

Using the Wiktionary Graph Structure for Synonym Detection

Timothy Weale, Chris Brew, Eric Fosler-Lussier

Speech and Language Technologies Lab
Department of Computer Science and Engineering
The Ohio State University
{weale,cbrew,fosler}@cse.ohio-state.edu

Aug. 7, 2009



In This Work

- Utilization of PageRank-inspired technique [Ollivier and Senellart, 2007] for word relatedness on:
 - Wiktionary graph
 - Synonym Detection
- Extension of the technique to include n -best list relatedness and a hybrid relatedness function



Synonym Detection

The Main Idea

Synonyms of a given a source word should return the highest relatedness values from a set of alternative words.

Source Word	Alternative Words
make	earn , print, trade, borrow
flawed	imperfect , tiny, lustrous, crude
solitary	alone , alert, restless, fearless



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Algorithm

- 1 For all four source/candidate word pairs, generate relatedness values for each pair of words using a given metric.
- 2 Return the candidate word that generates the highest relatedness value.
- 3 Evaluate by comparing the candidate word to the true synonym for the source word.

Established Evaluation Task: [Turney, 2001, Jarmasz and Szpakowicz, 2003, Zesch et al., 2008]



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Contributions

- Task-based evaluation of the PageRank-inspired technique proposed by [Ollivier and Senellart, 2007]
- Evaluation of extensions to the technique via n -best list comparisons and a hybrid combination
- Task results are competitive with published state-of-the-art
- Relatedness determined through graph structure alone



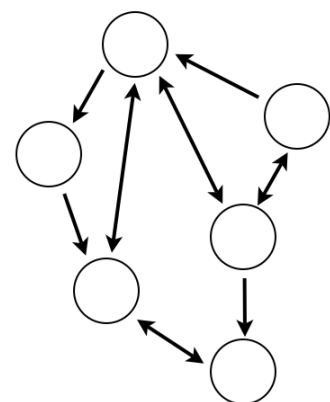
Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Data Source

2009.02.03 Wiktionary Data Set

- Collaboratively constructed dictionary of common words.
- Redirect vertices are pruned.
- Utilization of complete remaining graph structure, including foreign word definitions.
- Approximately 1.1 million vertices.
- Approximately 5.4 million page-page edges.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Word To Vertex Mapping

- Utilize the article titles for mapping
- All words are evaluated individually for multi-word sources/candidates.
 - $rel(\text{"chasm"}, \text{"deep fissure"})$ becomes:
 - $max(rel(\text{"chasm"}, \text{"deep"}), rel(\text{"chasm"}, \text{"fissure"}))$
- More complicated mappings are available [Mihalcea, 2007]



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Graph Structure and Wiktionary

Neighbor Lists [Milne and Witten, 2008]

- Comparison based vertex links.

Path Length [Zesch et al., 2008]

- Comparison based on distance between vertices

Concept Vectors [Zesch et al., 2008]

- Comparison based on word distribution of page text
- Utilizes no graph structure, but has comparable evaluations using Wiktionary



