Contextualized Embeddings Encode Monolingual and Cross-lingual Knowledge of Idiomaticity

Samin Fakharian and Paul Cook

University of New Brunswick

Fredericton, Canada

Introduction

- Multiword Expressions (MWEs)
    - MWEs are lexicalized combinations of multiple words, which display some form of idiomaticity
        - Fixed expressions: *by and large*
        - Light verb constructions: *take a walk*
        - Verb-noun combinations: *see stars*
    - Issues of MWEs
        - Learning the semantics of MWEs is a challenge due to their varying degrees of compositionality
    - MWEs’ importance in NLP
        - Commonly used in language and downstream applications like machine translation

Introduction

- Potentially idiomatic expressions (PIEs)
  - Ambiguous between non-compositional idiomatic interpretations and transparent literal interpretations
  - Used as idioms or as literal combinations
  - English Examples → *hit the road, skating on thin ice, off the hook*

Introduction

- VNCs are a common kind of MWE in English and cross-lingually
- Their meaning is often not predictable from the meanings of their component words
- Example → hit the road
  1. The marchers had hit the road before 0500 hours and by midday they were limping back having achieved success on day one.
     - Idiomatic → *hit the road* means ‘start a journey’.
  2. Two climbers dislodged another huge block which hit the road within 18 inches of one of the estate’s senior guides.
     - Literal

Introduction

- The idiomatic interpretations of English VNCs are typically lexico-syntactically fixed (canonical forms)
- Canonical forms are based on:
    - Voice of the verb
    - The determiner
    - Number of the noun
- Example → <u>hit the road</u> is in canonical form
- Usages that are not in their canonical form are often literal
    - E.g., *the road was hit, hit a road, hit the roads*

Introduction

- Research questions:
  1. Does an approach to identifying English and Russian PIEs that incorporates contextualized embeddings outperform prior approaches that do not use contextualized embeddings?
  2. Is an approach to identifying English and Russian PIEs that incorporates contextualized embeddings able to generalize to unseen expressions?
  3. Is an approach to identifying PIEs that incorporates contextualized embeddings able to generalize across languages?

Introduction

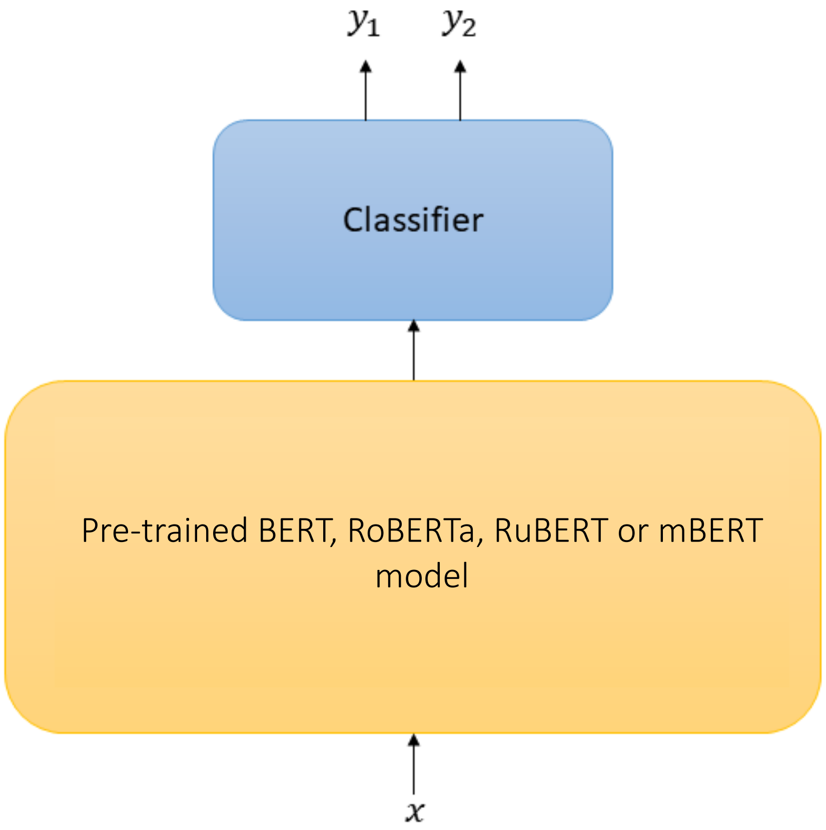
- Contributions:
  1. Propose an approach to identifying PIEs as idiomatic or literal that incorporates pre-trained contextualized embeddings and outperforms the previous state-of-the-art for this task
  2. Show that contextualized embeddings are able to capture the linguistic knowledge encoded in the canonical form feature in English VNCs
  3. Demonstrate that the proposed approach is able to generalize to unseen expressions
  4. Demonstrate that the proposed approach is able to generalize across languages

