Exploring Euphemism Detection in Few-Shot and Zero-Shot Settings

Sedrick Scott Keh
skeh@cs.cmu.edu

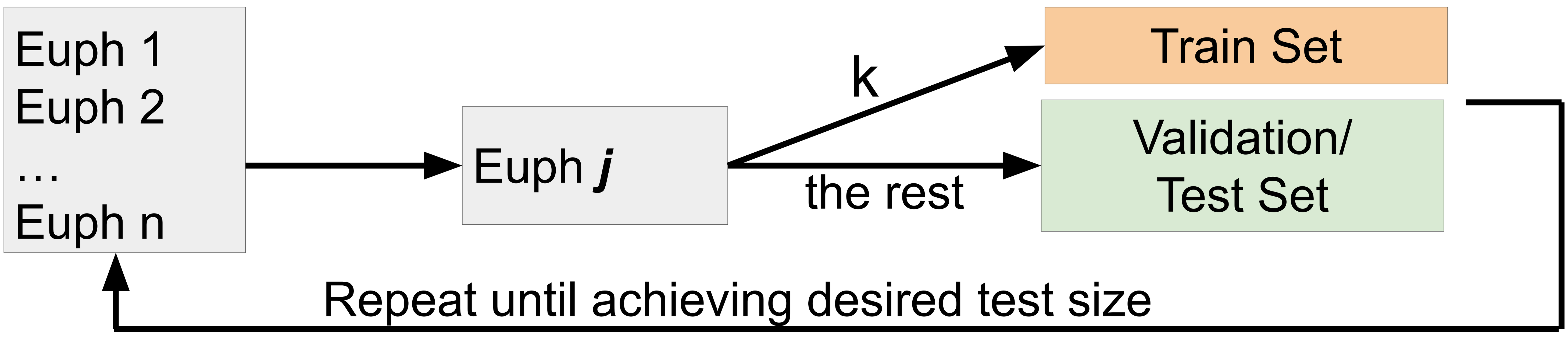
Carnegie Mellon University

I. Motivation

- Euphemism Detection Shared Task – detect euphemisms given a training and validation set
- **How do we ensure that the models are actually "learning" the euphemism-related concepts rather than simply memorizing the euphemistic terms?**
- Solution: Evaluate performance on euphemisms that were unseen during training time

II. Dataset Construction + Methodology

1. **Few(k)-Shot** – Randomly select a euphemistic phrase. Assign *k* of them to the train set, and the rest to the validation/test set. Repeat until desired test set size is achieved.



2. **Zero-Shot** – Similar to above, with k=0

3. **Zero-Shot (Categorical)** – Use categories defined in the dataset by Gavidia et al (2022). Select one category for validation and testing, and keep the rest for training.

	Ave. Test Size	Ave. # of unique PETs in test
Standard	295.0	93.3
Few-Shot (k=1)	279.6	35.0
Few-Shot (k=3)	281.2	35.4
0-shot (random)	280.6	34.3
Death	174.0	14.9
Sexual Activity	45.0	10.4
Employment	176.0	23.5
Politics	161.0	20.9
Bodily Functions	26.0	7.0
Physical/Mental	299.0	36.0
Substances	88.0	9.1

III. Experiments + Results

1. **RoBERTa** – Try both base and large. Fine-tune + predict.
2. **GPT-3 (davinci)** – Prompt with "*Is the word [PET] used euphemistically in the following sentence: [SENT]*", where [PET] is the euphemism and [SENT] is the sentence.

		RoBERTa-base			RoBERTa-large			GPT-3 (davinci)		
		P	R	F1	P	R	F1	P	R	F1
Standard Model	-	0.850	0.799	0.824	0.877	0.812	0.836	-	-	-
Few-Shot	k=1	0.802	0.744	0.759	0.818	0.748	0.769	0.565	0.551	0.546
	k=3	0.834	0.795	0.808	0.879	0.798	0.825	0.624	0.599	0.617
Zero-Shot (Random)	-	0.770	0.699	0.715	0.798	0.726	0.740	0.537	0.543	0.507
Zero-Shot (Type-based)	Death	0.782	0.735	0.742	0.803	0.748	0.761	0.453	0.457	0.448
	Sexual Activity	0.647	0.606	0.622	0.633	0.603	0.615	0.533	0.550	0.477
	Employment	0.778	0.790	0.781	0.782	0.817	0.792	0.537	0.532	0.479
	Politics	0.754	0.622	0.645	0.826	0.645	0.688	0.537	0.558	0.484
	Bodily Functions	0.500	0.240	0.324	0.500	0.416	0.480	0.500	0.192	0.278
	Physical/Mental	0.757	0.663	0.689	0.750	0.680	0.693	0.517	0.510	0.489
	Substances	0.897	0.858	0.878	0.913	0.883	0.895	0.553	0.551	0.486

IV. Discussion

- The overall **results are generally quite good** (i.e. few-shot and zero-shot performance is not far behind standard setting)
- Some categories of euphemisms (e.g. substances) performed quite well, while others (e.g. bodily functions) performed poorly
- GPT3 in general performed quite poorly
- GPT3 benefited from few-shot training particularly significantly

arXiv preprint: https://arxiv.org/pdf/2210.12926.pdf