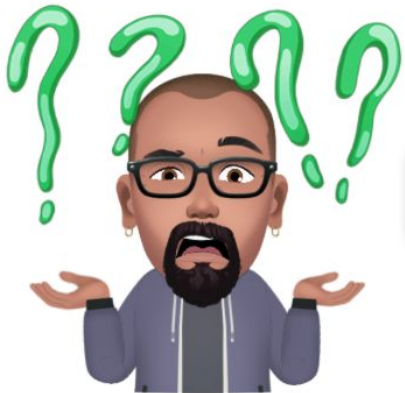
# GREEK BERT

## The Greeks visiting Sesame street

*John Koutsikakis,* ***Ilias Chalkidis,***
*Prodromos Malakasiotis and Ion Androutsopoulos*

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

# What is BERT
# and how does it work?

# Greek corpora



# Greek corpora



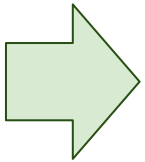

# Pre-training

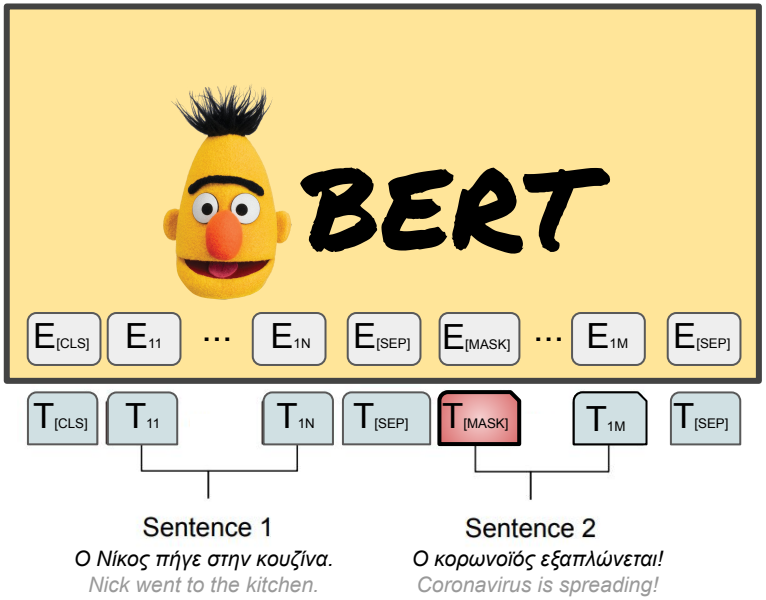
* similar to BERT-base (12 layers, 768 hidden units, 12 attention heads)

