Measuring Semantic Relatedness Across Languages

Alistair Kennedy Graeme Hirst

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

December 7, 2012

Introduction

- Distributional Measures of Semantic Relatedness (MSRs).
  - You shall know a word by the company it keeps – [Firth, 1957]
  - Degrees of Relatedness [Rubenstein and Goodenough, 1965]
- Some Questions:
  - How to make a distributional Cross-Language Measure of Semantic Relatedness (CL-MSR) without a parallel corpus?
  - How should we evaluate a CL-MSR?
- Why do we need a CL-MSR?
  - Useful for Machine Translation, Cross-Language Information Retrieval, Language Tutoring Systems, etc.
  - Large bilingual dictionaries and parallel corpora may not always be available

Cross-Language Semantic Relatedness

- So far we have only worked with French and English
- Unilingual Semantic Relatedness
  - "cat" and "feline" – very similar
  - "cat" and "animal" – definitely related
  - "cat" and "hairdryer" – very little in common
  - "cat" and "math" – probably nothing in common
- Between Languages
  - "cat" and "chat" – exact translation
  - "cat" and "féline" – very similar
  - "cat" and "animal" – definitely related
  - "cat" and "sèche-cheveux" – very little in common
  - "cat" and "mathématique" – probably nothing in common

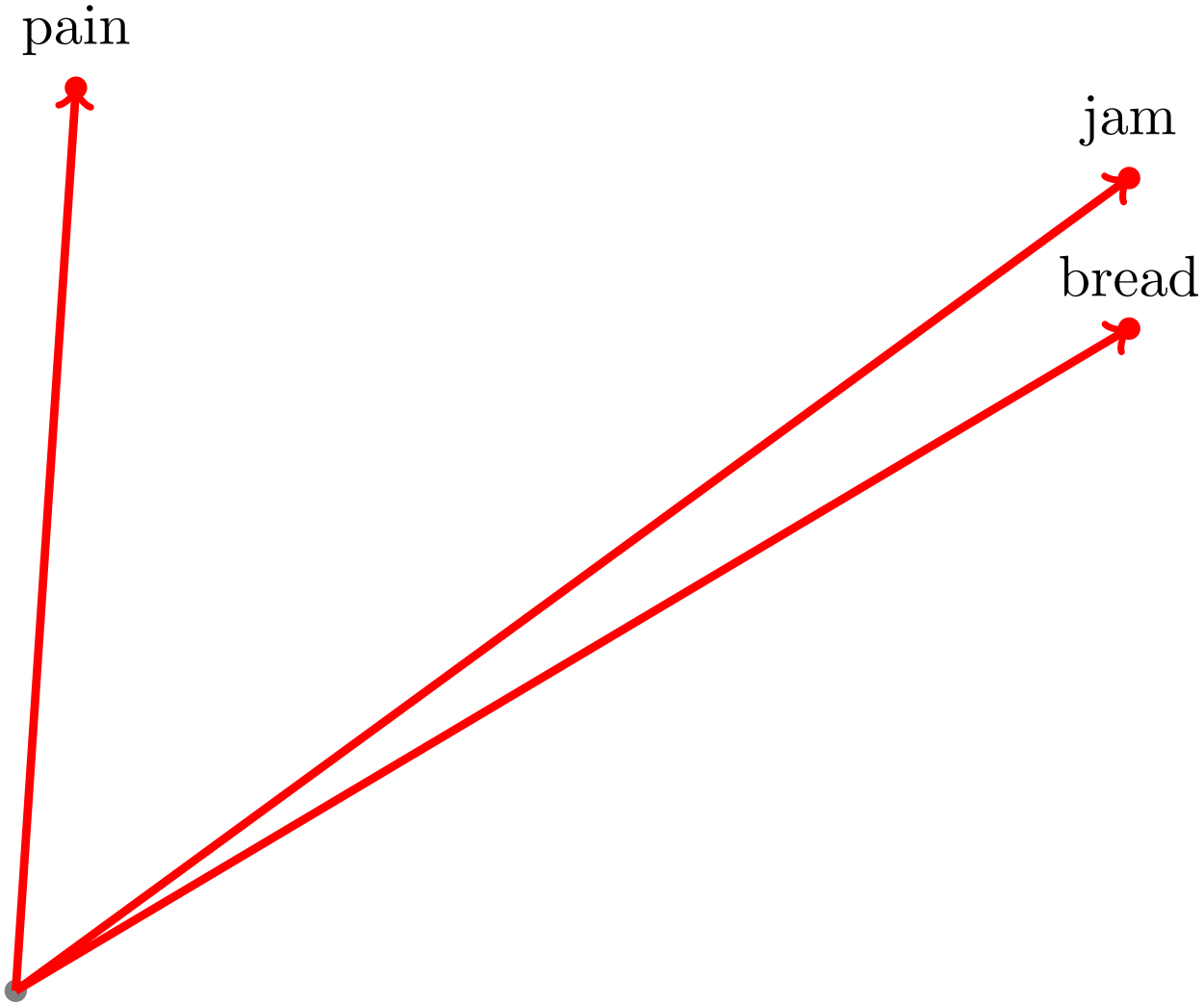
Unilingual Distributional Semantics

- Construct a word-context matrix
  - Used POS-tagged words as contexts
  - Sliding window of 5
- Re-weight matrix – PMI
- Cosine similarity
- Our corpora were the French and English Wikipedias

