CANINE vs mBERT: Why character-based tokenization?

Algorithms for Speech and NLP
MVA 2021-2022

David Soto
Elías Masquil
Nicolás Violante

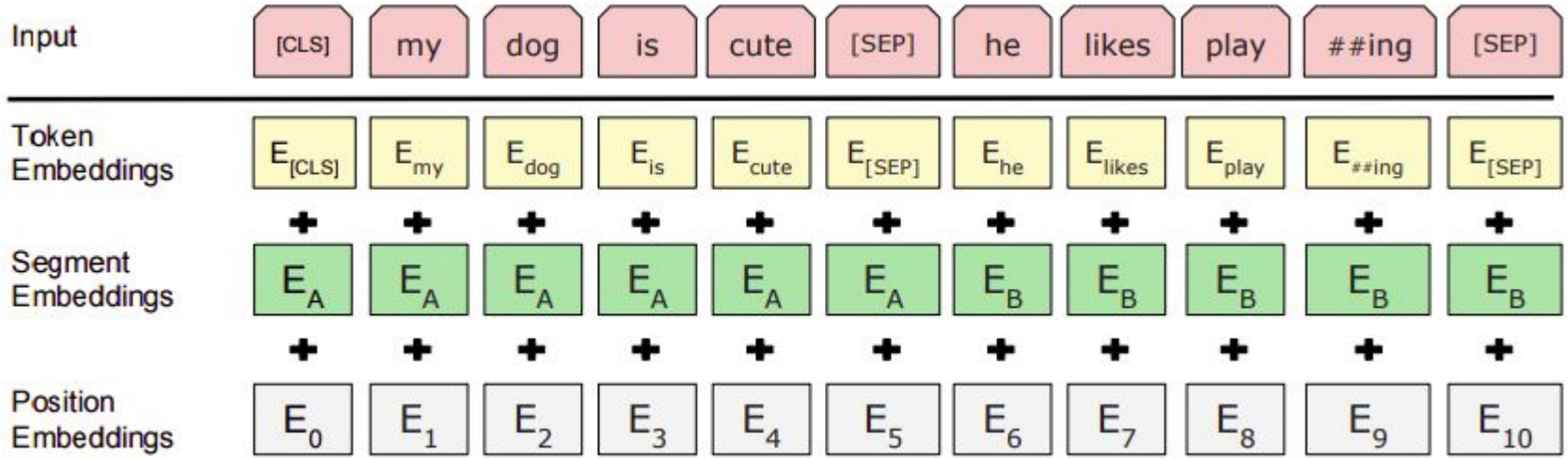
Agenda

1. Tokenization for LM
2. Datasets
3. Experiments
4. Discussion

Tokenization for LM

BERT tokenizer

- Subword approach: WordPiece:
  The vocabulary is initialized with individual characters in the language, then the most frequent combinations of symbols in the vocabulary are iteratively added to the vocabulary

