

Measuring Word Relatedness Using Wikipedia as a Data Source

Timothy Weale

Department of Computer Science and Engineering
The Ohio State University
<http://www.cse.ohio-state.edu/~weale>

2009.02.20



Outline

- 1 Background Information
 - Corpus-Based Measures
 - Ontology-Based Measures
 - Graph-Based Measures
- 2 Current Work
 - Data
 - Word Pair Evaluation
 - Information Retrieval
- 3 Future Investigations



Motivation

- My research interest is in automatic measurement of word relatedness.
- Many current technologies cannot recognize words related to those already hard-coded in the system.
- Knowledge is hard-coded in a application and must be matched exactly to be used (web searches, ASR, Dialogue Systems).
- Inexact word matches hold information that may be applicable to our current information needs.
- {Dog, Canine}, {Dog, Puppy}
{Stock, Bond}, {Stock, Mutual Fund}



Motivation

Word relatedness measures can be applied to a variety of problems:

- **Natural Language Processing**
 - *Word Sense Disambiguation*:
"Apple" (the fruit) OR "Apple" (the company)? [Mih07]
 - *Coreference Resolution*: Resolving "Barack Obama", "Obama", "The President" to the same 'item' in the world. [YS07]
 - *Word Spelling Errors*: "poker" vs. "poke"
- **Information Retrieval**
 - *Query Expansion*:
Include search words related to the query string [QF93, Fan08]
 - *Text Categorization*:
Automatically categorize news documents, websites [Gab06]



Research Question In A Slide...

Develop a function to measure the relatedness of two words.

$$f(w_1, w_2) = x \quad (1)$$

w_1 and w_2 are the input words

x is the measure of relatedness between the two input words and

$$0 \leq x \leq 1$$

At this time, we are not interested in WHY the two words are related, just that they ARE related in some sense.



Research Variables

- 1 **Data Source** – the data set mined for relatedness. Sources differ in many aspects that may affect the resulting weights, including topic coverage, labeling methodology and internal organization.
- 2 **Function Form** – the internals of the relatedness function. Different functions weigh data features differently and are applicable only for certain types of data sources.
- 3 **Target Task** – the application domain for the relatedness metric. Word relatedness has been used in fields such as Natural Language Processing (NLP) and Information Retrieval (IR). For example, is relatedness alone sufficient $\{good, bad\}$ or do they have to be semantically similar $\{excellent, great\}$?



Overview

The Main Idea

Related words will have related distributions in a corpus.

With a large enough corpus, words that are related are more likely to be found in identical contexts.

“The *dog* chased him.”
“The *puppy* chased him.”
“The *puma* chased him.”
“The *Microsoft* chased him.”



Pattern-Based

Hearst¹ provided one of the first corpus-based algorithms.

Look for patterns that indicated lists of related items:

- “ NP_0 such as $\{NP_1, NP_2, \dots, (and/or)\} NP_n$ ”
- “such NP_0 as $\{NP_1, NP_2, \dots, (and/or)\} NP_n$ ”

These discover more descriptive relationships than my research requires – simply knowing a relationship exists is sufficient for my needs.

¹[Hea92]



Distribution-Based

- Words are considered co-occurring if they are consistently and uniquely found within a window of size k words.
- **Bi-gram** [DLP99]
Relatedness based on bi-gram distribution relationship.
- **Webpage** [Tur01]
Relatedness based on document co-occurrence



Overview

The Main Idea

Related words will be 'close' to each other in an ontology.

Ontology is an organized hierarchy of concepts and entities used to categorize and organize elements in the world.

Items close to each other in the ontology tree will be more related than those farther away in the tree.

