Measuring Semantic Relatedness Across Languages

Alistair Kennedy Graeme Hirst

University of Toronto, Department of Computer Science, Toronto, Ontario, Canada

Abstract

Measures of Semantic Relatedness determine the degree of relatedness between two words. Most of these measures work only between pairs of words in a single language. We propose a novel method of measuring semantic relatedness between pairs of words in two different languages. This method does not use a parallel corpus but is rather seeded with a set of known translations. For evaluation we construct a cross-language dataset of French-English word pairs with similarity scores. Our new cross-language measure correlates more closely with averaged human scores than our unilingual baselines.

1. Distributional Semantics

"You shall know a word by the company it keeps" – Firth (1957)

- Construct a word-context matrix
  - Corpora: French and English Wikipedias
  - Used POS-tagged words as contexts
  - Re-weight matrix – Pointwise Mutual Information (PMI)
- Cosine similarity
- Evaluate correlation on Rubenstein and Goodenough (1965) style dataset

2. Translating a Word-Context Matrix

- Translate French matrix to English
- Build Translation Matrix
  - For each English-French context pair $\langle c_e, c_f \rangle$
    - Find all words $w_e \in c_e$ and $w_f \in c_f$
    - Find translations of w_e and w_f from aligned Wordnet Libre du Francais (WOLF) v0.1.5 (Sagot and Fišer, 2008) and Princeton WordNet v2.0 (Fellbaum, 1998)
    - Calculate PMI between c_e and c_f using translations
- Map contexts from French to English
  - Minimum PMI threshold, $\tau = 1.0, 2.0, ..., 5.0$
- Merge French and English matrices

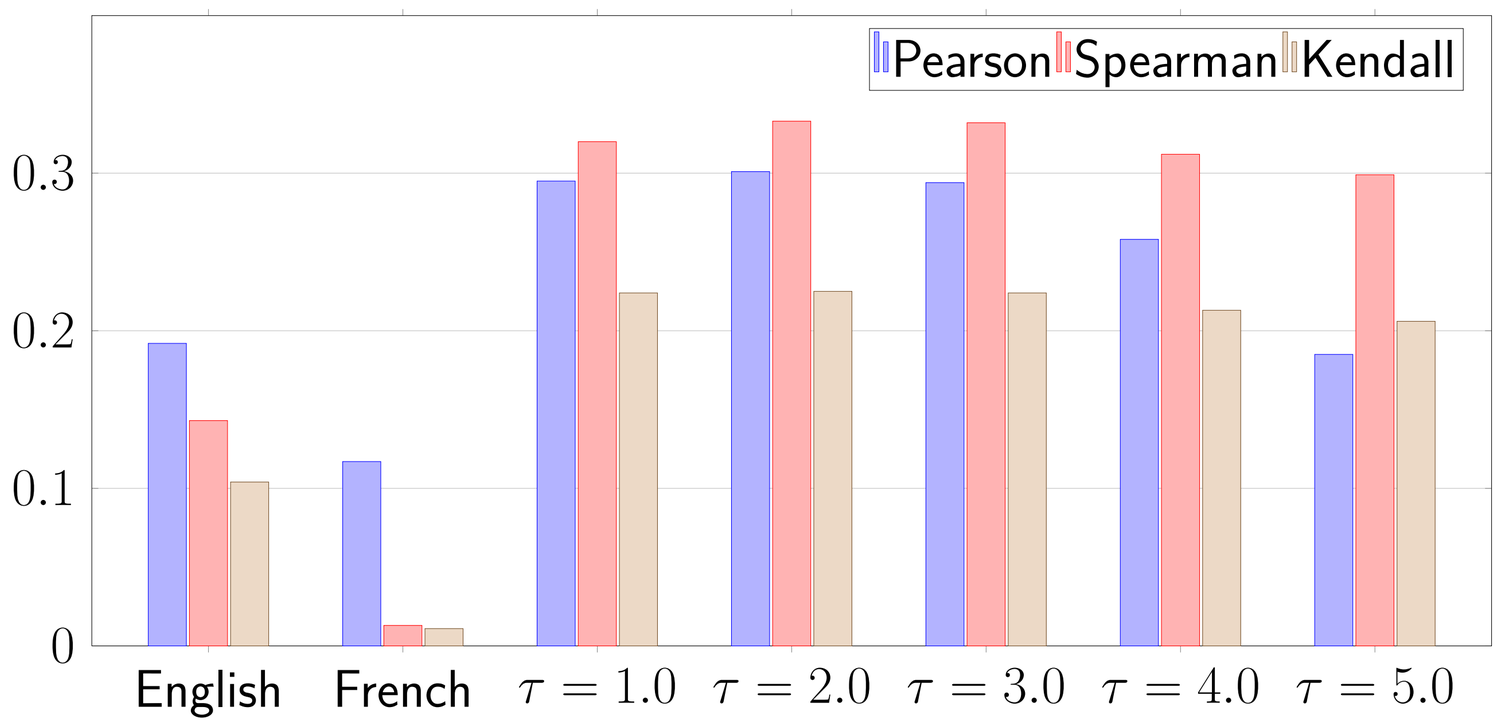
	jaune A	*pain* N	*anglais* N	
yellow A	1.2	0.3	1.1	⋯
bread N	0.3	4.1	0.9	⋯
english N	2.1	1.2	3.2	⋯
	⋮	⋮	⋮	⋱

3. Cross-lingual Rubenstein & Goodenough Dataset

- Merge the English dataset (Rubenstein and Goodenough, 1965) with the French version (Joubarne and Inkpen, 2011) when scores are within ± 1.

English			French			Bilingual		
word1	*word2*	*score*	*word1*	*word2*	*score*	*English*	*French*	*average*
gem	jewel	3.94	joyau	bijou	3.22	gem jewel	bijou joyau	3.58 3.58
car	journey	1.55	auto	voyage	0.33	-	-	-
noon	string	0.04	midi	ficelle	0.00	noon string	ficelle midi	0.02 0.02

4. Results – Three Measures of Correlation



Conclusions and Future Work

- Cross-language measures outperformed the unilingual baselines
- Best PMI threshold was $\tau = 2.0$
- Future Work: Other languages, LSA, new applications, etc.

References

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA; London.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Joubarne, C. and Inkpen, D. (2011). Comparison of semantic similarity for different languages using the Google N-gram corpus and second-order co-occurrence measures. In *Canadian Conference on Artificial Intelligence*, pages 216–221.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sagot, B. and Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Marocco.