Proposed Model

- Our model:
  - A supervised approach
  - Based on contextualized embeddings
    - BERT, RoBERTa, RuBERT, mBERT
- Approaches to represent a PIE token instance
  - “CLS”
    - [CLS] token for the sentence in which it occurs for English experiments
    - [CLS] token for context of up to 300 characters to left and right of the target expression
    - 768-dimensional vector
- For English monolingual experiment we consider incorporating the canonical form feature (CF)

Proposed Model

- Fine-tuning pre-trained BERT, RoBERTa, RuBERT and mBERT models for binary classification of PIE token instances
  - Pre-trained model
    - 12 layers
    - Last layer of the pre-trained model (i.e., 12-th layer)
  - Classifier
    - Two fully-connected layers
    - Inputs
      - Representation of the VNC (with or without canonical form feature for English)
    - Labels
      - Idiomatic / Literal



Experimental Setup

- English → VNC-Tokens dataset
  - Contains 28 VNC types, and their instances are extracted from the British National Corpus
  - Manually labelled at the token level for whether they are literal or idiomatic usages
  - We used DEV and TEST parts of the dataset
  - Idiomatic and literal instances are roughly balanced across DEV and TEST
- Russian
  - A range of syntactic constructions including preposition+noun, preposition+adj+noun, and VNCs
  - Three sections containing classical prose, modern prose, and text from Russian Wikipedia
  - We consider only the Russian Wikipedia
  - Each instance is accompanied by a context window of up to three paragraphs
  - Idiomatic and literal instances are roughly balanced across RUSSIAN dataset

Experimental Setup

Set	# VNC Types	# Instances	% of Idiomatic Instances
EN - DEV	14	594	61%
EN - TEST	14	613	63%
Russian	37	775	54%

Experimental Setup

- “All Expressions” experimental setup
  - We randomly partition the instances of EN-DEV, EN-TEST, and RUSSIAN into training (roughly 75%) and testing (roughly 25%) sets, keeping the ratio of idiomatic to literal usages of each expression balanced across the training and testing sets
  - We repeat this random partitioning ten times
- Do we always have annotated instances of all PIE types?
  - No!

Experimental Setup

- “Unseen Expressions” experimental setup
  - Here we hold out all instances of one PIE type for testing and train on all instances of the remaining types (within either EN-DEV, EN-TEST and RUSSIAN)
  - We repeat this 14 times for each of EN-DEV and EN-TEST, holding out each VNC type once for testing and 37 times for RUSSIAN, holding out each PIE type once for testing
- Train and test models on EN-DEV → preliminary experiments and setting parameters
- Train and test models on EN-TEST → English final results
- Train and test models on RUSSIAN → Russian final results

Experimental Setup

- “Cross-lingual Expressions” experimental setup
  - Extension of the monolingual unseen expressions experimental setup
  - We evaluate on instances of PIEs in a language that was not observed during training
- Train models on EN-DEV or EN-TEST → Test models on RUSSIAN
- Train models on RUSSIAN → Test models on EN-DEV or EN-TEST

