Measuring Semantic Relatedness Across Languages

Alistair Kennedy

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

February 12, 2013

Overview

- Introduction
  - What is Semantic Relatedness?
  - What is Cross-Language Semantic Relatedness?
  - How our work differs from previous work
- Building Measures of Semantic Relatedness
  - Unilingual Measure of Semantic Relatedness (MSR)
  - Cross-Language Measure of Semantic Relatedness (CL-MSR)
- Evaluation
  - Measuring degrees of relatedness
  - Selecting the best translation
- Conclusion and Future Work

Cross-Language Semantic Relatedness

- Unilingual Semantic Relatedness
  - "cat" and "cat" – identical
  - "cat" and "feline" – highly related
  - "cat" and "animal" – related
  - "cat" and "hairdryer" – mostly unrelated
  - "cat" and "math" – completely unrelated
- We have worked with French, English and German
- Between Languages
  - "cat" and "chat" – translation
  - "cat" and "féline" – highly related
  - "cat" and "animal" – related
  - "cat" and "sèche-cheveux" – mostly unrelated
  - "cat" and "mathématique" – completely unrelated

Cross-Language Semantic Relatedness

Continued

- Why do we need a CL-MSR?
  - Machine Translation
  - Cross-Language Information Retrieval
- How to build a CL-MSR?
  - Measure Semantic Relatedness between words without the use of a parallel corpus
- How to evaluate a CL-MSR?
  - Measure degrees of relatedness
  - Select the best translation from a set of candidates

General Methods for Measuring Semantic Relatedness

- Resource based approaches
  - Relatedness between two words is measured by how close the appear in a resource
  - Unilingual measures use resources such as *WordNet*
  - Cross-language wordnets or bilingual dictionaries
- Distributional approaches [Firth, 1957]
  - Words that regularly appear in the same contexts will often have the same meaning
  - A problem: Two languages rarely contain overlapping contexts
- Hybrid approaches
  - Mixes distributional and resource based sources of relatedness
  - **Our Method**: Using a set of known translations we can map distributional representations between two languages

Evaluating a Measure of Semantic Relatedness

- Datasets in the style of [Rubenstein and Goodenough, 1965]

Word 1	Word 2	Score
gem	jewel	3.94
midday	noon	3.94
cemetery	mound	1.69
car	journey	1.55
noon	string	0.04
cord	smile	0.02

Distributional Semantics

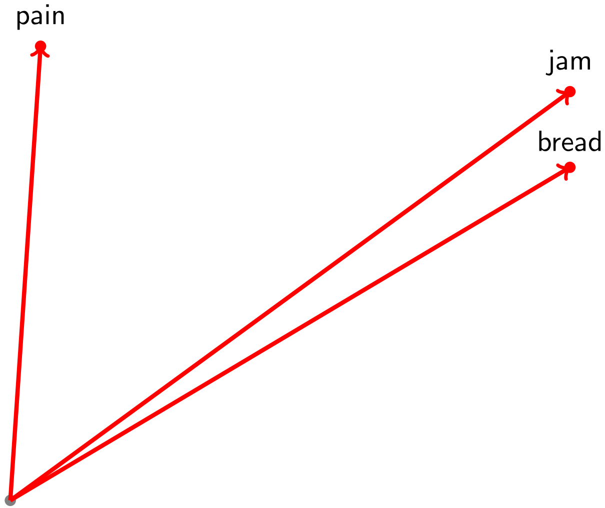
- Construct a word-context matrix
  - Used POS-tagged words as contexts
  - Sliding window of 5
- Re-weight matrix
- Measure distance between pairs of vectors

