

Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness

Siddharth Patwardhan

09/10/2003

Semantic Relatedness

- Is *needle* more related to *thread* than it is to *pie*?
- Humans agree on the relatedness of most word pairs.
- Miller and Charles (1991) suggest that relatedness is based on the overlap of contextual representation of words.

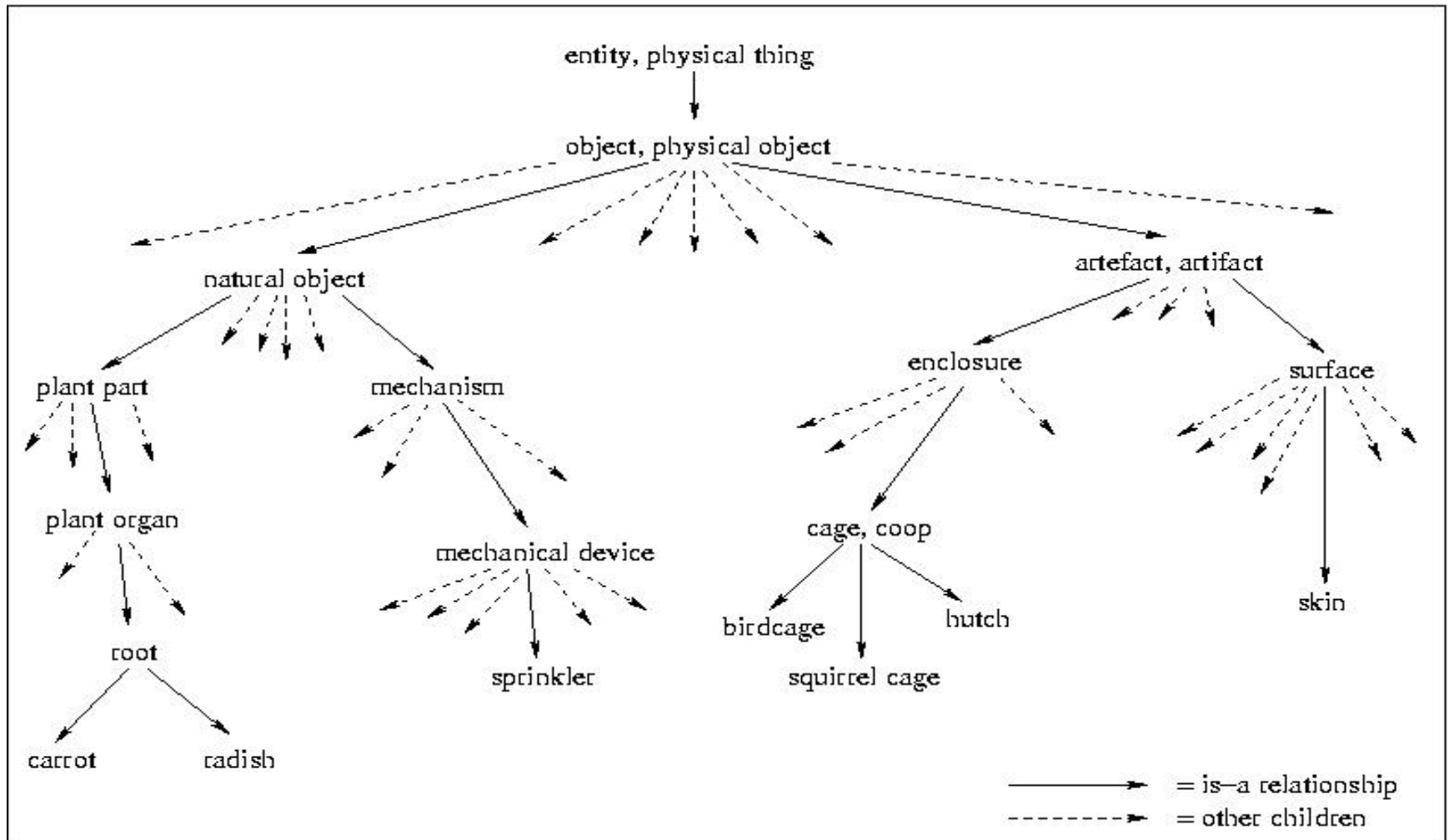
Measuring Relatedness

- **Automatic measures that attempt to imitate the human perception of relatedness of words and concepts.**
- **Number of measures of relatedness, based on WordNet and corpus data have been proposed.**
- **We compare various of measures of semantic relatedness.**

