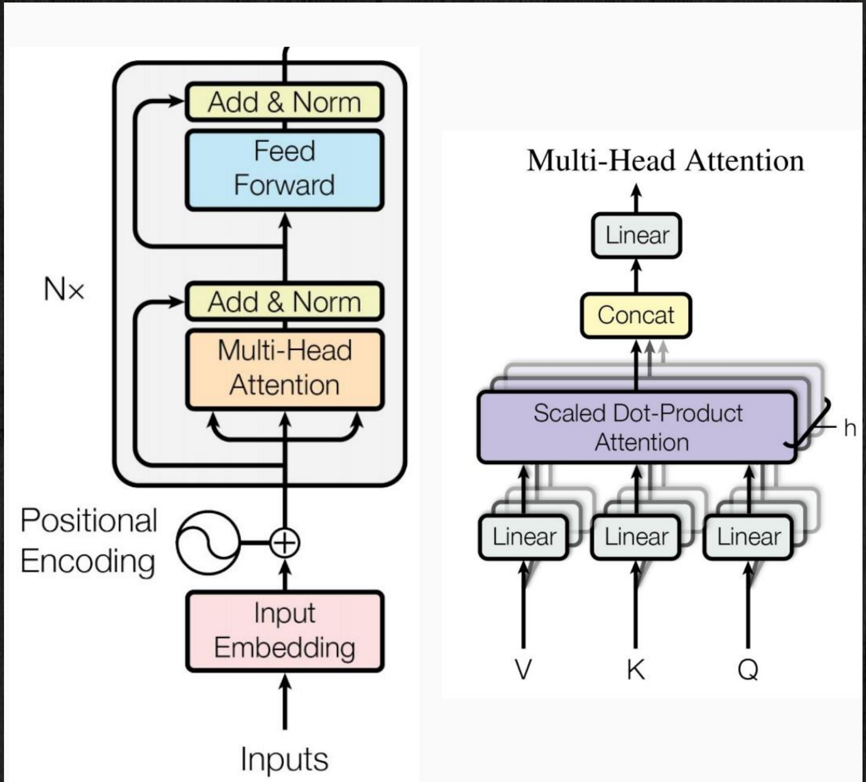
BERT input



Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

BERT Architecture

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning



https://nlp.stanford.edu/seminar/details/jdevlin.pdf

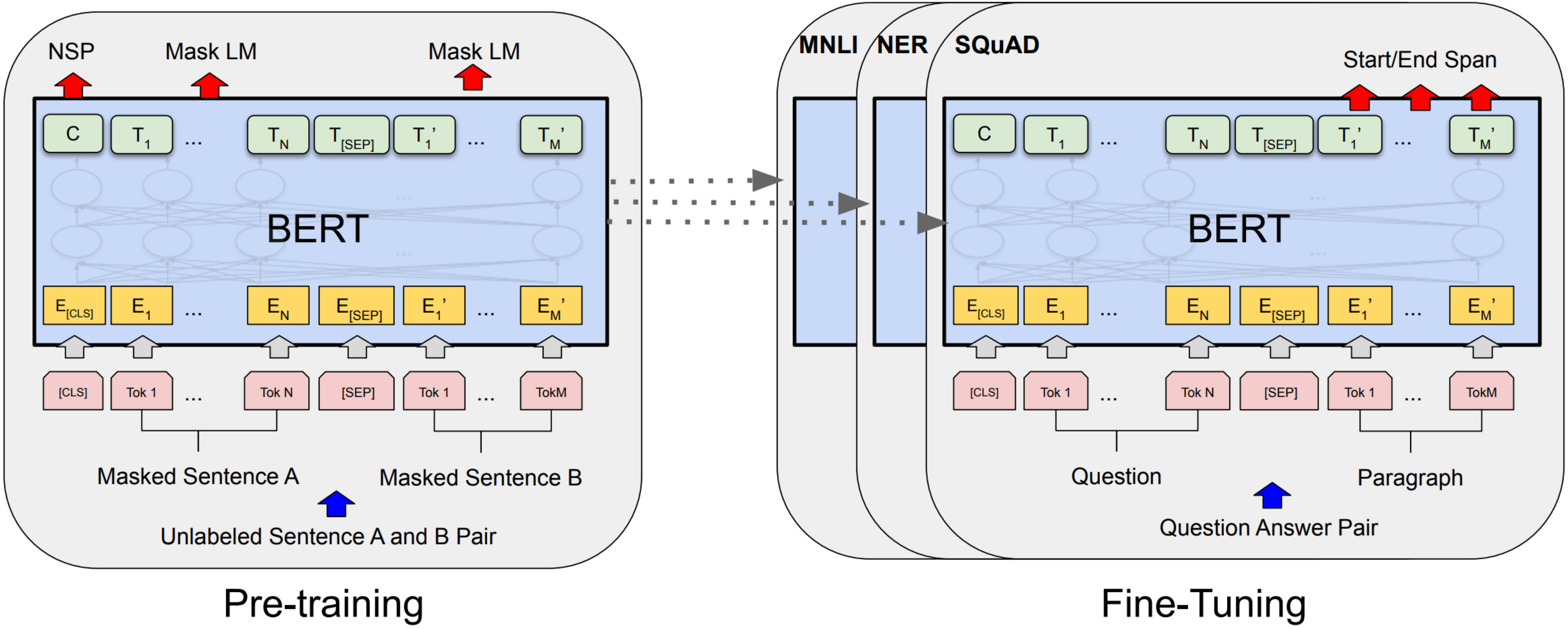
Base and Large

- BERT-Base: 12-layer, 768-hidden, 12-head
- BERT-Large: 24-layer, 1024-hidden, 16-head

Model Training data

- Wikipedia (2.5B words)
- BookCorpus (800M words)

BERT – Fine-tuning



GLUE Results

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Multilingual BERT

- Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary

System	English	Chinese	Spanish
XNLI Baseline - Translate Train	73.7	67.0	68.8
XNLI Baseline - Translate Test	73.7	68.4	70.7
BERT - Translate Train	81.9	76.6	77.8
BERT - Translate Test	81.9	70.1	74.9
BERT - Zero Shot	81.9	63.8	74.3

Multilingual BERT

- Trained single model on 104 languages from Wikipedia. Shared 110k WordPiece vocabulary

Fine-tuning \ Eval	EN	DE	NL	ES
EN	**90.70**	69.74	77.36	73.59
DE	73.83	**82.00**	76.25	70.03
NL	65.46	65.68	**89.86**	72.10
ES	65.38	59.40	64.39	**87.18**

Table 1: NER F1 results on the CoNLL data.

(Telmo at al., 2019)

Multilingual BERT - Problem

- Low-resource language are under-represented
- Wikipedia is not the only source

Language-specific BERT models

- These models are trained:
  - on different languages,