$$\cos(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

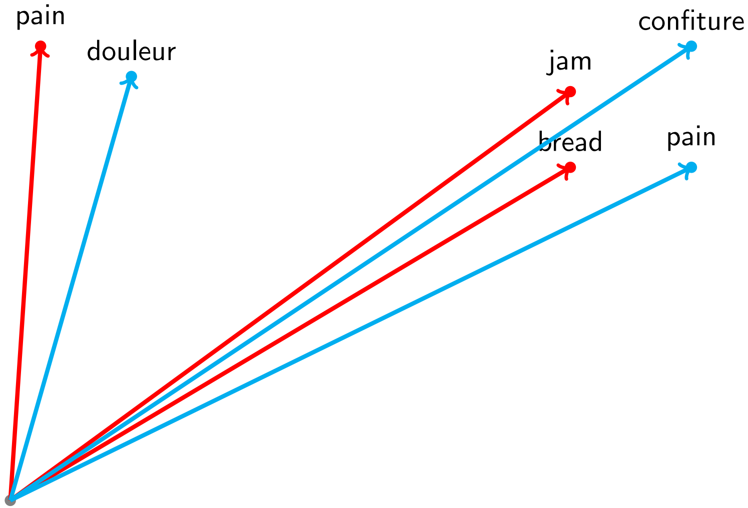
TOAST

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

English Vectors



French and English Vectors



Our CL-MSR

- Use a set of seed translations T between words
- Deduce mapping between context space in source and target languages
- For each pair of contexts c_{source} and c_{target} in two languages:
  - Find pairs of words w_{source} and w_{target} that appear in the respective contexts
  - Identify whether $\langle w_{source}, w_{target} \rangle$ is a valid translation
  - Measure association between c_{source} and c_{target}
  - Pointwise Mutual Information (PMI)
- Many-to-many context mapping
- Extract known pairs of translations from Wiktionary
  - http://www.dicts.info/uddl.php
  - Previously experimented with using aligned wordnets

Previous work on CL-MSRs

- Parallel corpora or directly mapping contexts
  - Use a parallel corpus to learn mappings between languages
  - Machine Translation [Agirre et al., 2009]
  - Map the context space directly using known context translations
  - [Rapp, 1999, Garera et al., 2009]
- Graph based approaches
  - [Etzioni et al., 2006, Michelbacher et al., 2010, Mausam et al., 2010, Flati and Navigli, 2012]
  - Build a Graph where nodes are words and edges like closely related words
  - Add edges between nodes of two languages for each known translation
  - Graph matching between languages to infer known translations

Previous work on CL-MSRs

Continued

- Latent Representations
  - Canonical Correlation Analysis (CCA) [Haghighi et al., 2008, Daumé and Jagarlamudi, 2011]
  - Finds a maximum bipartite matching
  - Word contexts and character n-grams used as features
  - Cross Language Latent Dirichlet Allocation (LDA) [Vulić et al., 2011, Vulić and Moens, 2012]
  - Generative model – topics generate words in two languages
- Use other bilingual resources
  - Bilingual Explicit Semantic Analysis (ESA) [Hassan and Mihalcea, 2009]
  - Bilingual resources like cross-language Wikipedia links to map words into a single representation
  - Mapping is between known translations of contexts, not known translations of words

Building a Unilingual MSR

Unilingual Term-Context Matrices

- Corpora
  - French, German and English Wikipedias – July 2012
  - Part-of-speech (POS) tagged with Stanford POS tagger [Toutanova and Manning, 2000, Toutanova et al., 2003]
- Unique Matrix for each language
  - POS tagged unigram matrix
  - Use sliding window of 5
  - Only use other nouns, verbs and adjectives as contexts
  - Keep only words and contexts that appear > 100 times

Language	Nouns	Contexts	Non-zero entries
English	62,169	106,581	88,662,507
French	28,530	53,658	31,048,865
German	105,989	89,883	52,532,551

Weighted Word-Context matrix

- Unilingual matrices are built for all three languages
- Three versions of each matrix
  - count only
  - PMI
  - PMI + LSA

$$\begin{array}{r c}
 & \begin{array}{cccc} \text{red A} & \text{drive V} & \text{wheel N} & \end{array} \\
\begin{array}{r} \text{apple} \\ \text{car} \\ \text{cheese} \\ \\ \end{array} &
\begin{bmatrix} 6.1 & 1.3 & 0.1 & \cdots \\ 3.3 & 5.1 & 1.9 & \cdots \\ 0.1 & 0 & 3.2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}
\end{array}$$

Reweight Matrix

- Pointwise Mutual Information (PMI)
  - Measures how much more often a word-context pair are observed together than would be expected
    - Maximizes scores for word-context pairs that usually co-occur
    - Minimizes scores for word-context pairs where the word/context co-occur with many other contexts/words
- Latent Semantic Analysis (LSA)
  - Use Singular Value Decomposition (SVD) – Divisi package
  - Low-rank approximation of the word-context matrix X
    - Reduces noise and dimensionality of the matrix
  - Decompose X into $X = U\Sigma V^T$
    - U and V are orthogonal matrices Σ is a diagonal matrix made up of singular values
    - Find the top $k = 500$ singular values: $X_k = U_k \Sigma_k V_k^T$
    - Distance between words is distance between rows of U_k [Turney and Littman, 2003]