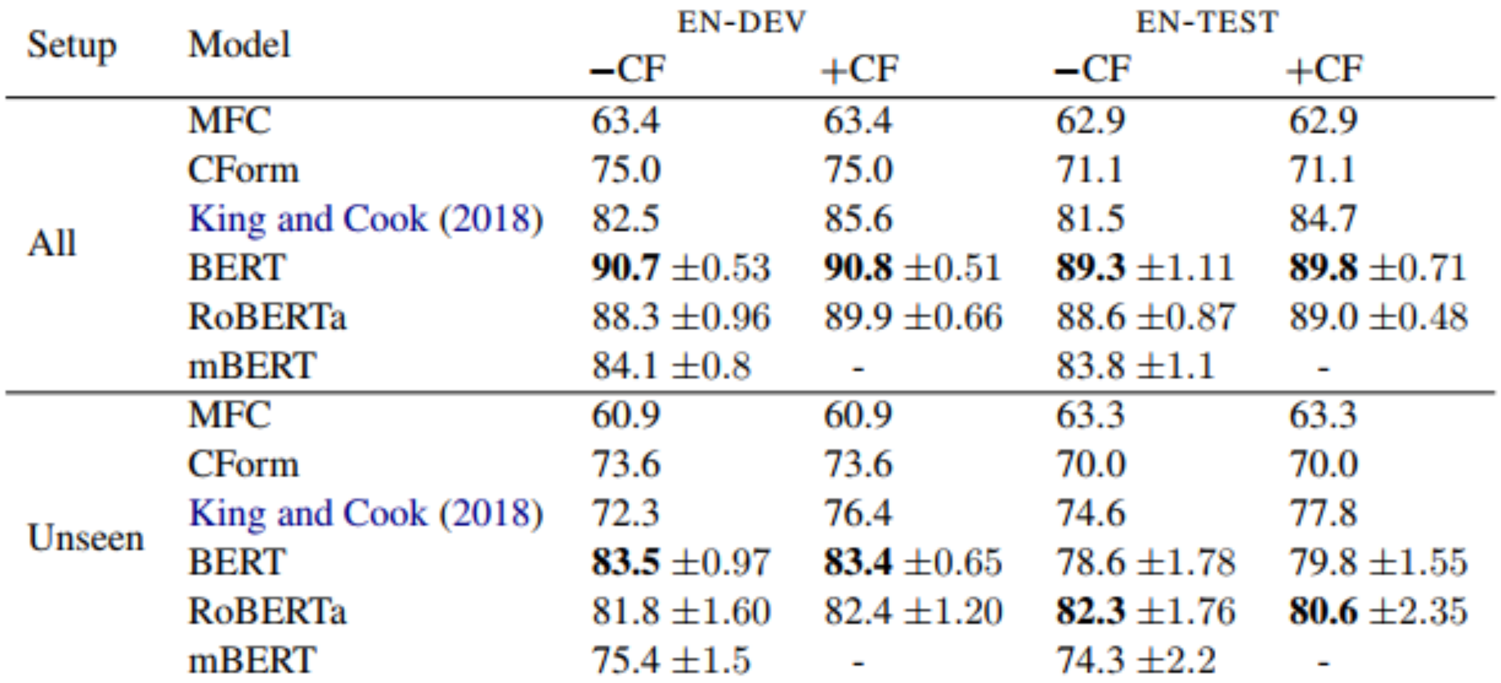
Experimental Setup

- Implementation and Parameter Settings
  - Huggingface implementations of BERT (bert-base-uncased), RoBERTa (roberta-base), RuBERT (rubert-base-cased) and mBERT (bert-base-multilingual-cased)
  - Adam optimizer to minimize cross-entropy loss
  - Default dropout
  - Batch sizes: 8, 16, 32
  - Epochs: 2, 3, 4
  - Learning rate: 2e-5, 3e-5, 5e-5
  - 10 runs with different random seeds