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Three Evaluated Functions

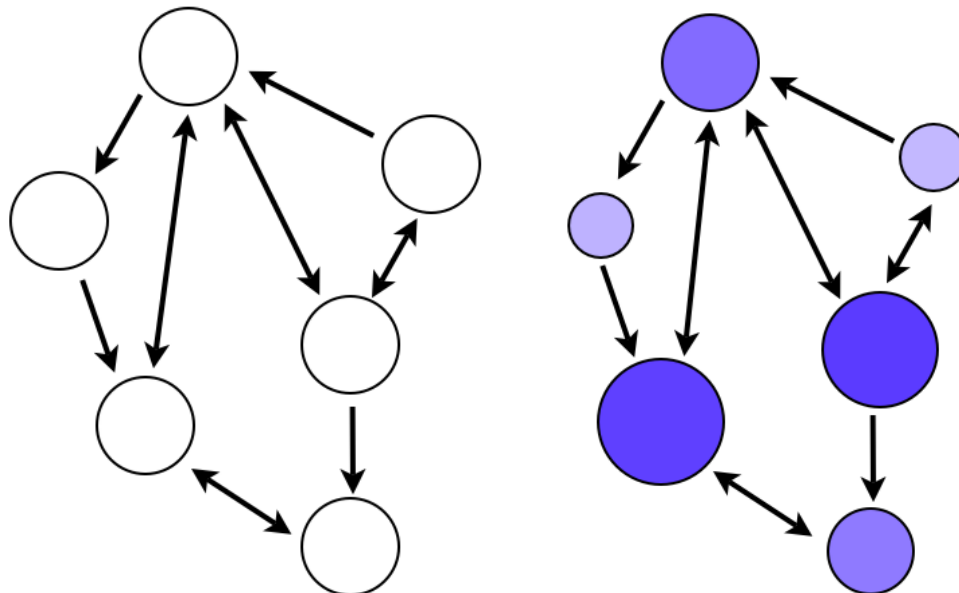
- **Global Graph Relatedness:** Page-Rank inspired graph-based relatedness. Originally developed by Ollivier and Senellart [Ollivier and Senellart, 2007].
- **N-Best Relatedness:** Comparison of the of n -Best vertex lists for the given vertices.
- **Hybrid Relatedness:** Average of the Global Graph Relatedness and N-Best Relatedness.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

PageRank



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Global Graph Relatedness [Ollivier and Senellart, 2007]

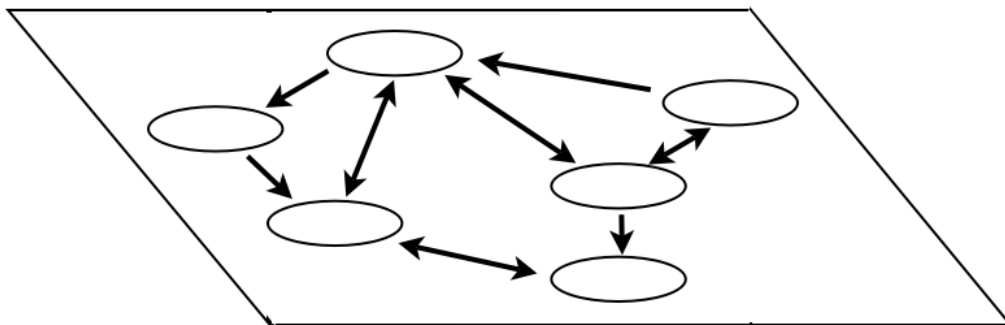
- PageRank gives an a priori indication of importance.
“How important is this vertex to the graph?”
- We’re really interested in:
“How important is this vertex, given a starting point?”
- To determine this, we “pour” weight into the graph at our given vertex and “drain” weight from the graph at each vertex.
- The amount to drain at each vertex is determined by its a priori importance (PageRank value)



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

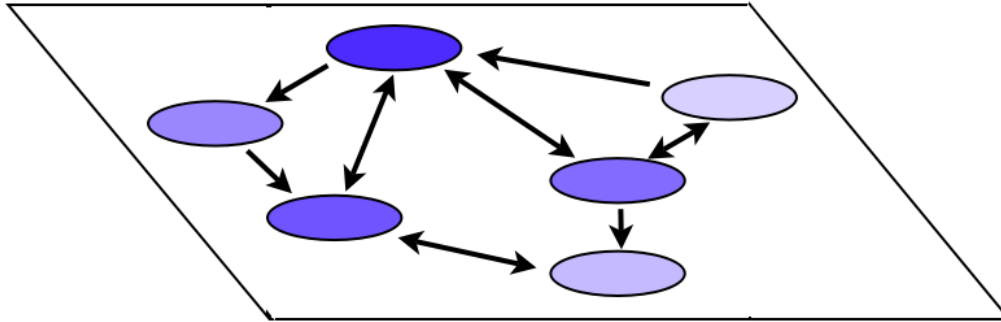
Global Graph Relatedness [Ollivier and Senellart, 2007]



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Global Graph Relatedness [Ollivier and Senellart, 2007]



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

N-Best Algorithm

Related words will have many highly related words in common.

source *n*-best alt1 *n*-best alt2 *n*-best

A		E
B	B	
C		
D	A	F
E		G
F		
G	D	

Consider not only the *number* of commonly related words, but also their *position* in the list.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

N-Best Algorithm

Take the source list to be 'truth'

Finding words highly related to the source should contribute more evidence than marginally related words.