Pointwise Mutual Information

Observed and Expected Values

$$\begin{array}{cc} & \begin{array}{cc} y \in Y & y \notin Y \end{array} \\ \begin{array}{c} x \in X \\ x \notin X \end{array} & \begin{bmatrix} O_{0,0} & O_{0,1} \\ O_{1,0} & O_{1,1} \end{bmatrix} \end{array} \Longrightarrow \begin{bmatrix} E_{0,0} & E_{0,1} \\ E_{1,0} & E_{1,1} \end{bmatrix}$$

$$E_{i,j} = \frac{\sum_y O_{i,y} \quad \sum_x O_{x,j}}{\sum_{x,y} O_{x,y}}$$

$$PMI(x \in X, y \in Y) = \log \frac{O_{0,0}}{E_{0,0}}$$

Building a Cross-Language MSR

Measuring Association between Contexts in two Languages

Measure association between context pairs

For each Source context c_{source}, Target context c_{target} and a set of translation pairs $\langle w_{source}, w_{target} \rangle$:

- $O_{0,0}$ [True Positive] $[x \in X \wedge y \in Y]$: number of translations $\langle w_{source}, w_{target} \rangle$ where $w_{source} \in c_{source}$ and $w_{target} \in c_{target}$;
- $O_{0,1}$ [False Negative] $[x \in X \wedge y \in Y]$: number of translations $\langle w_{source}, w_{target} \rangle$ where $w_{source} \in c_{source}$ but $w_{target} \notin c_{target}$;
- $O_{1,0}$ [False Positive] $[x \in X \wedge y \in Y]$: number of translations $\langle w_{source}, w_{target} \rangle$ where $w_{target} \in c_{target}$ but $w_{source} \notin c_{source}$;
- $O_{1,1}$ [True Negative] $[x \in X \wedge y \in Y]$: number of translations $\langle w_{source}, w_{target} \rangle$ where $w_{source} \notin c_{source}$ and $w_{target} \notin c_{target}$.

Example

- E.g. $c_{source} = \langle yellow, A \rangle$, $c_{target} = \langle jaune, A \rangle$ and word pair $\langle w_{source}, w_{target} \rangle$
  - TP $\langle flower, fleur \rangle$ *flower* is found in context *yellow* and *fleur* is found in context *jaune*
  - FN $\langle lilac, fleur \rangle$ *lilac* is not found in context *yellow* and *fleur* is found in context *jaune*
  - FP $\langle flower, lilas \rangle$ *flower* is found in context *yellow* and *lilas* is not found in context *jaune*
  - TN $\langle lilac, lilas \rangle$ *lilac* is not found in context *yellow* and *lilas* is not found in context *jaune*

Weighting Translations

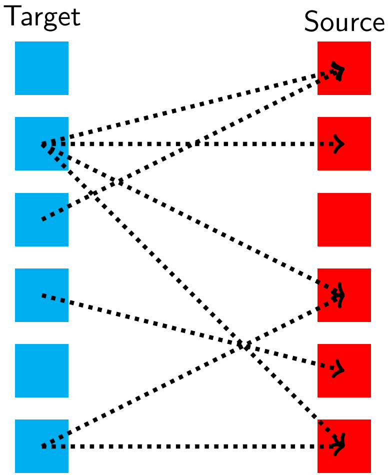
- Each translation $\langle w_{source}, w_{target} \rangle$ in translation set T will be counted as either a TP, FN, FP or TN
  - Should all translations receive the same weight?
  - Assign weights based on values of each word-context pair
- Counts
  - each translation $\langle w_{source}, w_{target} \rangle \in T$ gets a score of 1
  - $weight(c_{source}, c_{target}, w_{source}, w_{target}) = 1$
- Products of PMI socores
  - Each translation $\langle w_{source}, w_{target} \rangle \in T$ receives a unique weight for each context pair $\langle c_{source}, c_{target} \rangle$
  - $weight(c_{source}, c_{target}, \langle w_{source}, w_{target} \rangle) = PMI(c_{source}, w_{source}) * PMI(c_{target}, w_{target})$

Translation Matrix

- Translation matrix generated from PMI-weighted unilingual matrices
- Number of Translations
    - English-French: 1448
    - English-German: 1693
    - French-German: 1869

$$
\begin{array}{rc}
 & \begin{array}{cccc} \text{jaune A} & \text{pain N} & \text{francais N} & \end{array} \\
\begin{array}{r} \text{yellow A} \\ \text{bread N} \\ \text{english N} \\ \\ \end{array} & \begin{bmatrix} 6.2 & 0.0 & 0.0 & \cdots \\ 0.0 & 4.1 & 0.9 & \cdots \\ 0.0 & 1.2 & 2.2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}
\end{array}
$$