WordNet

- **Semantic network.**
- **Nodes represents real world concepts.**
- **Rich network of relationships between these concepts.**
- **Relationships such as “*car is a kind of vehicle*”, “*high opposite of low*”, etc exist.**
- **Node = Synonyms + Definition (gloss)**

WordNet – *Is a* Hierarchy



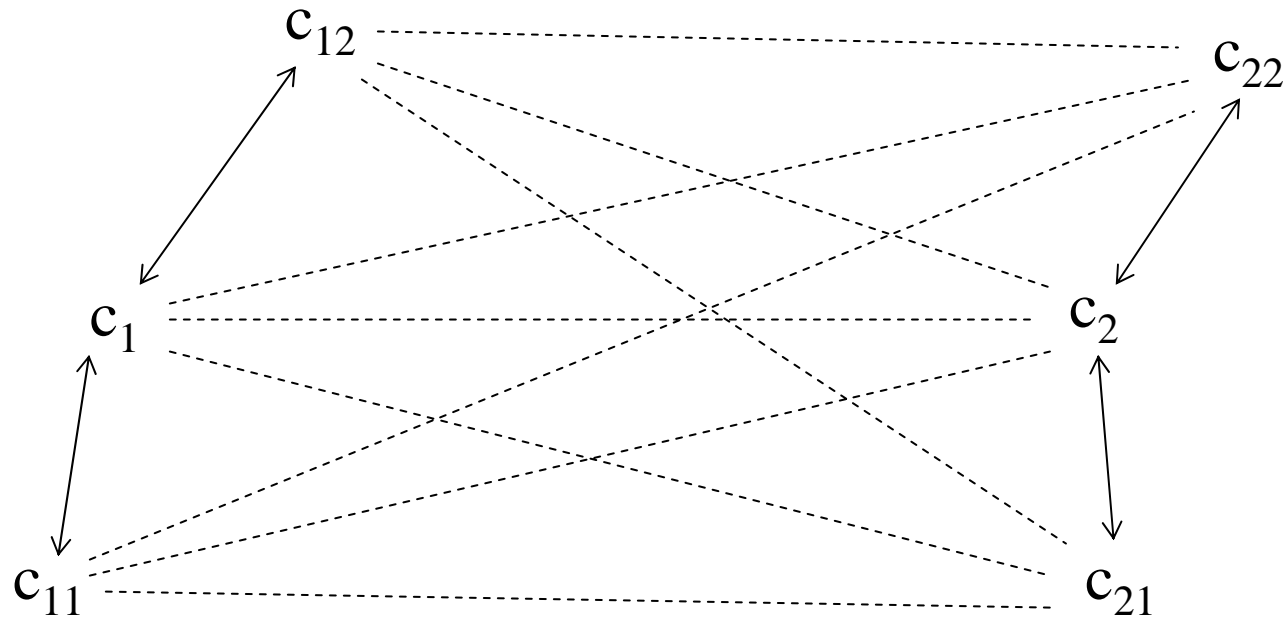
WordNet – Overview

- **Four parts of speech – nouns, verbs, adjectives and adverbs.**
- **~ 111,400 concepts.**
- **~ 13 types of relationships.**
- **9 noun *is a* hierarchies with an average depth of 13.**
- **628 verb *is a* hierarchies with an average depth of 2.**