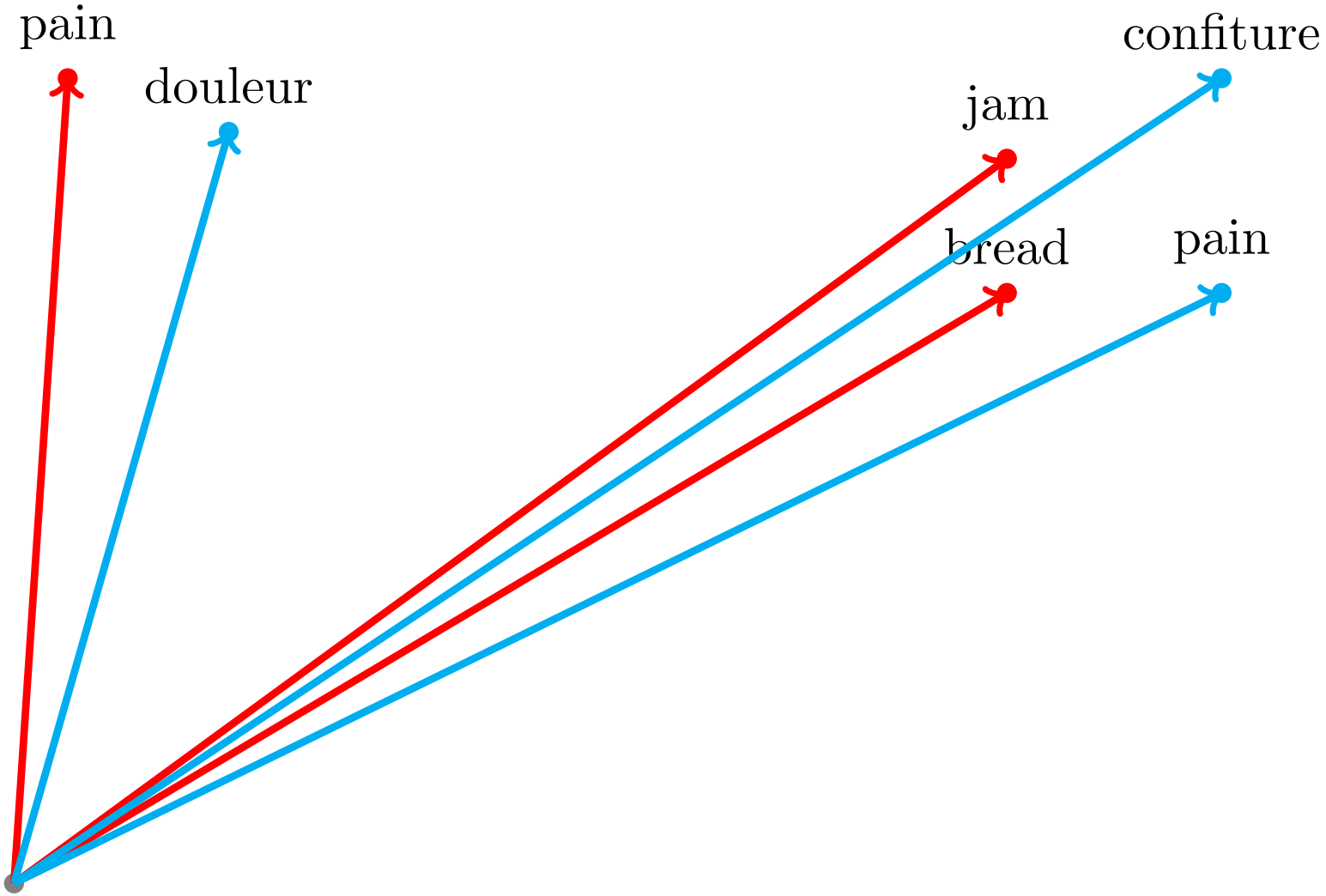
TOAST

0	burnt ADJ	6
1	delicious ADJ	3
2	butter N	9
⋮	⋮	⋮
n	jam N	3

English Vectors



French and English Vectors



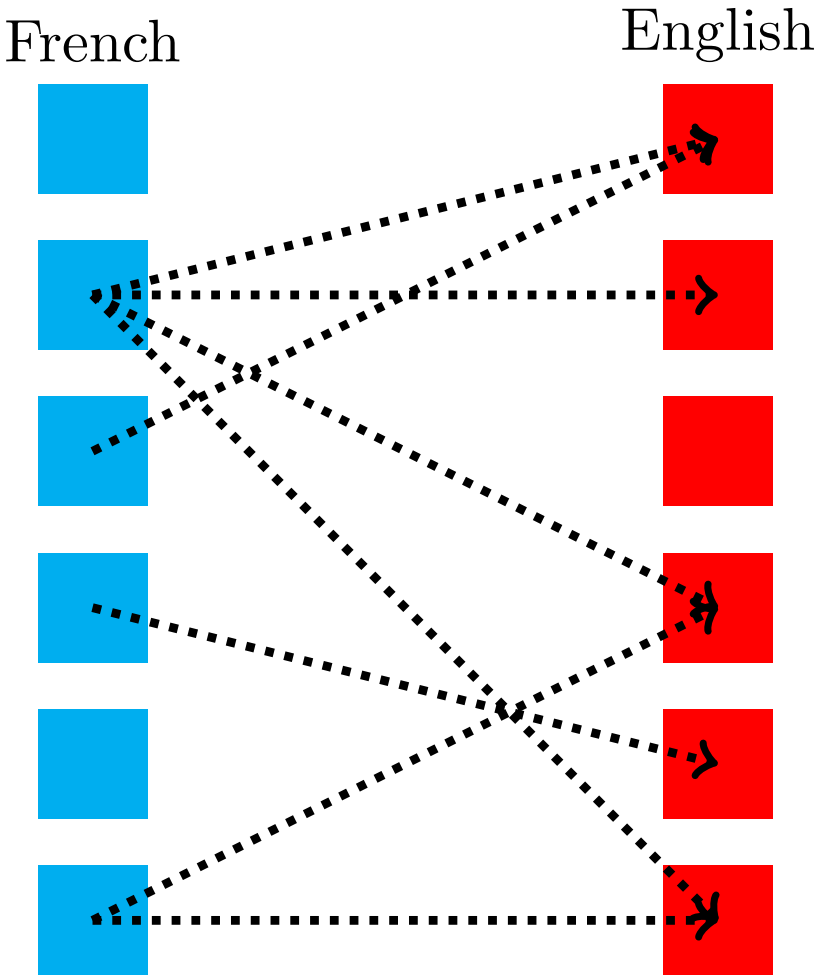
Building a Translation Matrix

- Get translations: $\langle w_e, w_f \rangle$
  - Wordnet Libre du Francais (WOLF) v0.1.5 [Sagot and Fišer, 2008]
  - Princeton WordNet v2.0 [Fellbaum, 1998]

- 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
  - Calculate PMI between c_e and c_f using translations

$$
\begin{array}{r}
 \\
\textit{yellow A} \\
\textit{bread N} \\
\textit{english N} \\
 \\
\end{array}
\begin{array}{c}
\begin{array}{cccc}
\textit{jaune A} & \textit{pain N} & \textit{anglais N} & \\
\end{array} \\
\left[\begin{array}{cccc}
1.2 & 0.3 & 1.1 & \cdots \\
0.3 & 4.1 & 0.9 & \cdots \\
2.1 & 1.2 & 4.2 & \cdots \\
\vdots & \vdots & \vdots & \ddots
\end{array}\right]
\end{array}
$$

Mapping Between Contexts



- Map contexts from French to English using the Translation Matrix
- Two thresholds
  - Minimum PMI score – 1.0, 2.0, 3.0, 4.0 and 5.0
  - Minimum weight of mapping – 0.05
- Merge French and English matrices
  - Label each word with "en" or "fr"