Mapping Between Contexts



Mapping Matrices

- Target context is distributed into multiple source contexts
- Source contexts receive weight from multiple targets
- Two translation thresholds
  - Minimum PMI score – tune for threshold
  - Minimum source weight – 0.2
- French, German and English matrices
  - Label each word with "fr", "de" or "en"
- The target languages portion of the matrix is far more dense than the source part
  - Optionally use LSA – 500 dimensions

Tuning Minimum PMI score

- Evaluate on seed translation set T
- Randomly select 1000 source-target translations $\langle w_{source}, w_{target} \rangle \in T$
- For each pair randomly select a Source word $w_{sourceX}$ and an English word $w_{targetX}$ such that
  - $\langle w_{source}, w_{targetX} \rangle \notin T$
  - $\langle w_{sourceX}, w_{target} \rangle \notin T$
- Create two triples $\langle w_{source}, w_{target}, w_{targetX} \rangle$ and $\langle w_{target}, w_{source}, w_{sourceX} \rangle$
- Evaluate CL-MSRs generated using thresholds 1.0, 2.0, 3.0, 4.0 and 5.0
  - Generally a minimum PMI threshold of 2.0 was best

Some Questions

- Will this method work for all language pairs?
- Will applying LSA to the merged cross-language matrices improve scores?
- Does the direction of context mapping matter?
  - E.g. French to English *vs* English to French
- Can we use a hub language for context representation?
  - E.g. French-English CL-MSR represented in German context space
- How many seed translations are needed?
- What are reasonable high/low baselines for the CL-MSR?

Evaluation

Evaluation – Degrees of Relatedness

- Unilingual Rubenstein & Goodenough style datasets
  - English version [Rubenstein and Goodenough, 1965]
  - German version [Gurevych, 2005]
  - 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
- Cross-language Rubenstein & Goodenough style datasets
  - Select matching pairs with scores ± 1
  - 100 French-English pairs
  - 126 English-German pairs
  - 94 German-French pairs

Cross-Language Rubenstein & Goodenough Dataset

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

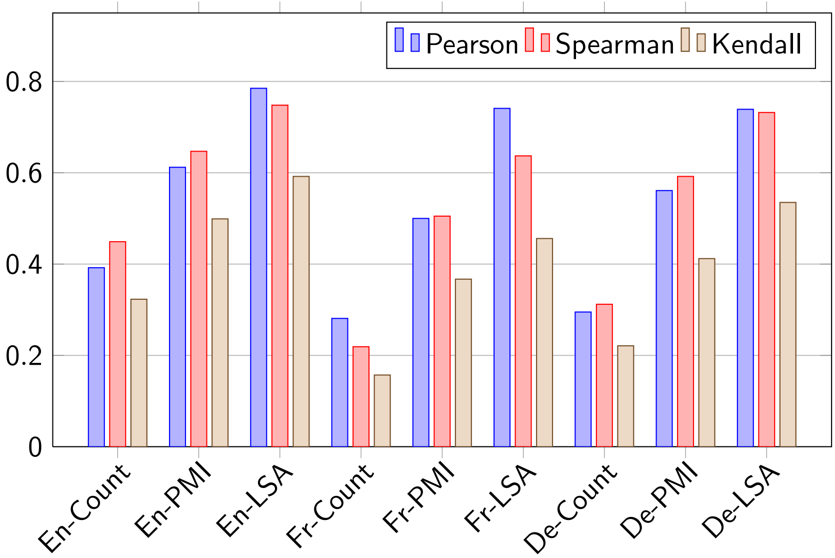
Bilingual		
English	*French*	*average*
gem	bijou	3.58
jewel	joyau	3.58
~~car~~	~~voyage~~	~~0.94~~
~~journey~~	~~auto~~	~~0.94~~
noon	ficelle	0.02
string	midi	0.02

Evaluation – Metrics

- Evaluate with:
  - Pearson's product-moment correlation coefficient – Score based correlation
  - Spearman's rho – Rank based correlation
  - Kendall's tau – Rank based correlation, measures number of concording and discording pairs
- Baselines – unilingual MSRs
  - Many cognates between these language pairs