The Adapted Lesk Algorithm

- **Performs *Word Sense Disambiguation* (Banerjee and Pedersen 2002).**
- **Uses the overlaps of dictionary definitions of word senses to determine sense of the target.**
- **Basic hypothesis – correct sense of the target word is most related to the senses of the words in its context.**
- **Overlaps measure the relatedness.**

Extended Gloss Overlaps



↔ WordNet relation

----- Gloss Overlap

Gloss Overlaps – Scoring

$$\begin{aligned} &1^2 + 2^2 \\ &= 1 + 4 \\ &= 5 \end{aligned}$$

The fruit of a coniferous tree used in salad.

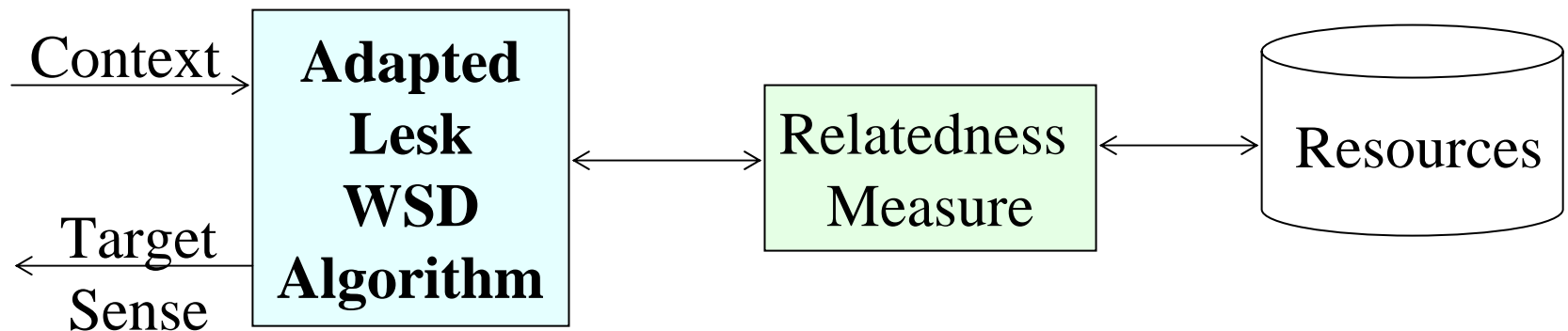
A fruit bearing coniferous tree that grows in hilly regions.

0

an artificial source of visible illumination

Our Extension of Adapted Lesk

- Use any measure in place of gloss overlaps and perform *WSD*.
- **Extended Gloss Overlaps** is a measure of semantic relatedness.



Leacock-Chodorow Measure (1998)

- **Based on simple edge counts in the *is a* hierarchy of WordNet.**
- **Deals with nouns only.**
- **The path length is scaled by the depth of the taxonomy.**

$$\textit{Relatedness}(c_1, c_2) = -\log(\textit{path_length} / 2D)$$

where c_1 and c_2 are the concepts and D is the depth of the taxonomy.

The Resnik Measure (1995)

- Deals with nouns only and is based on the *is a* hierarchy of WordNet
- Uses *Information Content* of concepts.
- Information Content of a concept indicates the specificity of the concept.

$$IC(\textit{concept}) = -\log(P(\textit{concept}))$$

- Probability of occurrence of concept in a corpus is calculated using its frequency in the corpus.