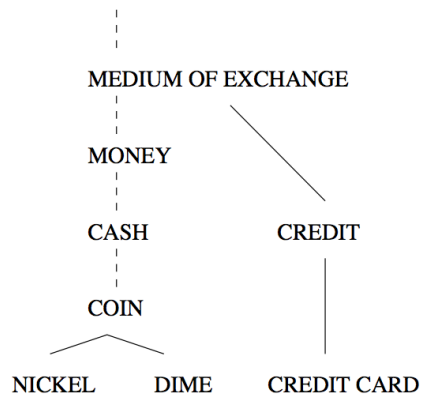


Ontology-based Measures

Path-Based: Relatedness is based on distance between nodes in hierarchy. [HSO98, Jar03]

$$f_{HS}(w_1, w_2) = C - \text{pathlength} - k * d$$

$$f_{LC}(w_1, w_2) = -\log \frac{\text{len}(w_1, w_2)}{2D}$$



Timothy Weale

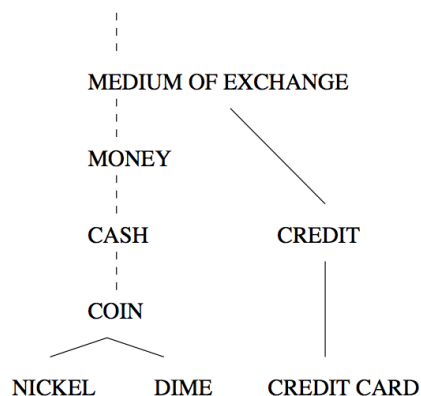
Measuring Word Relatedness Using Wikipedia as a Data Source

Ontology-based Measures

Information-Based: Relatedness is based on coverage of parent topic of both original nodes. [Res99, Lin98]

$$f_R(w_1, w_2) = -\log p(\text{Iso}(w_1, w_2))$$

$$f_{JC}(w_1, w_2) = 2 * \log p(\text{Iso}(w_1, w_2)) - (\log(p(w_1)) + \log(p(w_2)))$$



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Resources

Several existing human-created hierarchies:

- **WordNet:** Semantic lexicon; Groups words into sets of synonyms called *synsets*. Records the synsets in a hierarchy of conceptual organization.
- **Roget's Thesaurus:** Also groups sets of words into sets of synonyms. Geared more towards writing style than semantic information.
- **Wikipedia:** User-created categories give hierarchical organization to topic pages.



Timothy Weale

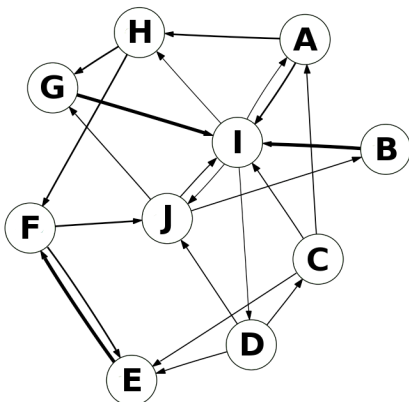
Measuring Word Relatedness Using Wikipedia as a Data Source

Overview

The Main Idea

Related words will share many common neighbors.

Wikipedia provides direct page-page links that create an overall Wikipedia graph.



- **Local:** Relationships are based on a comparison of the page-page links used by a vertex of the graph.
- **Global:** Treat the page-page graph as a searchable 'web' for related pages.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Local Graph Relatedness

- Related pages will have related link structures.
- “*Star Wars*” will have more neighbors in common with “*George Lucas*” than “*Potato*”.
- Derive relatedness based on two different metrics:
 - Cosine Metric [Mil07]
 - Google Normalized Distance [MW08]



Local Graph Relatedness [Mil07]

- Some links are more informative than others
 - Over 9000 pages link to the article on *water* (a common chemical substance)
 - Only 34 pages link to the article on *focaccia* (a type of bread)
- The less likely a link is, the more information it contains

$$w(s \rightarrow t) = \log \left(\frac{|W|}{|T|} \right) \text{ if } s \in T \quad (2)$$

- Creates a weighted vector of links: $\langle w(s, t_1), \dots, w(s, t_{|W|}) \rangle$
- $rel(a, b) = \cos(\vec{A}, \vec{B})$



Local Graph Relatedness [MW08]