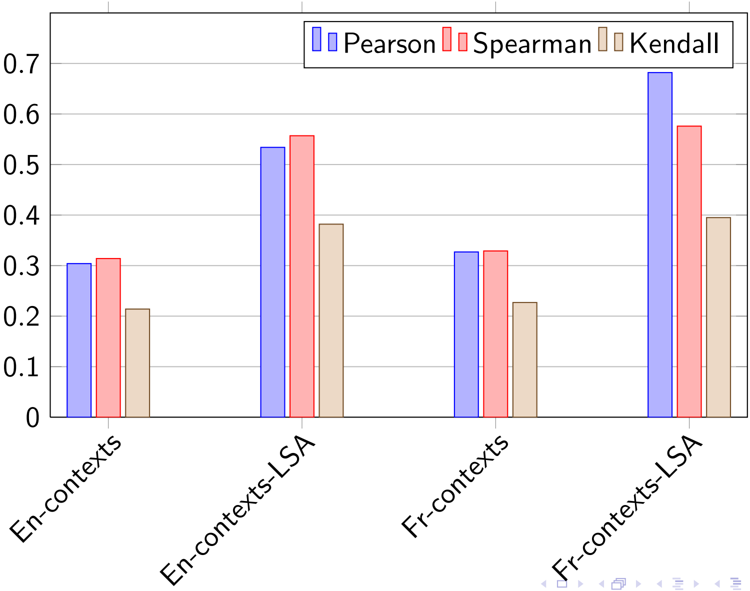
What is a reasonable upper bound for the CL-MSRs?

Correlations on Unilingual Data Sets



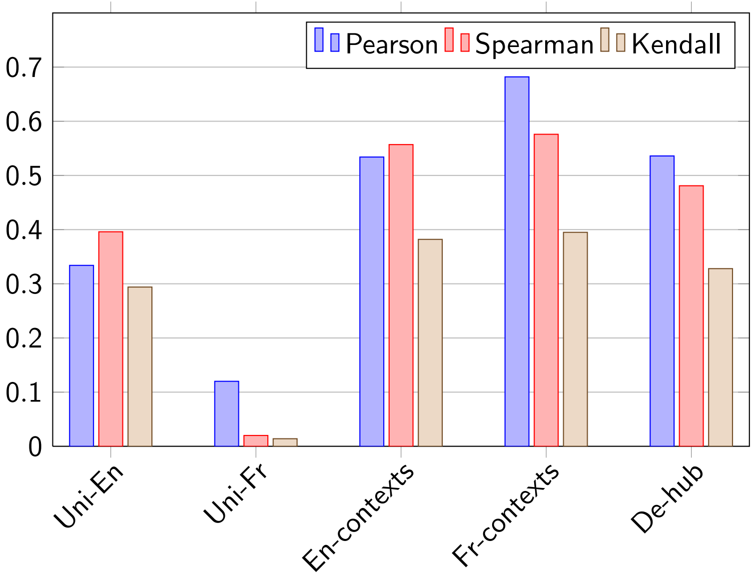
Will LSA improve the CL-MSRs as it does the unilingual MSRs?

LSA *vs* PMI – French-English example

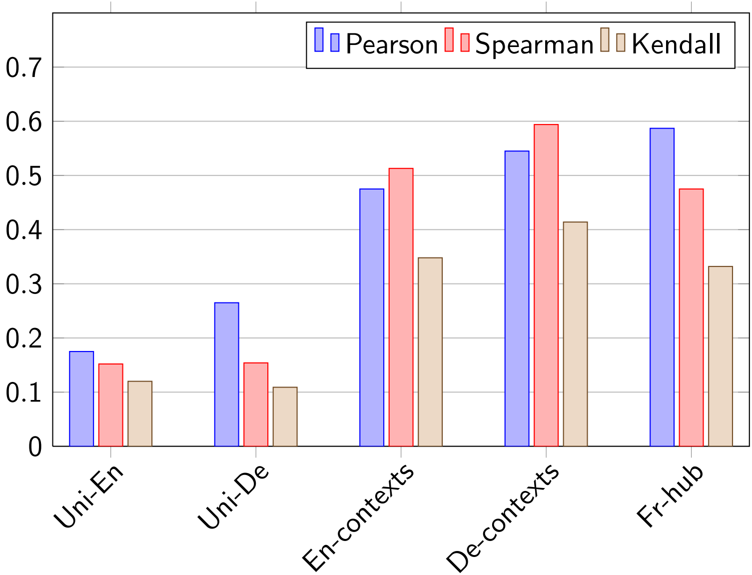


Does the CL-MSR work on all language pairs?
How do they compare to the unilingual baselines?
How does using a hub language affect the results?

