

Measuring Word Relatedness

Timothy Weale
weale.2@osu.edu

2008.03.28

Abstract

Intelligent applications require a basic understanding relatedness among words. For example, “dog” is highly similar to both “puppy” and “canine.” By ignoring related words and concepts, knowledge is unnecessarily limited in its scope and applicability. In my research, I work to automatically discover and measure the relatedness of concepts and words. In this talk, I will provide an overview of the existing research into word relatedness. I will start with a few motivating examples before giving a broad overview of the research topic. Finally, I conclude with some of my initial dissertation work on extending the existing state-of-the-art.

Outline

- 1 Introduction
- 2 Corpus-Based Measures
- 3 Ontology-Based Measures
- 4 Link-Based Measures
- 5 Preliminary Work
- 6 Conclusions

Introduction

Motivation

- My research interest is in the automatic determination and measurement of word relatedness.
- Many current technologies ignore words related to those already hard-coded in the system.
- Knowledge is hard-coded in a knowledge base and must be matched exactly to be used (Google searches).
- Inexact word matches hold information that may be applicable to our current information needs.
- {Dog, Canine, Puppy} or {Stock, Bond, 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)?
 - *Coreference Resolution*: Resolving “George Bush”, “Bush”, “The President” to the same ‘item’ in the world.
 - *Word Spelling Errors*:
“poker” vs. “poke”
- **Information Retrieval**
 - *Query Expansion*:
Include search words related to the query string
 - *Text Categorization*:
Automatically categorize news documents, websites

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 **Semantic Source** – the data set mined for semantic relatedness. Semantic sources differ in many aspects that may affect the semantic weights, including topic coverage, labeling methodology and internal organization.
- 2 **Function Form** – the internals of the semantic similarity function. Different functions weigh semantic features differently and are applicable only for certain types of semantic sources.
- 3 **Target Task** – the application domain for the similarity metric. Semantic 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.”

Origins

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]

Seed new patterns based on known relationships:

- ① Decide on a relation to discover (*group/member*)
- ② Gather a list of terms for which this relationship holds (*England-country*), (*tank-vehicle*)
- ③ Find places in the corpus where these words occur near each other
- ④ Generalize the commonalities among the discovered sentence fragments
- ⑤ Use these rules to discover new relationships in the corpus

More recent work has been done in learning relationships for analogy comparison²

A:B::C:D

Given the words in the analogy, search a corpus for statistics based on 64 different joining patterns (“ for ”, “ is ”, “ instead of ”). This yields a vector of 128 terms. Then compare the pairs of vectors (A and B, C and D) using the cosine similarity metric (more on this later).

About 47% correct on their test set (n=374). This would score about the 29th percentile.

²[TL05]

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.

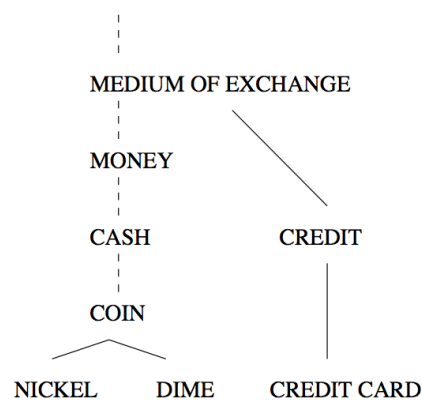
Items close to each other in the ontology tree will be more related than those farther away in the tree.

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' and 'portals' give hierarchical organization to topic pages.

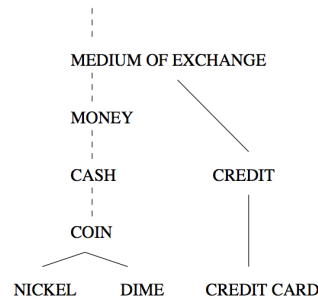
Path-based Measures³



$$rel_{HS}(c_1, c_2) = C - pathlength - k * d$$

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2D}$$

³[BH06]

Information-based Measures⁴

$$sim_R(c_1, c_2) = -\log p(ISO(c_1, c_2))$$

$$dist_{JC}(c_1, c_2) = 2 * \log p(ISO(c_1, c_2)) - (\log(p(c_1)) + \log(p(c_2)))$$

$$sim_L(c_1, c_2) = \frac{2 * \log p(ISO(c_1, c_2))}{\log(p(c_1)) + \log(p(c_2))}$$

⁴[BH06]

Ontology-based similarity metrics have been utilized with all three datasets⁵.

Highly dependent upon on how the ontology was constructed and the original purpose of the ontology.