Measure relatedness based on the position of words related to the source word found in the alternative word list.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

N-Best Algorithm

Use the Reciprocal Rank of the word position in the source list.

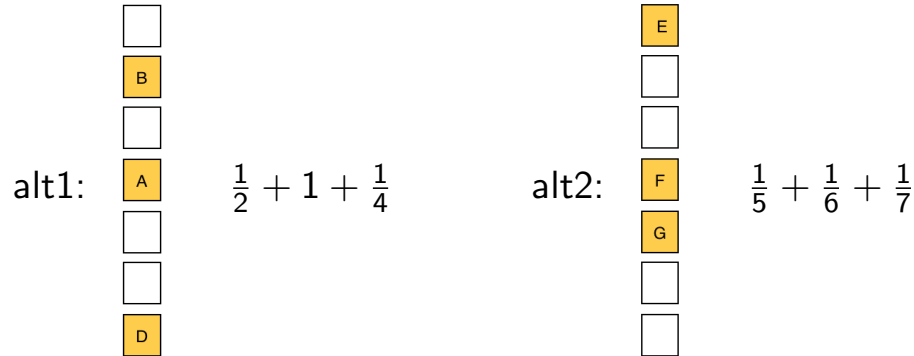
source n -best	relatedness contribution
A	
B	Matching word A contributes 1
C	Matching word B contributes $\frac{1}{2}$
D	
E	Matching word C contributed $\frac{1}{3}$
F	
G	...



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

N-Best Algorithm



Score: **1.75**

Score: **0.51**



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Combined Algorithm

Can we achieve better performance by leveraging information from both metrics?

Average the results from the OS algorithm and the n -best algorithm.

Use normalized results as different metrics have different ranges.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Data Sets

- 50 questions from the English as a Second Language [Turney, 2001]
- 80 questions from the Test Of English as a Foreign Language [Landauer and Dumais, 1997]
- 300 of the Reader's Digest Word Power questions [Jarmasz and Szpakowicz, 2003]

ESL and TOEFL tend to have single-word source/candidate words. RDWP is more complicated and includes multi-word candidates.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

50 ESL

	<i>PL</i>	<i>IR</i>	OS_{Raw}	NB	OS+NB
Correct	41	39	43	40	44
Percent	82.0%	78.0%	86.0%	80.0%	88.0%

PL is a path-length metric using Roget's Thesaurus. [Jarmasz and Szpakowicz, 2003]

IR is a co-occurrence based metric. [Higgins, 2004]



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

80 TOEFL

	<i>PL</i>	<i>IR</i>	<i>OS_{Raw}</i>	NB	OS+NB
Correct	63	65	71	71	75
Percent	78.8%	81.3%	88.8%	88.8%	93.8%

Behind the combined performance of [Turney et al., 2003] (97.5%).
 However, our individual modules outperform their modules.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

300 Reader's Digest Word Power

Metric	Source	Attempted	Score	# Ties	Raw	Prec
<i>PL</i>	Roget's	300	223	0	.74	.74
<i>IR</i>	Web	300	224.33	-	.75	.75
<i>PL_Z</i>	Wiktionary	201	103.7	55	.35	.52
<i>CV_Z</i>		174	147.3	3	.49	.85
<i>OS</i>	Wiktionary	300	234	0	.78	.78
<i>NB</i>		300	212	0	.71	.71
<i>OS + NB</i>		300	227	0	.76	.76



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

300 Reader's Digest Word Power

Metric	Source	Attempted	Score	# Ties	Raw	Prec
<i>PL</i>	Roget's	300	223	0	.74	.74
<i>IR</i>	Web	300	224.33	-	.75	.75
<i>PL_Z</i>	Wiktionary	201	103.7	55	.35	.52
<i>CV_Z</i>		174	147.3	3	.49	.85
<i>OS</i>	Wiktionary	300	234	0	.78	.78
<i>NB</i>		300	212	0	.71	.71
<i>OS + NB</i>		300	227	0	.76	.76