# Greek corpora

# Pre-training

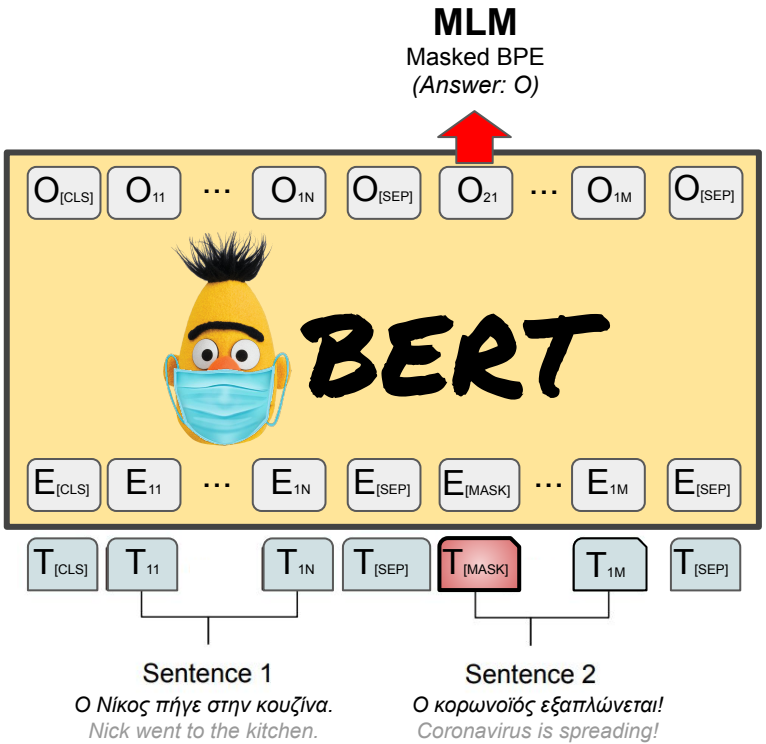




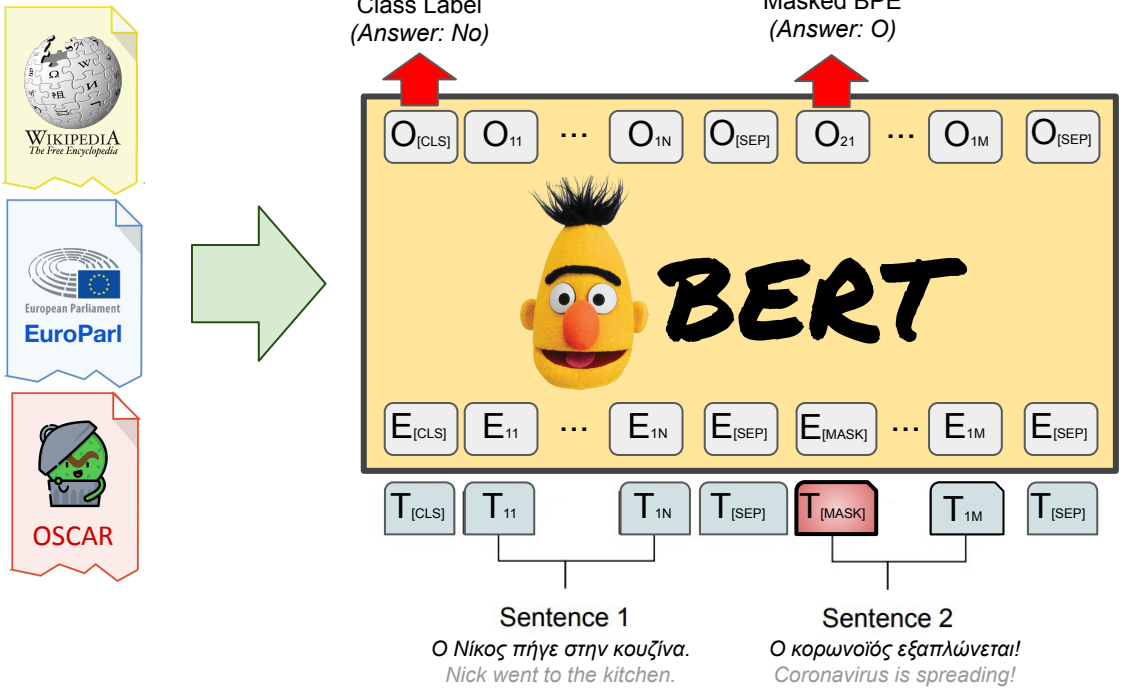
# Greek corpora


# Pre-training




# Greek corpora

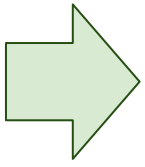
# Pre-training

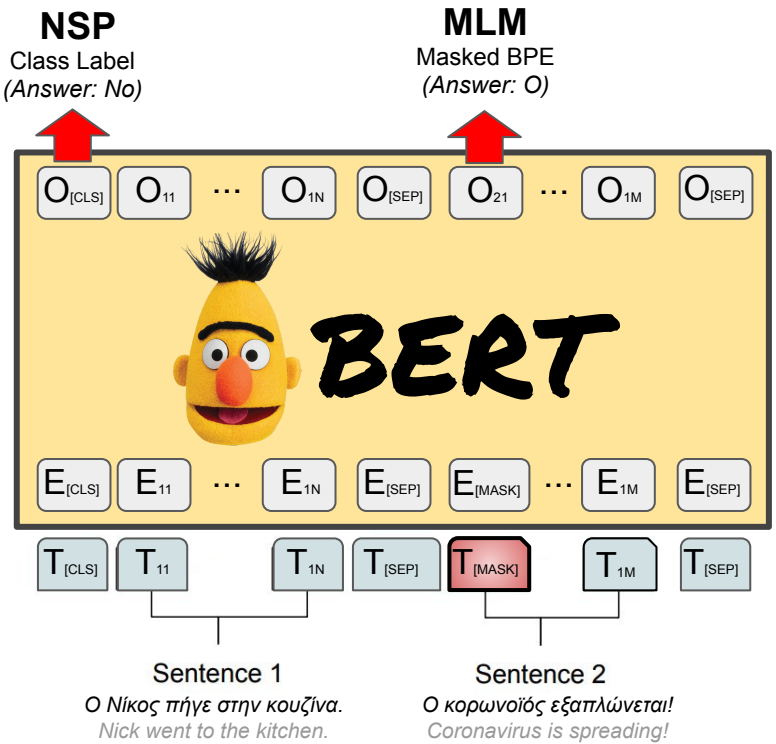


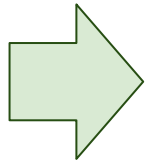
# Greek corpora

# Pre-training

# Fine-tuning



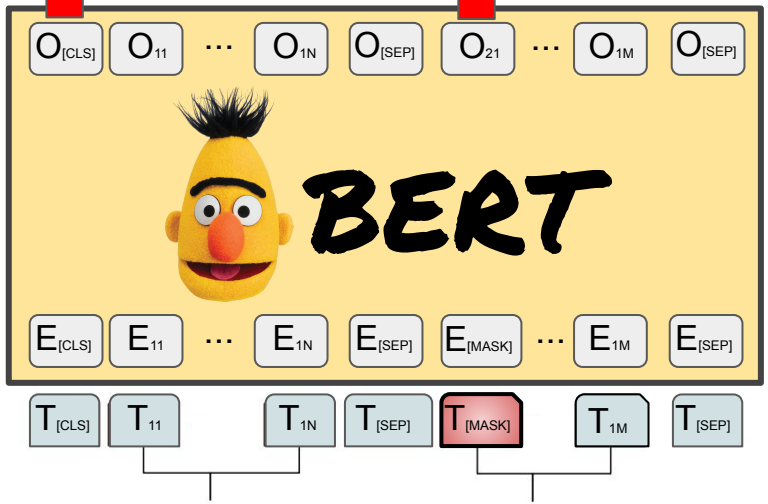
# Greek corpora

# Pre-training

# Fine-tuning


**NSP**
Class Label
*(Answer: No)*

**MLM**
Masked BPE
*(Answer: O)*




Sentence 1
*Ο Νίκος πήγε στην κουζίνα.*
*Nick went to the kitchen.*

Sentence 2
*Ο κορωνοϊός εξαπλώνεται!*
*Coronavirus is spreading!*

Downstream task
Gold Dataset

NER: Ο **[ Παναθηναϊκός - ORG ]** *κέρδισε 2-0.*
*[Panathinaikos - ORG] won 2-0.*
NLI: *Έφαγε ένα μήλο. Έφαγε ένα φρούτο.*
*He ate an apple. He ate a fruit.*
→ ***ENTAILMENT***

# Related Work - Multilingual models

| Model | Languages | Vocabulary Size | Data | Tasks | Greek coverage |
|---|---|---|---|---|---|
| **M-BERT** *Devlin et al. (2019)* | 100 | 100k | Wikipedias | MLM + NSP | ~1% |
| **XLM** *Lample and Conneau (2019)* | 15 | 100k | Wikipedias | MLM + TLM | ~1% |
| **XLM-R** *Conneau et al. (2019)* | 100 | 250k | Wikipedias + CommonCrawl | MLM | ~2% |

