An Automatic Approach to Discover Lexical Semantic Differences in Varieties of English

Priyal Nagra and Paul Cook

Faculty of Computer Science, University of New Brunswick

Lexical Semantic Variation

- Speakers of different language varieties use certain words differently– same words with different meanings.
- For example: *Boot* in American English is a type of footwear, whereas in British English it's a covered space at the back of a car, for storing things in**.**
- Research Question: Does a word embedding model outperform a traditional distributional semantic model and a method of keywordness in detecting lexical semantic differences in varieties of English?

Models

- **Keywordness**: Measure how much more frequent a word is in one corpus vs another corpus based on **frequency ratio**, **log-likelihood ratio** and **chi-square**.
- **Distributional Semantic model**: Represent a word as a vector based on frequency of co-occurrence with other words.
- **Word2Vec**: Neural network inspired model for representing a word as a vector.
  - *Word2Vec1*: Train model on each corpus separately, and then learn a transformation.
  - *Word2Vec2*: Train model on one corpus and continue training on another.
- **Cosine:** Measure how similar a word is in two corpora for DSM and Word2Vec approaches.

Evaluation Methodology

- **Corpora**: Built 3 pairs of English corpora based on a web crawl.

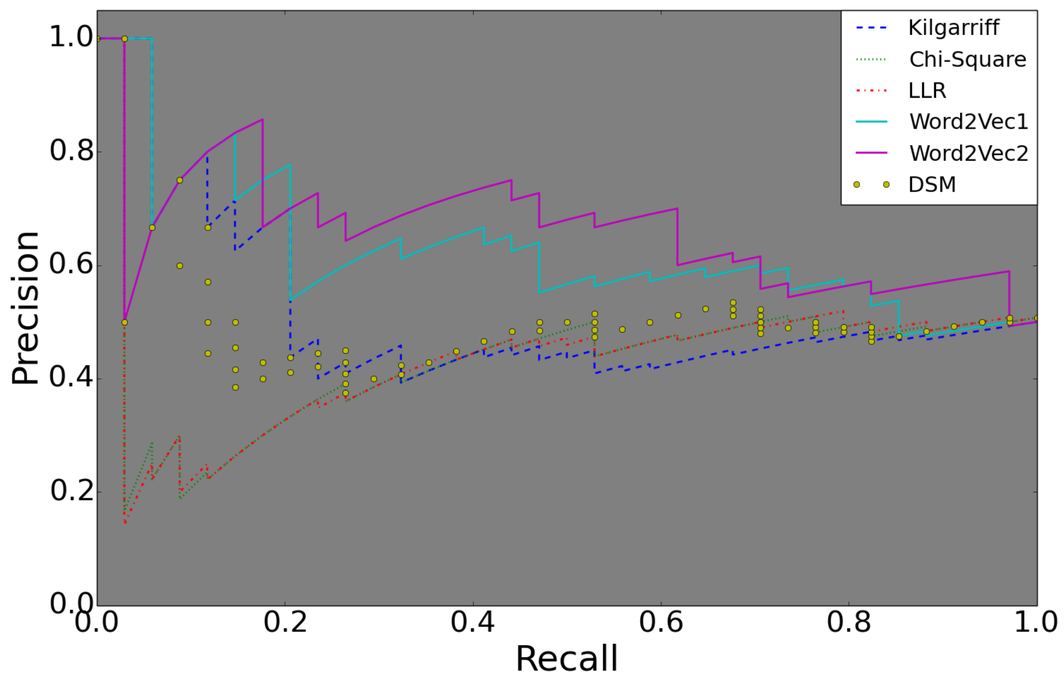
Corpora	# of Words
AU NOT- AU	57M 1B
CA NOT-CA	218M 1B
US NOT-US	1B 1B

- **Regionalisms:** 100 Americanisms, 100 Australianisms and 34 Canadianisms from dictionaries.
- **Distractors:** Same number of non-regionalisms**.**
- **Evaluation:** Measure extent to which regionalisms are more marked than distractors**.**

Precision-Recall Area Under Curve

Methods	Australian	Canadian	American
Kilgarriff	0.58	0.51	0.52
Chi-Square	**0.61**	0.44	0.47
LLR	0.60	0.44	0.48
DSM	0.49	0.50	0.48
Word2Vec1	0.51	0.62	0.46
Word2Vec2	0.43	**0.66**	**0.55**

Precision-Recall Curves



A PR curve for detecting lexical variation in Canadian English

Conclusion and Future Work

- Word2vec2 outperforms a DSM and measures of keywordness for identifying regionalisms in 2 out of 3 cases.
- **Future work**: Conduct a larger-scale evaluation with more known regionalisms.