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

300 Reader's Digest Word Power

Metric	Source	Attempted	Score	# Ties	Raw	Prec
<i>PL</i>	Roget's	300	223	0	.74	.74
<i>IR</i>	Web	300	224.33	-	.75	.75
<i>PL_Z</i>	Wiktionary	201	103.7	55	.35	.52
<i>CV_Z</i>		174	147.3	3	.49	.85
<i>OS</i>	Wiktionary	300	234	0	.78	.78
<i>NB</i>		300	212	0	.71	.71
<i>OS + NB</i>		300	227	0	.76	.76



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

300 Reader's Digest Word Power

Metric	Source	Attempted	Score	# Ties	Raw	Prec
<i>PL</i>	Roget's	300	223	0	.74	.74
<i>IR</i>	Web	300	224.33	-	.75	.75
<i>PL_Z</i>	Wiktionary	201	103.7	55	.35	.52
<i>CV_Z</i>		174	147.3	3	.49	.85
<i>OS</i>	Wiktionary	300	234	0	.78	.78
<i>NB</i>		300	212	0	.71	.71
<i>OS + NB</i>		300	227	0	.76	.76



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Discussion

Wiktionary provides adequate and growing coverage for most common words.

- 682,982 Entries used in [Zesch et al., 2008] (Feb, 29 2008)
- About 1.1 Million vertices in this work (Feb 2009)
- 1,339,871 Entries as of Aug 5, 2009



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Discussion

There are cases where OS algorithm gives definitive separation:

- $OS('zenith', 'pinnacle') > OS('zenith', 'decline')$
- Value generated by $OS('zenith', 'pinnacle')$ is 126,000 times greater

OS algorithm provides little separation in some of the wrong cases:

- $OS('consumed', 'bred') > OS('consumed', 'eaten')$
- Value generated by $OS('consumed', 'bred')$ only 1.2 times larger

Addition of information from n -Best algorithm improves results on border cases.



New Data Integration

Our work utilizes only the graph structure.

Wiktionary contains interesting data that's not being used in our current metrics.

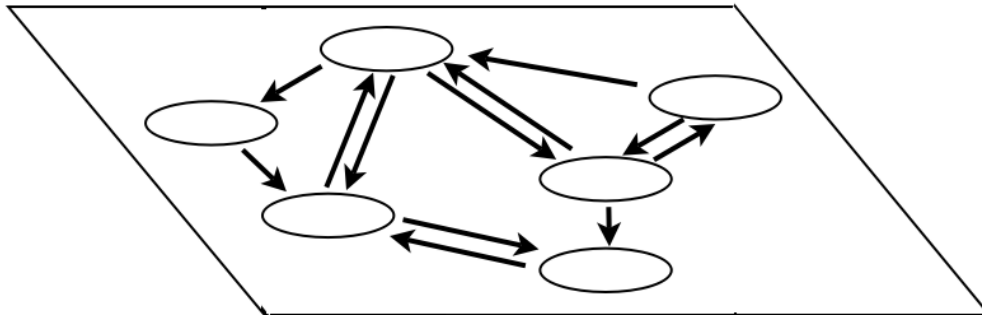
Future work will focus on integrating new data:

- Category Structure
- Article Text
- Graph Changes Over Time

Utilize Wiktionary, Wikipedia for Query Expansion in Information Retrieval [Fang, 2008]



