LREC 2022

Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms

Tosin Adewumi¹, Roshanak Vadoodi*, Aparajita Tripathy¹, Konstantina Nikolaidou¹, Foteini Liwicki¹ and Marcus Liwicki¹

¹Machine Learning Group, EISLAB, ^{1*}Luleå University of Technology, Sweden.

INTRODUCTION

- Idioms pose strong challenges to NLP systems, whether with regards to tasks such as MT, WSD, or IR.
- Idioms, or figures of speech, are common in everyday speech of human languages.
- Previous work have focused on datasets with only the literal and general idioms classification
- Objective: Create a relatively large English idioms corpus with multiple classes of idioms to train ML models in idiom identification & classification.

METHOD

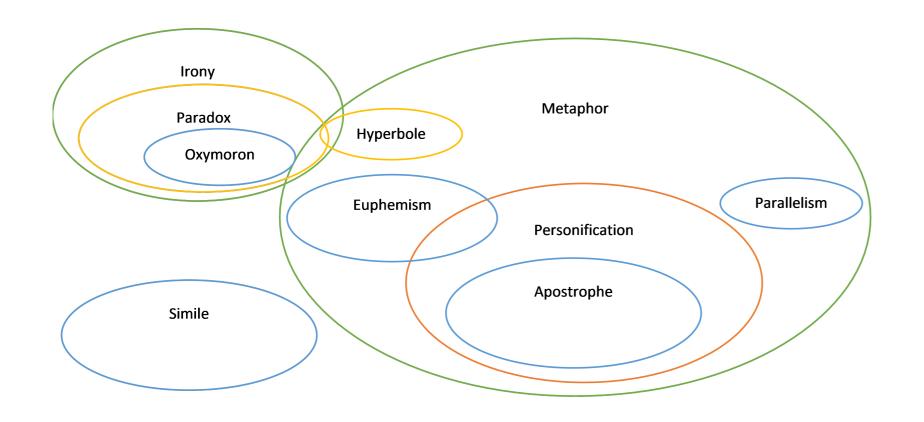
- We selected idioms from the dictionary by Easy Pace Learning in an alphabetical manner and samples were selected from the BNC & UKWaC based on the first to appear in both corpora.
- There are 4 contributors (second/L2 English speakers) and a nearnative speaker, who reviewed the corpus, for quality control (QC).
- Two independent annotators are involved in this work, with IAA of 88.89%.
- Classification
 experiments are
 performed with 3
 models, plus BERT.

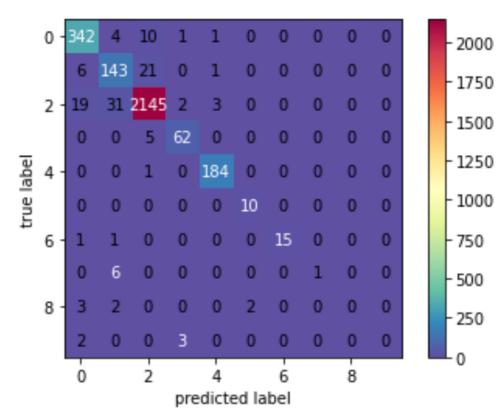
RESULTS

Classes	% of Samples	Samples	Acc	F1
Euphemism	11.82	2,384	0.935	0.93
Literal	5.65	1,140	0.813	0.78
Metaphor	72.7	14,666	0.975	0.98
Personification	2.22	448	0.811	0.81
Simile	6.11	1,232	0.996	0.98
Parallelism	0.32	64	0.667	0.62
Paradox	0.56	112	0.725	0.82
Hyperbole	0.24	48	0.048	0.08
Oxymoron	0.24	48	0.095	0.15
Irony	0.16	32	0	0
Total		20,174		
		Model	Acc	F1
		mNB	0.747	0.66
		SVM	0.766	0.67
		BERT	0.934	0.948

Distribution of samples, BERT average accuracy & F1 results over the 10 classes & weighted average accuracy & F1 of classification over the entire corpus for mNB, SVM & BERT models.

Depiction of classes of idioms & their relationships





BERT Confusion Matrix (0=Euphemism, 1=Literal, 2=Metaphor, 3=Personification, 4=Simile, 5=Parallelism, 6=Paradox, 7=Hyperbole, 8=Oxymoron, 9=Irony)

SCAN ME



CONCLUSION

- We created, possibly the first, labelled idioms corpus with 10 classes, based on the BNC and the UKWaC corpora.
- The dataset may be extended to meet specific NLP needs by researchers.
- We make the dataset and the relevant codes for working with it publicly available.