Evaluation

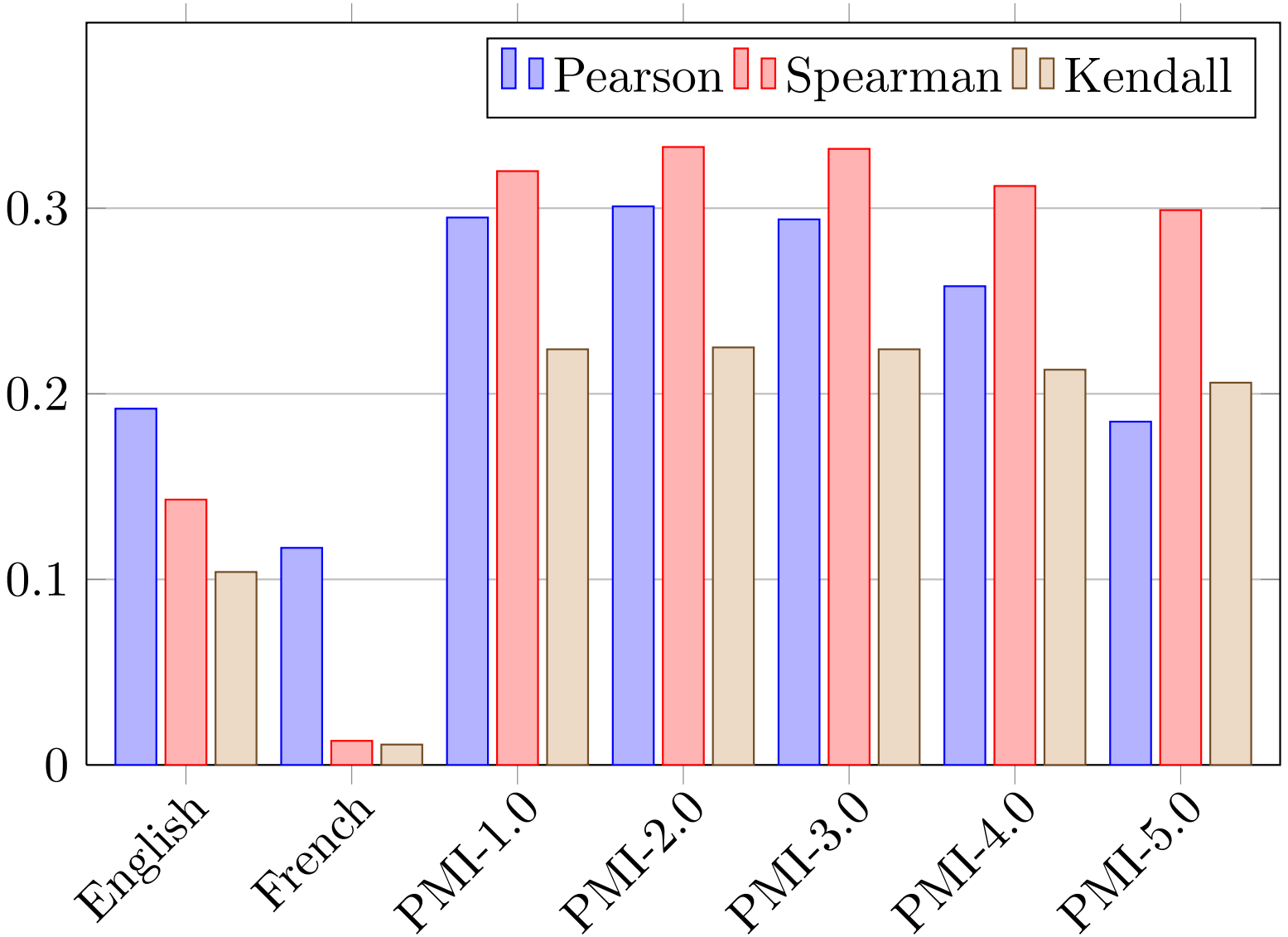
- Rubenstein & Goodenough style data sets
  - English version [Rubenstein and Goodenough, 1965]
  - French version [Joubarne and Inkpen, 2011]
  - 65 word pairs with human scores ranging from 0..4
  - Scores are not identical between the two data sets
- Select matching pairs with scores ± 1
  - 100 French-English pairs
- Evaluate with:
  - Pearson's product-moment correlation coefficient – Score based correlation
  - Spearman's rho – Rank based correlation, measures squares of the deviation
  - Kendall's tau – Rank based correlation, measures number of concording and discording pairs
- Baselines – French and English unilingual MSRs

Cross-Language Rubenstein & Goodenough Data Set

English			French		
gem	jewel	3.94	joyau	bijou	3.22
midday	noon	3.94	midi	dîner	2.17
cemetery	mound	1.69	cimetière	monticule	0.22
cord	smile	0.02	corde	sourire	0.00

Bilingual		
gem	bijou	3.58
jewel	joyau	3.58
midday	*dîner*	*3.05*
noon	*midi*	*3.05*
cemetery	*monticule*	*0.96*
mound	*cimetière*	*0.96*
cord	sourire	0.01
smile	corde	0.01

Bilingual Correlations



Conclusion

- Created a new CL-MSR and Rubenstein and Goodenough style data set
- Our CL-MSRs outperformed the baseline on all three evaluation measures
  - Best PMI threshold was 2.0
- Future work
  - Other languages – German version of Rubenstein & Goodenough
  - LSA – French translated matrix was much more dense than the English one
  - New applications – Cross-Language Information Retrieval, Parallel Corpus Discovery, etc.
  - Experiment with different sources and quantities of translations

Thank You
Questions?

Bibliograpy I

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.