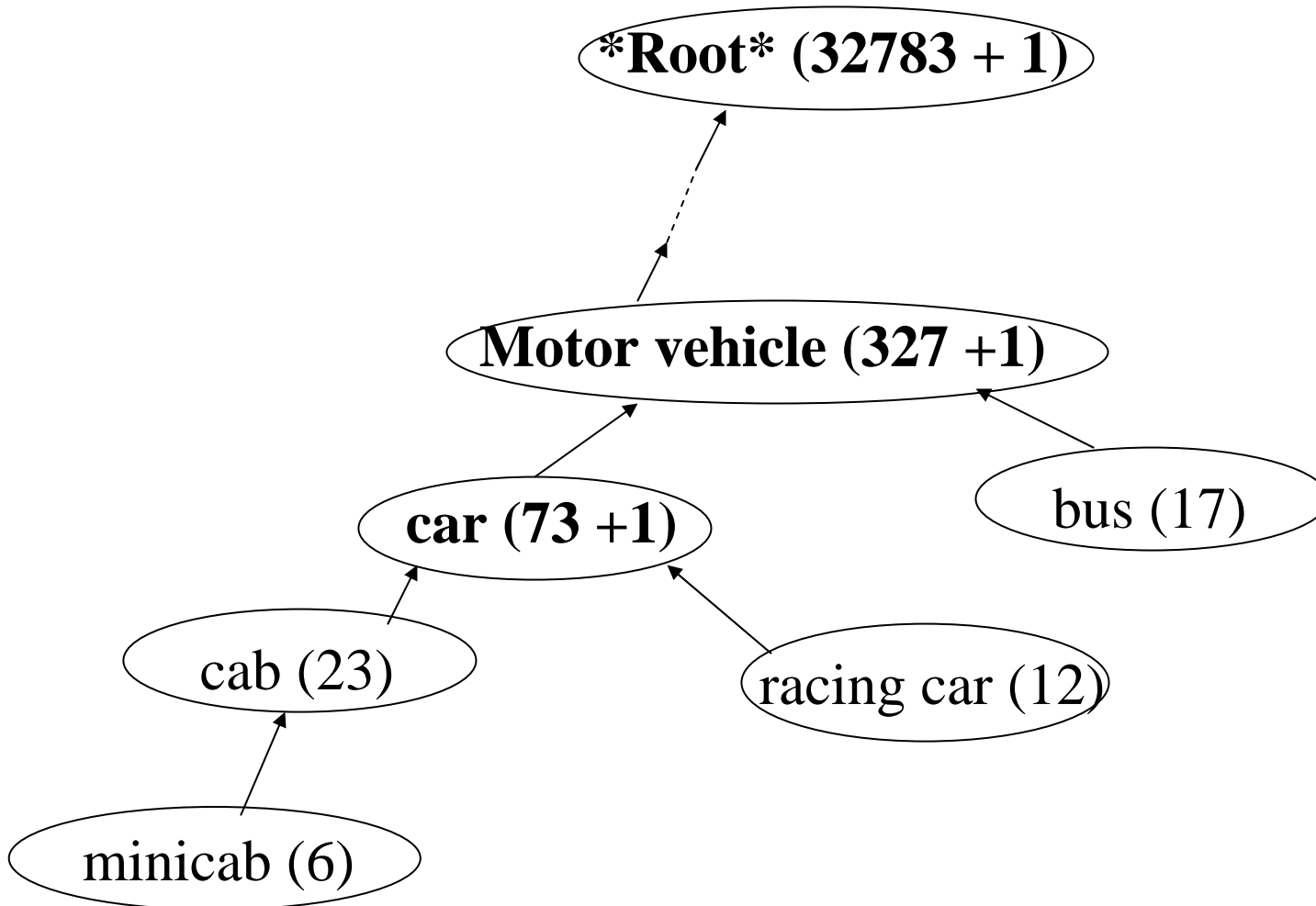
$$P(\textit{concept}) = \textit{freq}(\textit{concept})/\textit{freq}(\textit{root})$$