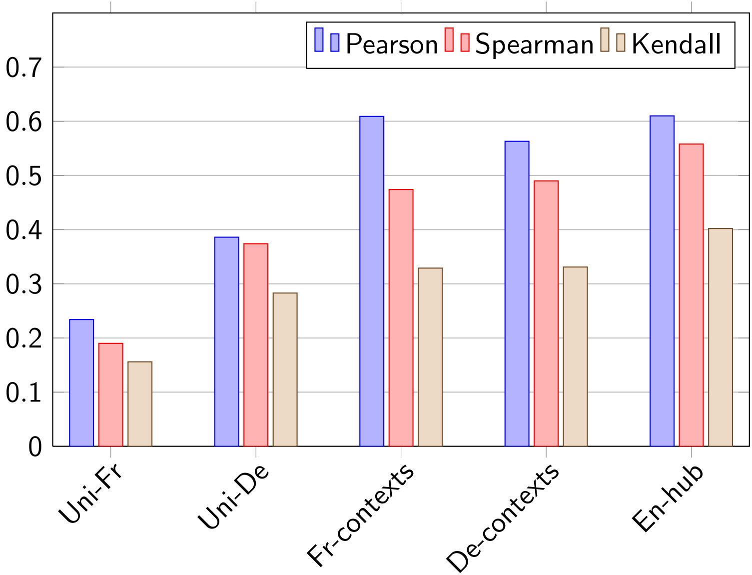
French-English Correlations



German-English Correlations



German-French Correlations



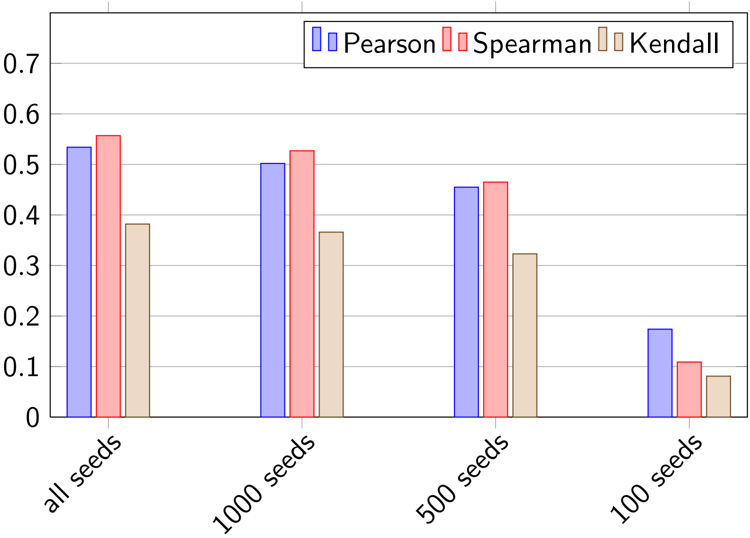
Number of Seed Translations

- How many seed translations are needed?
- Rank seed translations $\langle w_{source}, w_{target} \rangle \in T$
  - $Score(\langle w_{source}, w_{target} \rangle) = Pr(w_{source}) + Pr(w_{target})$
- In order select: all, 1000, 500, or 100 seed translations
- Examples:

French	English	Score
partie	part	0.00482
fois	time	0.00467
nom	name	0.00437
ville	city	0.00377
ville	town	0.00345
nombre	number	0.00290
nom	surname	0.00284