  - on different data sets,
  - use different architectural variants.

There are at least two French, two Dutch, and four Italian models. Which one is the best?

Bert Lang Street | Home | About

https://bertlang.unibocconi.it/

BERT LANG STREET

Lost in (language-specific) BERT models? We are here to help!

We currently have indexed **31** BERT-based models, **19** Languages and **28** Tasks.

We have a total of **178** entries in this table; we also show **Multilingual Bert (mBERT)** results if available! (see our paper)

Curious which BERT model is the best for named entity recognition in Italian? Just type *"Italian NER"* in the search bar!

Show 10 entries

Search:

Language	Model	NLP Task	Dataset	Dataset-Domain	Measure	Performance	mBERT	Difference with mBERT	Source
Arabic	Arabert v1	SA	AJGT	twitter	Accuracy	93.8	83.6	10.2	
Arabic	Arabert v1	SA	HARD	hotel reviews	Accuracy	96.1	95.7	0.4	
Arabic	Arabert v1	SA	ASTD	twitter	Accuracy	92.6	80.1	12.5	
Arabic	Arabert v1	SA	ArSenTD-Lev	twitter	Accuracy	59.4	51.0	8.4	
Arabic	Arabert v1	SA	LABR	book reviews	Accuracy	86.7	83.0	3.7	

Models comparison

Task	Metric	LS BERT	mBERT	Δ
SA	Acc	90.49	83.80	6.69
SA	F1	73.67	68.42	5.25
NER	F1	86.09	82.96	3.13
NLI	Acc	82.54	74.60	7.94
TC	Acc	87.93	85.22	2.72
TC	F1	70.65	54.49	16.16
POS	Acc	97.41	95.87	1.54
POS	F1	91.36	88.88	2.48
POS	UPOS	98.28	97.33	0.95
PI	Acc	88.44	87.74	0.69

(Nozza at al., 2020)