- Based on Google Normalized Distance

$$\text{dist}(a, b) = \frac{\log(\max(|A|, |B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|, |B|))} \quad (3)$$

Where A , B are the sets of pages that link to page a , b respectively



Global Graph Relatedness [OS07]

Leverage PageRank for vertex relatedness determination.

$$\vec{P}_n = \alpha \times (P_{n-1} \times T) + (1 - \alpha) \times \vec{W} \quad (4)$$

- \vec{P} : **PageRank Vector**. Holds the measure of vertex 'importance'.
- T : **Transition Matrix**. Contains edge transition probabilities.
- \vec{W} : **Walk Vector**. Uniform vector used to model 'random walk' behavior.
- α : **Model Weighting**. Tradeoff between strict PageRank behavior and Random Walk behavior (usually 0.85)



Global Graph Relatedness [OS07]

- 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)



Global Graph Relatedness [OS07]

$$\vec{R}_n = R_{n-1} \times T + (\vec{S} - \vec{P}) \quad (5)$$

- **\vec{R} : Relatedness Vector.**
- **T : Transition Matrix.** Edge transition probabilities (same as PageRank).
- **\vec{S} : Source Vector.** Contains a 1 at the source vertex and zeros in all other locations.
- **\vec{P} : Sink Vector.** Subtracts a value from the vertex values based on their PageRank value.

$$rel(a, b) = \vec{R}_a[b] \times \log \left(\frac{1}{\vec{P}[b]} \right) \quad (6)$$



Wikipedia Data

- Organized by “page”
- Every page belongs to one of several potential types:
 - **Main:** Encyclopedia-type pages. “Content” pages.
 - **Talk:** Discussion and coordination pages for updating main pages.
 - **Category:** User-defined category structure for organization/navigation.
 - **User:** Information about Wikipedia contributors.
- Links may be made between pages of the same type or different types.
- Additionally, pages may be flagged as ‘redirect’ pages whose sole purpose is to forward a user to another page.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Implementation

- 2008.01.03 Wikipedia Data Set
- Graph-based Metrics implemented in Java
 - Graph creation from raw Wikipedia SQL files
 - Metric implementation from resulting graph
- Competing WordNet-based metrics implemented in Perl²

²<http://wn-similarity.sourceforge.net/>

Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source



Implementation

2008.01.03 Wikipedia Data Set

Page Counts

| | |
|----------------------|------------|
| All Pages | 11,372,877 |
| Main Namespace | 4,610,749 |
| Final Graph Vertices | 2,179,265 |

Link Counts

| | |
|---------------------|-------------|
| All Links | 208,173,450 |
| Main Namespace | 136,492,176 |
| Non-Redirected Main | 80,710,797 |



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Word Pair Evaluation

| Data Set | Sample Size | Scale |
|-----------------------|-------------|------------|
| Rubenstein-Goodenough | 65 | 0.0 – 4.0 |
| Miller-Charles | 30 | 0.0 – 4.0 |
| WordSim-353 | 353 | 0.0 – 10.0 |

Evaluation is then done by finding the Pearson's Correlation Coefficient using the human-evaluated pairs and the results of the relatedness function.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Miller Charles Evaluation

Created a hand-crafted data set from Wikipedia.

| Algorithm | Correlation |
|--------------------------|-------------|
| WikiRelate | 0.59 |
| Green[†] | 0.63 |
| WLM | 0.70 |
| ESA | 0.73 |
| Resnik | 0.75 |
| Path | 0.79 |

It's a good start for our first pass.
 Improvements are possible with additional information and better data sets.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Top-n Evaluation

| Weale | Ollivier[OS07] |
|-------------------------|-----------------------|
| Clique (graph theory) | Clique (graph theory) |
| Graph theory | Graph (mathematics) |
| NP-complete | Graph theory |
| Graph (mathematics) | Category:Graph theory |
| Clique problem | NP-complete |
| Independent set | Complement graph |
| Complement graph | Clique problem |
| Complete graph | Complete graph |
| Mcsip ³ | Independent set |
| Clique (disambiguation) | Mcsip |