Evaluation

- Evaluation metric
    - Accuracy
- English baselines
    - Most frequent class baseline
    - Unsupervised approach by Fazly et al. (2009)
        - Unsupervised approach based on canonical form feature
    - Supervised approach by King and Cook (2018)
        - Supervised approach based on conventional word embeddings

Results

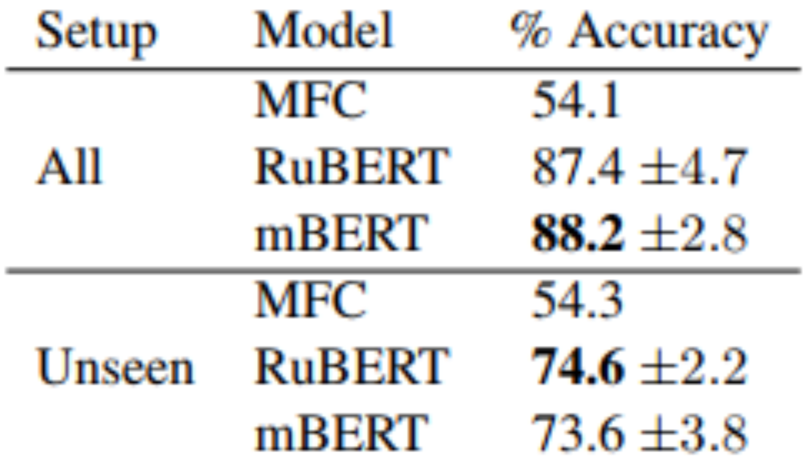


Setup	Model	EN-DEV		EN-TEST	
		−CF	+CF	−CF	+CF
All	MFC	63.4	63.4	62.9	62.9
	CForm	75.0	75.0	71.1	71.1
	King and Cook (2018)	82.5	85.6	81.5	84.7
	BERT	**90.7** ±0.53	**90.8** ±0.51	**89.3** ±1.11	**89.8** ±0.71
	RoBERTa	88.3 ±0.96	89.9 ±0.66	88.6 ±0.87	89.0 ±0.48
	mBERT	84.1 ±0.8	-	83.8 ±1.1	-
Unseen	MFC	60.9	60.9	63.3	63.3
	CForm	73.6	73.6	70.0	70.0
	King and Cook (2018)	72.3	76.4	74.6	77.8
	BERT	**83.5** ±0.97	**83.4** ±0.65	78.6 ±1.78	79.8 ±1.55
	RoBERTa	81.8 ±1.60	82.4 ±1.20	**82.3** ±1.76	**80.6** ±2.35
	mBERT	75.4 ±1.5	-	74.3 ±2.2	-

Results

- Findings (All expressions)
  - Contextualized embeddings can better capture knowledge of the idiomaticity of PIEs than previous approaches
  - Contextualized embeddings can better capture the linguistic knowledge encoded in the canonical form feature than conventional word embeddings
- Findings (Unseen Expressions)
  - The classifiers can capture information about the idiomaticity of PIEs
  - Information is not restricted to specific expressions, as in the case of the all expressions setup