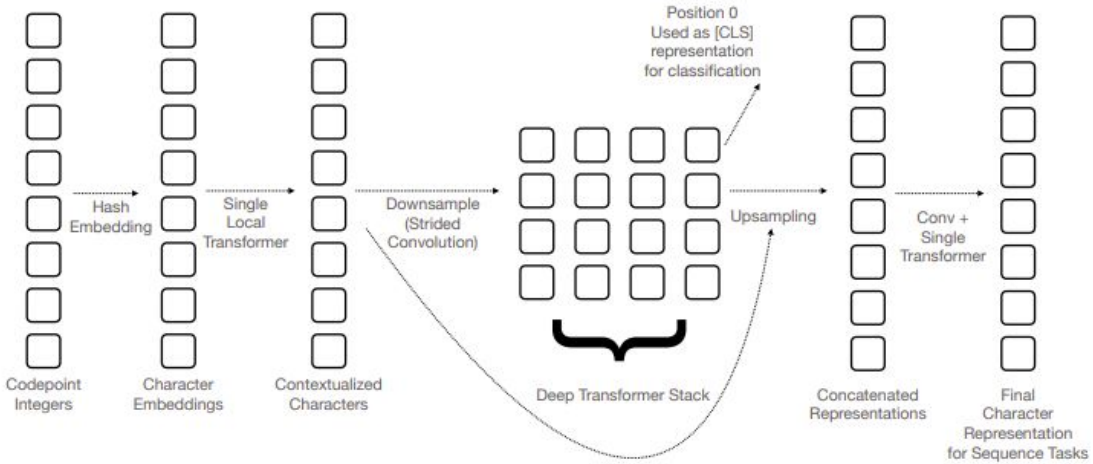


extracted from BERT

CANINE tokenizer

- Character-based tokenization

character -> unicode codepoint -> hash functions -> indexes -> linear combination of embeddings



extracted from CANINE

Why go character-based?

“heildarraforkuþörf” (total electric energy requirement)

- Constituent structure: [heildar, raf, orku, þörf] (total, electric, energy, need)
- Cased tokenization: ['he', '##ild', '##arra', '##for', '##ku', '##þ', '##ör', '##f']

“El bebé se bebe la leche” (The baby drinks the milk)

- Cased tokenization: ['El', 'be', '##bé', 'se', 'be', '##be', 'la', 'lec', '##he']
- Uncased tokenization: ['el', 'bebe', 'se', 'bebe', 'la', 'lech', '##e']

Datasets

Datasets

- **glue_mrpc**: Corpus of pairs of sentences in English, annotated to whether the sentences are semantically equivalent or not.
- **ajgt_twitter_ar**: Tweets in Arabic annotated as positive or negative.
- **spanish_diagnostics**: Dataset composed of medical diagnostics, labeled according to whether the data is a dental diagnostic or not.
- **amazon_reviews_multi**: Amazon product reviews of products written by users. They also include a ranking from 1 to 5. Multilingual dataset, we only considered the Spanish split.

Experiments

Experiments: CANINE vs mBERT

	English		Arabic		Spanish (diagnostics dataset)	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
canine-s	0.762	0.841	0.762	0.777	0.722	0.712
canine-c	0.809	0.860	**0.822**	**0.817**	**0.896**	**0.879**
mBERT-cased	**0.860**	**0.899**	**0.822**	0.810	0.854	0.831

Experiments: mBERT-cased vs mBERT-uncased

	spanish_diagnostics		amazon_reviews_multi	
	Accuracy	**F1 Score**	**Accuracy**	**F1 Score**
mBERT-cased	0.854	0.831	0.424	0.443
mBERT-uncased	**0.942**	**0.938**	**0.494**	**0.498**

Discussion

Discussion: availability of datasets

- Language-specific datasets for downstream tasks are harder to find for languages other than English.
- Less curated data
- Lack of documentation

Discussion: CANINE vs mBERT

- Using CANINE instead of mBERT did not yield a significant improvement in the downstream tasks, even in languages bad-suited for subword tokenization.

- Although mBERT might have worse encodings that destroy part of the linguistic structure of the sentence, given enough data, this information can still be recovered by exploiting statistical information around word contexts

Discussion: CANINE-S vs CANINE-C

- CANINE-C outperformed CANINE-S on all the experiments
- Pre-training with character loss seems to be better than pre-training with subword loss

Discussion: mBERT-cased vs mBERT-uncased

- Counterintuitively, mBERT-uncased outperformed mBERT-cased in Spanish
- This may be explained by the fact that we fine-tuned with a small dataset and a cased-model in this setup is harder to train that the uncased-model (that has a smaller vocabulary)
- The other explanation is that the case-sensible information was not necessary to solve the task.

Discussion: CANINE reported performance

- CANINE conference paper reports irregular results across some languages
- In particular, Arabic and Spanish results are worse than those of mBERT
- For more information please refer to this link

Thank you