³Maximum common subgraph isomorphism problem



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Top-n Evaluation

| Weale | Ollivier[OS07] |
|-----------------------|-------------------------------|
| Star Wars | Star Wars |
| Star Wars Episode IV | Dates in Star Wars |
| Star Wars Episode III | Palpatine |
| Star Wars Episode II | Jedi |
| George Lucas | Expanded Universe (Star Wars) |
| Star Wars Episode I | Star Wars Episode I |
| Luke Skywalker | Star Wars Episode IV |
| Jedi | Obi-Wan Kenobi |
| Palpatine | Star Wars Episode III |
| Darth Vader | Coruscant |



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Query Expansion

$$Q = \{q_1, q_2, \dots, q_i\}$$

Can we augment the query with set of words that will improve the returned results of the query?

$$Q' = \{q_1, q_2, \dots, q_i, q_{i+1}, \dots, q_{i+n}\}$$



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

Query Expansion

For each query word find top k related words

$$K_1 = \text{topN}(q_1, W, f(\cdot)); \dots$$

Add the top- k words to a candidate list

$$C = \{K_1, K_2, \dots, K_n\}$$

Re-rank the candidate list

$$\text{rerank}(C)$$

Pull out the top n words for expansion



To Do:

Wikipedia 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

Can adding this information help improve performance?

Where can this information be added?
Larger graph? Smarter edge weights?



To Do:

Translation of Green measures into values more suitable for application domains.

Wikipedia XML data can be used to provide better surface form to page translation.

(?) Integrate ranking intuitions into relatedness metrics (?)



Conclusions

Word relatedness metrics are established tools in Natural Language Processing and Information Retrieval.

Our work looks at the potential for using Wikipedia as a source for developing relatedness metrics.

Preliminary results are promising, but there is more work to be done in both the application domain and also in investigating ways to integrate the additional information available to us in the data set.



-  [Ido Dagan, Lillian Lee, and Fernando C. N. Pereira.](#)
Similarity-Based Models of Word Cooccurrence Probabilities.
Machine Learning, 32(1):13–47, 1999.
-  [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.
-  [Evgeniy Gabrilovich.](#)
Feature Generation for Text Categorization Using World Knowledge.
PhD thesis, Israel Institute of Technology, December 2006.
-  [Marti A. Hearst.](#)
Automatic Acquisition of Hyponyms from Large Text Corpora.
In *Proceedings of COLING-92*, 1992.
-  [Graeme Hirst and David St-Onge.](#)
Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms.
In *Christiane Fellbaum*, editor, *WordNet*. MIT Press, 1998.
-  [Mario Jarmasz.](#)
Roget's Thesaurus as a Lexical Resource for Natural Language Processing.
Master's thesis, Ottawa-Carleton Institute for Computer Science, July 2003.
-  [Dekang Lin.](#)
An Information-Theoretic Definition of Similarity.
In *Proceedings of ICML 1998*, 1998.
-  [Rada Mihalcea.](#)
Using Wikipedia for Automatic Word Sense Disambiguation.
In *Proceedings of NAACL HLT 2007*, pages 196–203, 2007.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source

-  [David Milne.](#)
Computing Semantic Relatedness using Wikipedia Link Structure.
In *New Zealand Computer Science Research Student Conference (NZCSRSC'2007)*, 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.
-  [Yonggang Qui and H. P. Frei.](#)
Concept Based Query Expansion.
In *Proceedings of SIGIR 1993*, 1993.
-  [Philip Resnik.](#)
Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.
JAIR, 11:95–130, 1999.
-  [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.
-  [Xiaofeng Yang and Jian Su.](#)
Coreference Resolution Using Semantic Relatedness Information from Automatically Discovered Patterns.
In *Proceedings of ACL 2007*, 2007.



Timothy Weale

Measuring Word Relatedness Using Wikipedia as a Data Source