Results



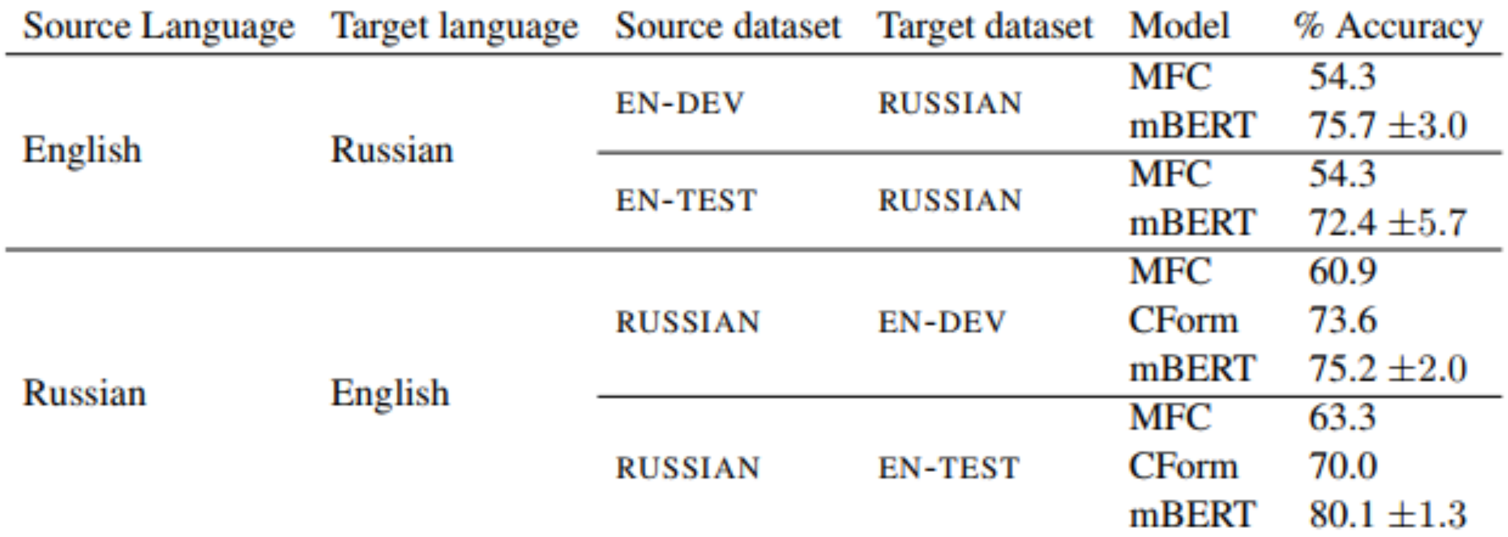
Setup	Model	% Accuracy
All	MFC	54.1
	RuBERT	87.4 ±4.7
	mBERT	**88.2** ±2.8
Unseen	MFC	54.3
	RuBERT	**74.6** ±2.2
	mBERT	73.6 ±3.8

Results

- Findings (All expressions)
  - Contextualized embeddings can capture knowledge of the idiomaticity of PIEs that are not specific to any syntactic constructions
- Findings (Unseen Expressions)
  - The classifiers can capture information about the idiomaticity of PIEs that is not restricted to expressions that were observed during training

Results

- We train on instances of PIEs in a source language, and evaluate on instances of PIEs in a target language



Source Language	Target language	Source dataset	Target dataset	Model	% Accuracy
English	Russian	EN-DEV	RUSSIAN	MFC	54.3
				mBERT	75.7 ±3.0
		EN-TEST	RUSSIAN	MFC	54.3
				mBERT	72.4 ±5.7
Russian	English	RUSSIAN	EN-DEV	MFC	60.9
				CForm	73.6
				mBERT	75.2 ±2.0
		RUSSIAN	EN-TEST	MFC	63.3
				CForm	70.0
				mBERT	80.1 ±1.3

Results

- Findings (Cross-lingual)
  - The classifiers can capture information about the idiomaticity of PIEs cross-lingually
  - Information is not restricted to specific expressions, nor to a specific language

Conclusion

- Contributions:
  1. Proposed an approach to identifying PIE idioms as idiomatic or literal that incorporates pre-trained contextualized embeddings and outperforms the previous state-of-the-art for this task
  2. Showed that contextualized embeddings are able to capture the linguistic knowledge encoded in the canonical form feature in English VNCs
  3. Demonstrated that the proposed approach is able to generalize to unseen expressions
  4. Showed that the proposed approach is able to generalize across languages

Conclusion

- Future work:
  - Further explore cross-lingual idiomaticity prediction
  - Include more languages in the analysis to be able to measure the impact of training on multiple source languages
  - Consider cross-lingual approaches for other MWE prediction tasks, such as predicting noun compound compositionality

Thank you!

Any Questions?