Information Content

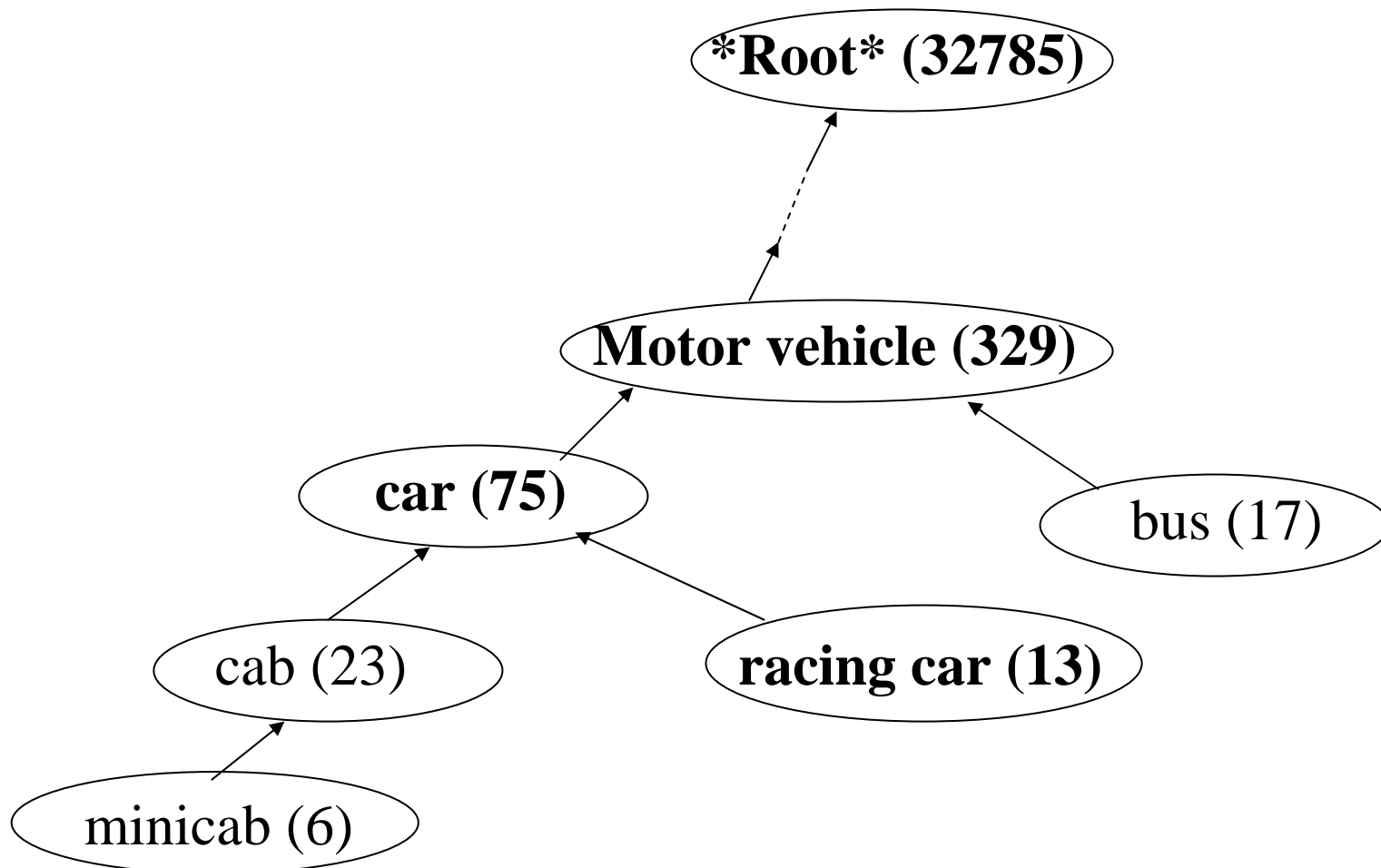
Counting the frequency of concepts:

- **Occurrence of a concept implies occurrence of all its subsuming concepts.**
- **Root node includes the count of all concepts in the hierarchy.**
- **Counting from non sense-tagged text raises some issues.**