# Related Work - Monolingual models

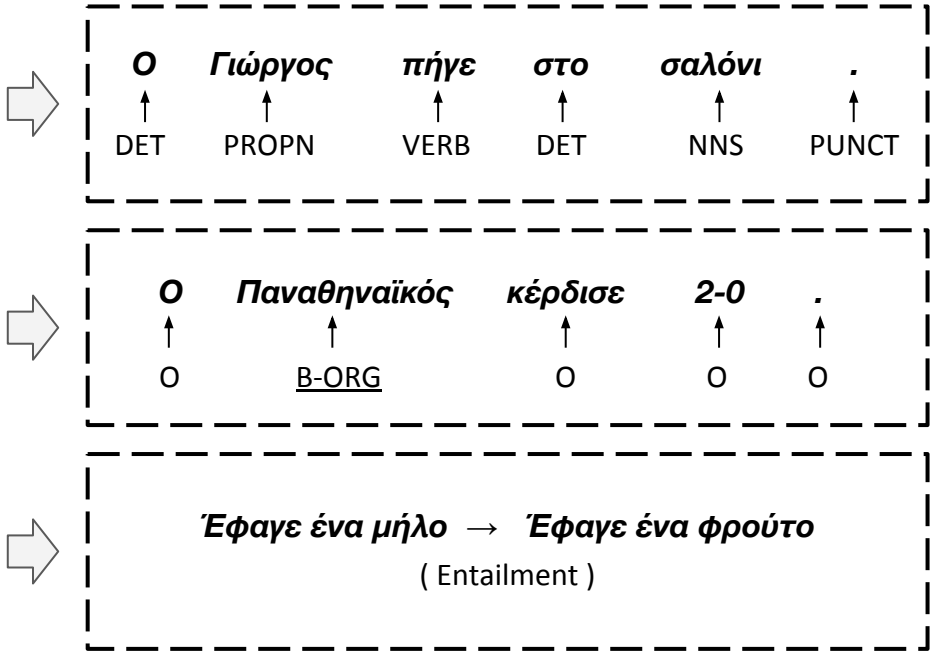
- **CAMEMBERT** - *Martin et al. (2019)* 🇫🇷
  - French, ROBERTA ⟶ MLM
  - SOTA in PoS tagging, dependency parsing, NER, and NLI
  - Comparison with M-BERT and XLM

- **FinBERT** - *Virtanen et al. (2019)* 🇫🇮
  - Finish, BERT ⟶ MLM + NSP
  - SOTA in PoS tagging, dependency parsing, NER, and text classification
  - Comparison with M-BERT

- **And many more…** 🇮🇹🇩🇪🇪🇸🇸🇦
  - e.g., for Italian, German, Spanish, Arabic, etc., still on development, no published work
  - Usually released on https://huggingface.co/models

# Benchmarks

- **Part of Speech (PoS) tagging**
  - Greek Universal Dependencies Treebank (GUDT) - Prokopidis et al. (2005, 2017)
  - 17 universal PoS tags (UPoS)
- **Named Entity Recognition (NER)**
  - Combination of 2 unpublished NER datasets - Ioannis Darras and Angeliki Romanou
  - 3 entity types (Person, Organization, Location)
- **Natural Language Inference (NLI)**
  - Cross-lingual Natural Language Inference (XNLI) - Conneau et al. (2018)
  - 3 categories (Entailment, Contradiction, Neutral)
  - Translated in 14 languages.