Number of Seed Translations

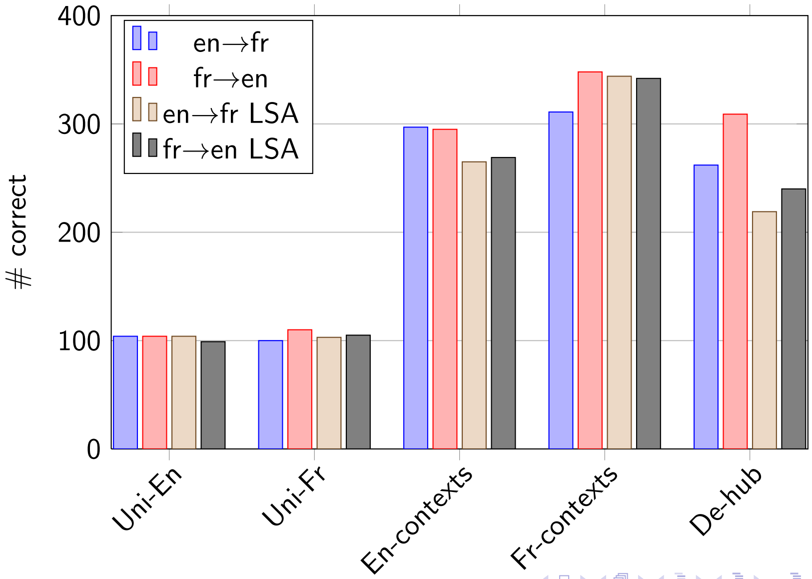
French to English



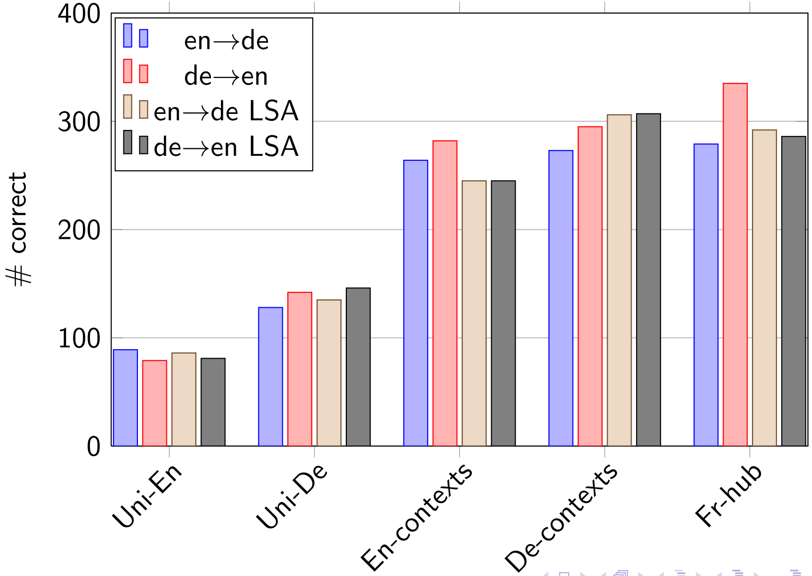
Evaluation – Select the Correct Translation

- Randomly select 400 Source-Target translations $\langle w_{source}, w_{target} \rangle \in T$
  - Use only translations with rank greater than 1000
- For each pair randomly select 3 Source words $w_{sourceX1}, w_{sourceX1}, w_{sourceX3}$ and an Target words $w_{targetX1}, w_{targetX2}, w_{targetX2}$ such that
  - $\langle w_{target}, w_{sourceX} \rangle \notin T$
  - $\langle w_{targetX}, w_{source} \rangle \notin T$
- Create two problems $\langle w_{source}, w_{target}, w_{targetX1}, w_{targetX2}, w_{targetX3} \rangle$ and $\langle w_{target}, w_{source}, w_{sourceX1}, w_{sourceX2}, w_{sourceX3} \rangle$
- Solve problem with the CL-MSR
  - All CL-MSRs trained with 1000 seed translations

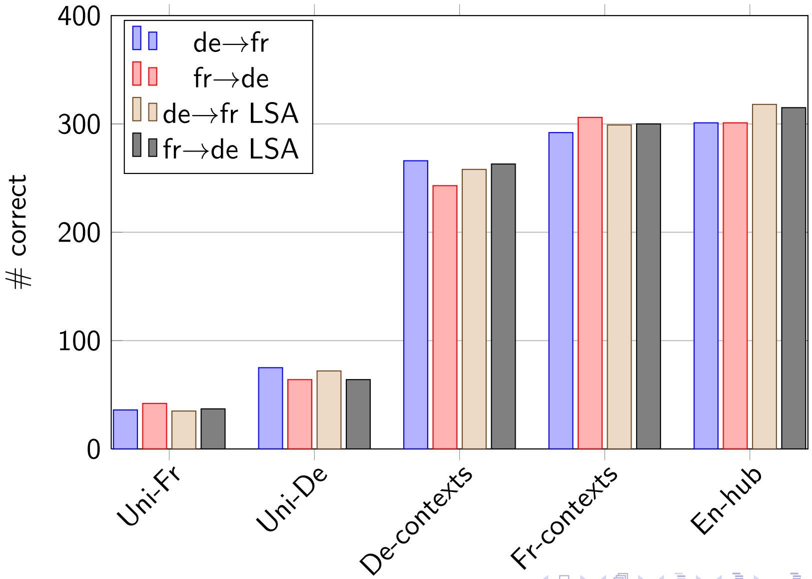
French-English Translations



German-English Translations



German-French Translations



Verbs and Adjectives

- Use similar evaluation methodology to measure relatedness between pairs of verbs and pairs of adjectives
- All experiments so far on French and English
- Comparable results for adjectives
- Poor results for verbs
  - Smaller training set – 600 examples
  - Verbs tend to be polysemous

