Token-level Identification of Multiword Expressions using Pre-trained Multilingual Language Models

Raghuraman Swaminathan and Paul Cook
University of New Brunswick
{rswamina,paul.cook}@unb.ca

EST. 1785 UNB UNIVERSITY OF NEW BRUNSWICK

1. Multiword-expressions (MWEs)

- Expressions containing two or more words that exhibit idiomaticity
- Examples: *ivory tower, kick the bucket*
- Important to identify and classify MWEs for applications:
  - Machine translation, opinion mining

2. Tasks

- SemEval 2022 task 2 subtask A:
  - Binary sentence-level classification of MWEs as either idiomatic or literal
  - 3 languages: English , Portuguese and Galician
  - Input: He is a **night owl** and I am a morning person
  - Output: 0 - idiomatic
- PARSEME 1.2 edition:
  - Token-level identification of Verbal MWEs into 9 categories for 14 languages
  - Input: En 966, elle donne naissance à un fils, nommé Edmond
  - Output: donne naissance - VID

3. Experimental Setup

- 3 setups for SemEval:
  - en: train only on English
  - pt: train only on Portuguese
  - en+pt: train on both
- 3 setups for Parseme:
  - Mono: Model for each language
  - All: One model for all languages
  - Heldout: Train on all languages except test language

4. Model

- Multilingual BERT (mBERT), a transformer-based multilingual language model is pre-trained on 104 languages:
  - SemEval: mBERT is fine-tuned on the training data
  - Parseme: VMWEs are identified using mBERT with a dependency CRF network

5. Results

SemEval

Model	Train	Test- F1 score			
		en	pt	gl	ALL
mBERT	en	0.717	0.583	0.420	0.587
	pt	0.355	0.578	0.478	0.482
	en+pt	0.700	0.662	0.550	0.665
Baseline		0.345	0.391	0.434	0.389

Parseme

Setting	F1
Mono	0.699
All	0.722
Heldout	0.331
Baseline	0.002

6. Conclusions and Future Work

- Conclusions
  - Models learn information about MWEs and idiomaticity that is not language-specific
  - Data from other languages can be leveraged to improve model performance
- Future Work
  - Investigate influence of language families in cross-lingual MWE identification
  - Investigate ability of models to generalize to languages that were unseen during pre-training