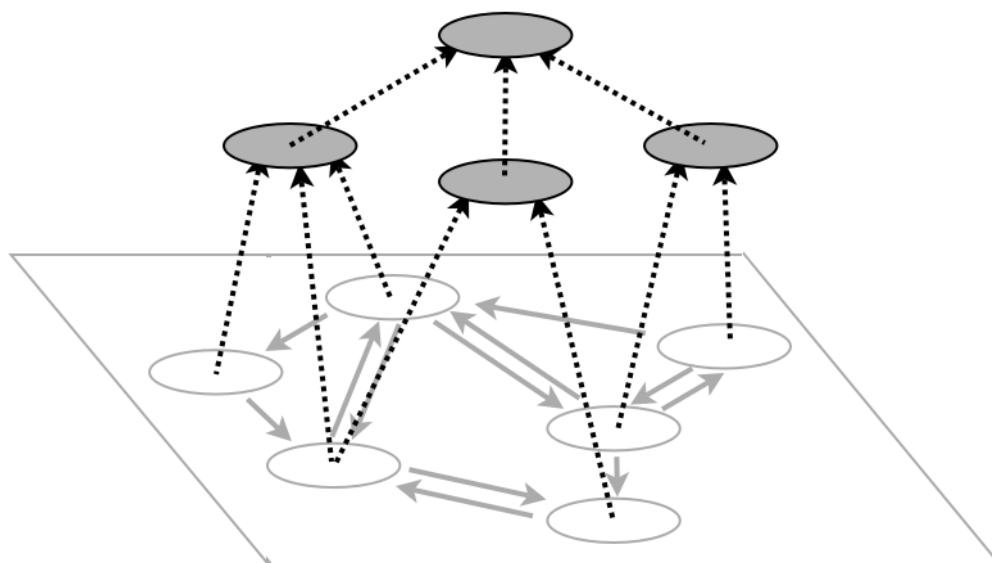
Category Integration



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

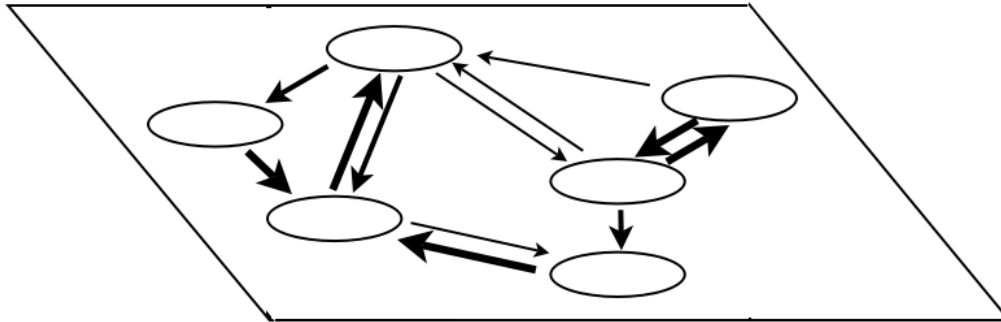
Category Integration



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Category Integration



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

Conclusions

PageRank-inspired metrics using link structure alone are competitive with state of the art metrics.

Wiktionary shows promise as a data source for developing relatedness metrics.

Future work involves:

- Investigating ways to integrate additional Wiktionary information for improved metric creation.
- Investigating alternative ranking methodologies.
- Using Wiktionary/Wikipedia resources for improved application performance.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

THANK YOU

Questions?



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection

- Hui Fang. A Re-examination of Query Expansion Using Lexical Resources. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, June 2008. Association for Computational Linguistics.
- Derrick Higgins. Which Statistics Reflect Semantics? Rethinking Synonymy and Word Similarity. In *Proceedings of the International Conference on Linguistic Evidence*, 2004.
- Mario Jarmasz and Stan Szpakowicz. Roget's Thesaurus and Semantic Similarity. In *Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, 2003.
- Thomas K. Landauer and Susan T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 1997.
- Rada Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of NAACL HLT 2007*, pages 196–203, 2007.
- David Milne and Ian H. Witten. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of AAAI 2008*, 2008.
- Yann Ollivier and Pierre Senellart. Finding Related Pages Using Green Measures: An Illustration with Wikipedia. In *Proceedings of AAAI 2007*, 2007.
- Peter D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freidburg, Germany, 2001.
- Peter D. Turney, Michael L. Littman, Jeffrey Bigham, and Victor Shnayder. Combining Independent Modules to Solve Multiple-Choice Synonym and Analogy Problems. *Proceedings of International Conference RANLP*, 2003.
- Torsten Zesch, Christof Müller, and Iryna Gurevych. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of AAAI 2008*, 2008.



Timothy Weale, Chris Brew, Eric Fosler-Lussier

Using the Wiktionary Graph Structure for Synonym Detection