# Experimental Setup - Baselines

- **We compare GREEK-BERT against multilingual language models:**
  - M-BERT (cased and uncased) of Devlin et al. (2019)
  - XLM-R of Conneau et al. (2019)
- **We also compare with neural baselines operating on word embeddings:**
  - BILSTM-CNN-CRF of Ma and Hovy (2016) for PoS tagging and NER
  - Decomposable Attention Model (DAM) of Parikh et al. (2016) for NLI
  - Use Greek 300-dimensional FastText embeddings of Bojanovski et al. (2017)

# Experimental Setup - Baselines

- **Possible drawbacks of multilingual language models!**
  - Greek are underrepresented (1% in M-BERT, 2% in XLM-R)
  - Over-fragmentation of Greek words, i.e., *κατηγορούμενος → [κ, _ατ, _η, _γο, _ρου, _μενος]*
  - To estimate the level of fragmentation, we introduce Word Fragmentation Ratio (WFR)

| Model | GUDT (PoS) | NER | XNLI |
|---|---|---|---|
| M-BERT-UNCASED | 2.38 | 2.43 | 2.22 |
| M-BERT-CASED | 2.58 | 2.65 | 2.40 |
| XLM-R | 1.82 | 1.92 | 1.64 |
| GREEK-BERT | **1.35** | **1.33** | **1.23** |

$$\mathrm{WFR} = \frac{\sum_{n=1}^{N} \text{subword-units(n)}}{N}$$

  - *Hypothesis:* Over-fragmentation may harm the performance of multilingual models.

# Experimental Setup - Denoising XNLI

- **XNLI includes 340k machine-translated training pairs:**
  - Many of those have been poorly translated.
  - All 340K pairs *VS.* a subset of 40K high-quality pairs.
  - We estimate the quality using GREEK-BERT as a language model.

| Premise | | Hypothesis |
|---|---|---|
| νοιολογικά κρέμα κρέμα έχει δύο βασικές διαστάσεις - ϊόν και γεωγραφία.<br>*The conceptual cream cream has two basic dimensions - product and geography.* | → | Το προϊόν και η γεωγραφία είναι αυτά που κάνουν την κρέμα να κλέβει.<br>*The product and the geography are what make the cream steal.* |
| παράδειγμα, ορισμένες ηλικιακές ομάδες φαίνεται να είναι ευαίσθητες στην ατμοσφαιρική ρύπανση από άλλες.<br>*For example, some age groups appear to be more sensitive to air pollution than others.* | → | Η ατμοσφαιρική ρύπανση δεν μπορεί να επηρεάσει όλες τις ηλικιακές ομάδες.<br>*Air pollution may not affect all age groups.* |
| -Βουητό πιο εσωτερική.<br>*Buzz more internal.* | → | Συνηθισμένη<br>*Ordinary* |

- **The 2,500 development and 5,000 test pairs translated by professional translators.**

# Experimental Results - PoS tagging

| Model | Accuracy |
|---|---|
| BILSTM-CNN-CRF (Ma and Hovy, 2016) | 97.0 ± 0.14 |
| M-BERT-UNCASED (Devlin et al., 2019) | 97.8 ± 0.03 |
| M-BERT-CASED (Devlin et al., 2019) | 98.1 ± 0.08 |
| XLM-R (Conneau et al., 2019) | **98.2 ± 0.07** |
| GREEK-BERT (ours) | 98.1 ± 0.08 |

- BILSTM-CNN-CRF performs clearly worse, but the difference from the other models is small (0.8-1.2%).
- XLM-R is marginally (+0.1%) better than GREEK-BERT and M-BERTs.
- PoS tags in Greek can be determined by considering mostly the word's suffix → context is not always required

# Experimental Results - NER

| Model | Micro-F1 |
|---|---|
| BILSTM-CNN-CRF (Ma and Hovy, 2016) | 76.4 ± 2.07 |
| M-BERT-UNCASED (Devlin et al., 2019) | 81.5 ± 1.77 |
| M-BERT-CASED (Devlin et al., 2019) | 82.1 ± 1.35 |
| XLM-R (Conneau et al., 2019) | 84.8 ± 1.50 |
| GREEK-BERT (ours) | **85.7 ± 1.00** |

- GREEK-BERT outperforms the rest of the methods, quite comparable with XLM-R.
- NER is clearly more difficult than PoS tagging → easier to distinguish better methods.