Nearest Neighbours – Pain

- pain_en – headaches (0.849), discomfort (0.835), fatigue (0.834)
  - **douleur (0.552)**, palpitations (0.274), asthénie (0.245), douleurs (0.244), sueurs (0.241), souffrance (0.238), vertiges (0.233)
- pain_fr – gâteau (0.542), farine (0.502), galette (0.487)
  - paratha (0.487), **bread (0.423)**, chung (0.407), matzo (0.385), jiaozi (0.381), flatbreads (0.380), onigiri (0.378)

Nearest Neighbours – Torpedo

- torpedo_en – replenishments (0.870), wolfpack (0.857), beaching (0.851)
  - bateau (0.202), avion (0.191), cody (0.184), troy (0.175), aéronef (0.173), richie (0.166), brent (0.162)
- torpille_fr – torpilles (0.699), destroyer (0.630), roquette (0.595)
  - portside (0.227), bomb (0.226), firebombs (0.221), shellfire (0.215), salvoes (0.213), airburst (0.286), salvos (0.199)

Conclusion

- When tuning the best minimum PMI threshold was 2.0
- LSA improved results for the Rubenstein & Goodenough style datasets but improvement was not so clear for selecting the best translation
- Correlations for cross-lingual Rubenstein & Goodenough datasets approach those found on the unilingual data sets
- The CL-MSR works comparably measuring distances across French, English and German
- Using a hub language did not strongly help or hurt results
  - Generally mapping larger matrices into the smaller matrices context space worked better
- The more seed translation, the better, though usually 1000 was sufficient
  - Comparable to [Haghighi et al., 2008] and subsequent work

Future Work

- New Applications – Cross Language Information Retrieval, Parallel Corpus Discovery, etc.
    - Compare results against other systems
- More detailed analysis with multiple parts-of-speech
    - verbs and adjectives

Thank You

Questions?

Bibliograpy I

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009).
A study on similarity and relatedness using distributional and wordnet-based approaches.
In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daumé, III, H. and Jagarlamudi, J. (2011).
Domain adaptation for machine translation by mining unseen words.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 407–412, Portland, OR. Association for Computational Linguistics.

Etzioni, O., Reiter, K., Soderl, S., and Sammer, M. (2006).
Lexical translation with application to image search on the web.
In *Proceedings of the 11th Machine Translation Summit*, pages 175–182.

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.

Flati, T. and Navigli, R. (2012).
The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary.
Journal of Artificial Intelligence Research, 43:135–171.

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009).
Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences.
In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 129–137, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bibliograpy II

Gurevych, I. (2005).
Using the structure of a conceptual network in computing semantic relatedness.
In *Proceedings of the Second International Joint Conference on Natural Language Processing*, IJCNLP'05, pages 767–778, Berlin, Heidelberg. Springer-Verlag.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008).
Learning bilingual lexicons from monolingual corpora.
In *Proceedings of The Association of Computational Linguistics: Human Language Technologies*, pages 771–779, Columbus, Ohio. Association for Computational Linguistics.

Hassan, S. and Mihalcea, R. (2009).
Cross-lingual semantic relatedness using encyclopedic knowledge.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, (EMNLP) 2009*, pages 1192–1201. ACL.

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.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Reiter, K., Skinner, M., Sammer, M., and Bilmes, J. (2010).
Panlingual lexical translation via probabilistic inference.
Artificial Intelligence, 174(9-10):619–637.

Michelbacher, L., Laws, F., Dorow, B., Heid, U., and Schütze, H. (2010).
Building a cross-lingual relatedness thesaurus using a graph similarity measure.
In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bibliograpy III

Rapp, R. (1999).
Automatic identification of word translations from unrelated english and german corpora.
In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, College Park, Maryland. Association for Computational Linguistics.

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

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003).
Feature-rich part-of-speech tagging with a cyclic dependency network.
In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Edmonton, Canada. Association for Computational Linguistics.

Toutanova, K. and Manning, C. D. (2000).
Enriching the knowledge sources used in a maximum entropy part-of-speech tagger.
In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics – Volume 13*, EMNLP '00, pages 63–70, Hong Kong. Association for Computational Linguistics.

Turney, P. D. and Littman, M. L. (2003).
Measuring praise and criticism: Inference of semantic orientation from association.
ACM Trans. Inf. Syst., 21(4):315–346.

Bibliograpy IV

Vulić, I., De Smet, W., and Moens, M.-F. (2011).
Identifying word translations from comparable corpora using latent topic models.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 479–484, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vulić, I. and Moens, M.-F. (2012).
Detecting highly confident word translations from comparable corpora without any prior knowledge.
In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 449–459, Stroudsburg, PA, USA. Association for Computational Linguistics.