- **WordNet, Thesaurus:** General textual information
{hello, run, bank}
- **Wikipedia:** Specific entity information
{Dayton Ohio, Angelina Jolie, Wilkins Sound}

⁵[SP06, BH06]

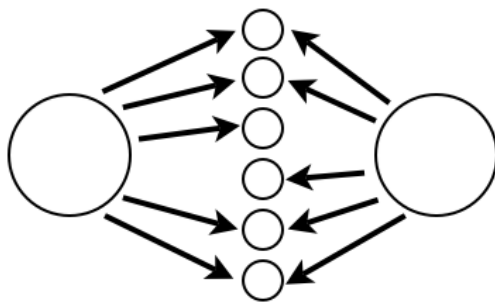
Overview

The Main Idea

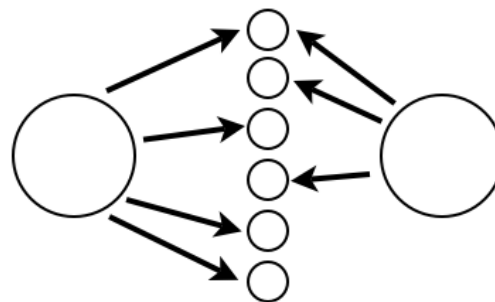
Related words will share many common neighbors.

Wikipedia provides direct page-page links which can be 'searched' like a web.

Link-Based Measures



Strongly Related Pages



Loosely Related Pages

With page-page links, we are able to use common similarity metrics to determine relatedness between the two pages:

$$\begin{array}{ll}
 \text{Matching} & |Q \cap S| \\
 \text{Dice} & 2|Q \cap S| / (|Q| + |S|) \\
 \text{Jaccard} & |Q \cap S| / |Q \cup S| \\
 \text{Overlap} & |Q \cap S| / \min(|Q|, |S|) \\
 \text{Cosine} & |Q \cap S| / \sqrt{|Q| \times |S|}
 \end{array}$$

$$\begin{array}{ll}
 W_0 = \{A, B\} & W_1 = \{B, C, D\} \\
 W_2 = \{A, E, F, G\} & W_3 = \{A, B, E, G, H\}
 \end{array}$$

	W_1	W_2	W_3
Matching	1	1	2
Dice	$\frac{2}{5}$	$\frac{2}{6}$	$\frac{4}{7}$
Jaccard	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{2}{5}$
Overlap	$\frac{1}{2}$	$\frac{1}{2}$	1
Cosine	$\frac{1}{\sqrt{6}}$	$\frac{1}{\sqrt{8}}$	$\frac{2}{\sqrt{10}}$

- Previous examples have assumed equal link weights.
- However, 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 contains⁶

$$w(a \rightarrow b) = |a \rightarrow b| \times \log \left(\frac{t}{\sum_{x=0}^t |x \rightarrow b|} \right) \quad (2)$$

⁶[Mil07]

Dissertation Data

- Working the Wikipedia dataset (2008.01.03)
- Freely available, downloadable wikipedia content⁷
- Article text data (About 12 Gig)
- Page-Page and Page-Category SQL data dumps (About 2.5/3 Gigs)

Preparing the system by implementing existing relatedness measures.

⁷<http://download.wikimedia.org/>

Future Work

Presence of multiple, overlapping data sets provided by the Wikipedia community allows us to investigate a variety of new opportunities for relatedness metrics.

Specifically, my dissertation work will focus on opportunities in two new dimensions of relatedness:

- Temporal Information
- Cross-linguistic Information

Temporal

One of the major arguments against using Wikipedia as a knowledge base is the possibility of incorrect or misleading information in the dataset.

By integrating information across several iterations of Wikipedia datasets, we are able to minimize the impact of erroneous data.

Human-checked data source, and so the assumption is that incorrect information picked up in a previous data dump will be corrected by subsequent dumps.

Also allows for automatic updating of information content by adding additional data sets.

Cross-Linguistic

Wikipedia datasets in multiple languages with cross-linguistic links between pages.

By integrating information across several languages, we should be able to strengthen the relationship measures across all languages and help fill in knowledge gaps.

Conclusions

Word relatedness measures are useful tools in Natural Language Processing and Information Retrieval

Data comes a variety of sources and can be arranged in multiple ways

The overlapping nature of Wikipedia allows us to investigate knowledge over time and across languages.

Acknowledgments

- University of Dayton Department of Computer Science and Dr. Saverio Perugini
- Dr. Eric Fosler-Lussier, Dr. Chris Brew
- Dr. Donna K. Byron
- Tireless Wikipedia Volunteers and Contributors!

THANK YOU!



Alexander Budanitsky and Graeme Hirst.

Evaluating WordNet-based Measures of Lexical Semantic Relatedness.

Computational Linguistics, 32(1):13–47, 2006.



Marti A. Hearst.

Automatic Acquisition of Hyponyms from Large Text Corpora.

In *Proceedings of COLING-92*, 1992.



David Milne.

Computing Semantic Relatedness using Wikipedia Link Structure.

In *New Zealand Computer Science Research Student Conference (NZCSRSC'2007)*, 2007.



Michael Strube and Simone Paolo Ponzetto.

WikiRelate! Computing Semantic Relatedness Using Wikipedia.

In *AAAI*, 2006.



Peter D. Turney and Michael L. Littman.

Corpus-based Learning of Analogies and Semantic Relations.

Machine Learning, 2005.