Supervised and Unsupervised Approaches to Word Usage Similarity

Milton King and Paul Cook
University of New Brunswick

Word Usage Similarity

- Measure the similarity of a given word in two different contexts.

Similarity	
3.3	A chain is only as strong as its weakest link. Norway must send a strong signal.
1.3	A country needs a strong leader. It has a strong fruity flavour.

- Determine meaning of words without the assistance of a dictionary

Word Embeddings and Model

- **Word embeddings:** Vector representations of words. Trained using the neural network based model, Skip-gram, which is implemented in Word2Vec.
- **Sentence Embedding:** Sum vectors for words in a sentence
- **Unsupervised Model:** Similarity = cosine between sentence embeddings.
- **Supervised Model:** Represent a sentence pair as the componentwise absolute difference and product of their embeddings.
  - Train ridge regression
  - 10-fold cross validation

Datasets and Evaluation

- **Datasets:** Pairs of sentences with the same target word and a similarity score given by human annotators.

Dataset	Lemmas	Pairs of Sentences	Part-of-speech
ORIGINAL	34	1512	N,V,Adj,Adv
TWITTER	10	550	N

- **TWITTER:** Tweets
- **ORIGINAL:** Text from web pages
- **Evaluation:** Spearman's ρ between the model's predictions and the human judgments.

Results

Method	ρ ORIGINAL	ρ TWITTER
Previous Best	0.202	0.290
Unsupervised	0.285	0.364
Supervised	0.440	0.442

Conclusion and Future Work

- **Conclusion:**
  - State of the art results with our unsupervised model.
  - Further improvements with a supervised approach.
- **Future Work:** Develop improved methods to represent the meanings of words in context.