# Experimental Results - NER

| Entity type | GREEK-BERT | XLM-R |
|---|---|---|
| PERSON | 88.8 ± 3.06 | 85.2 ± 1.25 |
| LOCATION | 88.4 ± 0.88 | 88.5 ± 0.86 |
| ORGANIZATION | 69.6 ± 4.28 | 68.9 ± 5.62 |

- Both are more accurate with persons and locations.
- GREEK-BERT is better on persons. Comparable on locations.
- Both struggle with organizations. Why?

Μπισκότα **Παπαδοπούλου** , Αθλέτικο **Μπιλμπάο**
PERSON LOCATION

# Experimental Results - NLI

*10% high quality*

| Model | Accuracy |
|---|---|
| DAM (Parikh et al., 2016) | 61.5 ± 2.07 |
| M-BERT-UNCASED (Devlin et al., 2019) | 65.7 ± 1.01 |
| M-BERT-CASED (Devlin et al., 2019) | 64.6 ± 1.29 |
| XLM-R (Conneau et al., 2019) | 70. 5 ± 0.69 |
| GREEK-BERT (ours) | **71.6 ± 0.80** |

- GREEK-BERT outperforms the rest of the methods.

# Experimental Results - NLI

| | *10% high quality* | *all train data* |
|---|---|---|
| Model | Accuracy | Accuracy |
| DAM (Parikh et al., 2016) | 61.5 ± 2.07 | 68.5 ± 1.71 |
| M-BERT-UNCASED (Devlin et al., 2019) | 65.7 ± 1.01 | 73.9 ± 0.64 |
| M-BERT-CASED (Devlin et al., 2019) | 64.6 ± 1.29 | 73.5 ± 0.49 |
| XLM-R (Conneau et al., 2019) | 70. 5 ± 0.69 | 77.3 ± 0.41 |
| GREEK-BERT (ours) | **71.6 ± 0.80** | **78.6 ± 0.62** |

- Performance improvement with all train data
  → noise acting as regularizer → improved generalization

# Experimental Results - NLI

| Class | GREEK-BERT | XLM-R |
|---|---|---|
| ENTAILMENT | 78.8 ± 1.20 | 78.0 ± 0.70 |
| CONTRADICTION | 81.2 ± 0.15 | 79.7 ± 0.53 |
| NEUTRAL | 75.9 ± 0.74 | 74.1 ± 0.50 |

- GREEK-BERT is better across classes.
- Both have difficulties on *neutral* pairs → often confused with *contradictions*.

# Conclusions

- Introduced GREEK-BERT, transformer-based language model pre-trained on Greek large corpora.

# Conclusions

- Introduced GREEK-BERT, transformer-based language model pre-trained on Greek large corpora.
- SotA in PoS tagging, Named Entity Recognition and Natural language inference.

# Conclusions

- Introduced GREEK-BERT, transformer-based language model pre-trained on Greek large corpora.
- SotA in PoS tagging, Named Entity Recognition and Natural language inference.
- Both GREEK-BERT and our code are publicly available ( https://github.com/nlpaueb/greek-bert , 🤗 )

```
import torch
from transformers import *

# Load model and tokenizer
tokenizer_greek = AutoTokenizer.from_pretrained('nlpaueb/bert-base-greek-uncased-v1')
lm_model_greek =
AutoModelWithLMHead.from_pretrained('nlpaueb/bert-base-greek-uncased-v1')

# ================ EXAMPLE ================
text_1 = 'Ο ποιητής έγραψε ένα [MASK] .'
# EN: 'The poet wrote a [MASK].'
input_ids = tokenizer_greek.encode(text_1)
print(tokenizer_greek.convert_ids_to_tokens(input_ids))
# ['[CLS]', 'ο', 'ποιητης', 'εγραψε', 'ενα', '[MASK]', '.', '[SEP]']
outputs = lm_model_greek(torch.tensor([input_ids]))[0]
print(tokenizer_greek.convert_ids_to_tokens(outputs[0, 5].max(0)[1].item()))
# the most plausible prediction for [MASK] is "song"
```

# Thank you!