Information Content



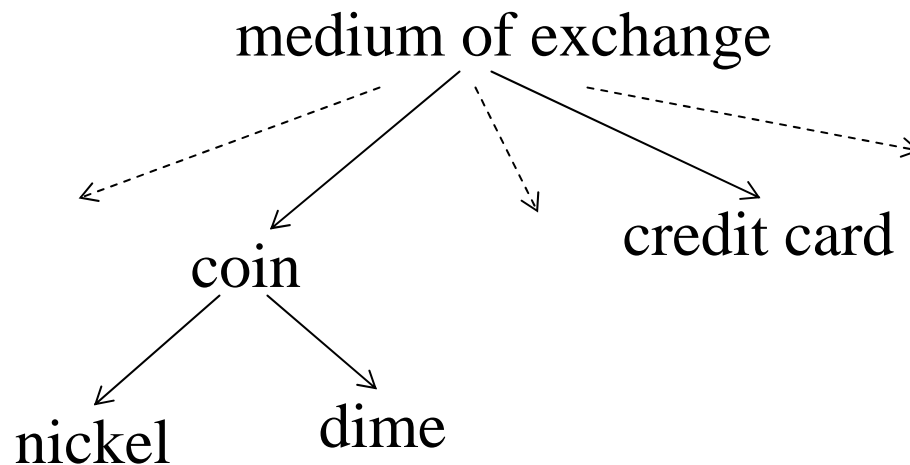
Information Content



Resnik Measure

$$\textit{Relatedness}(c_1, c_2) = IC(\textit{lcs}(c_1, c_2))$$

***lcs*(c_1, c_2) is the lowest concept in the *is a* hierarchy that subsumes both c_1 and c_2 .**



Jiang-Conrath Measure (1997)

- It is a measure of semantic *distance*:

$$\textit{distance} = IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1, c_2))$$

- We inverted it and used it as a measure of semantic relatedness:

$$\textit{Relatedness}(c_1, c_2) = 1 / \textit{distance}$$

- Also deals with nouns only.
- Has a lower bound of 0, no upper bound.

Lin Measure (1998)

$$\textit{Relatedness}(c_1, c_2) = \frac{2 \cdot \textit{IC}(\textit{lcs}(c_1, c_2))}{\textit{IC}(c_1) + \textit{IC}(c_2)}$$

- **Ranges between 0 and 1.**
- **If either $\textit{IC}(c_1)$ or $\textit{IC}(c_2)$ is zero, the relatedness is zero.**

Need for a New Measure

- **Preliminary results (Patwardhan, Banerjee and Pedersen 2003) show that Extended Gloss Overlaps does really well at WSD.**
- **Preliminary results also show that Jiang-Conrath does very well at WSD.**
- **One uses WordNet glosses, while the other is based on statistics derived from a corpus of text.**
- **Extended Gloss Overlaps – too exact.**

A Vector Measure

- **Represents glosses as multidimensional vectors of co-occurrence counts.**
- **Relatedness defined as the cosine of the *gloss vectors*.**
- **This alternate representation overcomes the “exactness” of the Extended Gloss Overlaps measure.**
- **Based on *context vectors* of Schütze (1998).**

Vector Measure – Word Space

- **We start by creating a “word space” – a list of words that will form the dimensions of the vector space.**
- **These words must be highly topical content words.**
- **We use a stop list and frequency cutoffs on the words in WordNet glosses to create this list (~54,000 words).**

Vector Measure – Word Vectors

- A word vector is created corresponding to every content word w in the WordNet glosses.
- The words of the Word Space are the dimensions of this vector.
- The vector contains the co-occurrence counts of words co-occurring with w in a large corpus.

coin = [*dollar* 35, *dime* 3, *movie* 0, *noon* 0, *cent* 56, *bank* 14]

Vector Measure – Gloss Vectors

- **Gloss vector for a concept is created by adding the word vectors for all the content words in its gloss.**

an artificial source of visible illumination

- **Gloss may be augmented by concatenating glosses of related concepts in WordNet (similar to Extended Gloss Overlaps).**

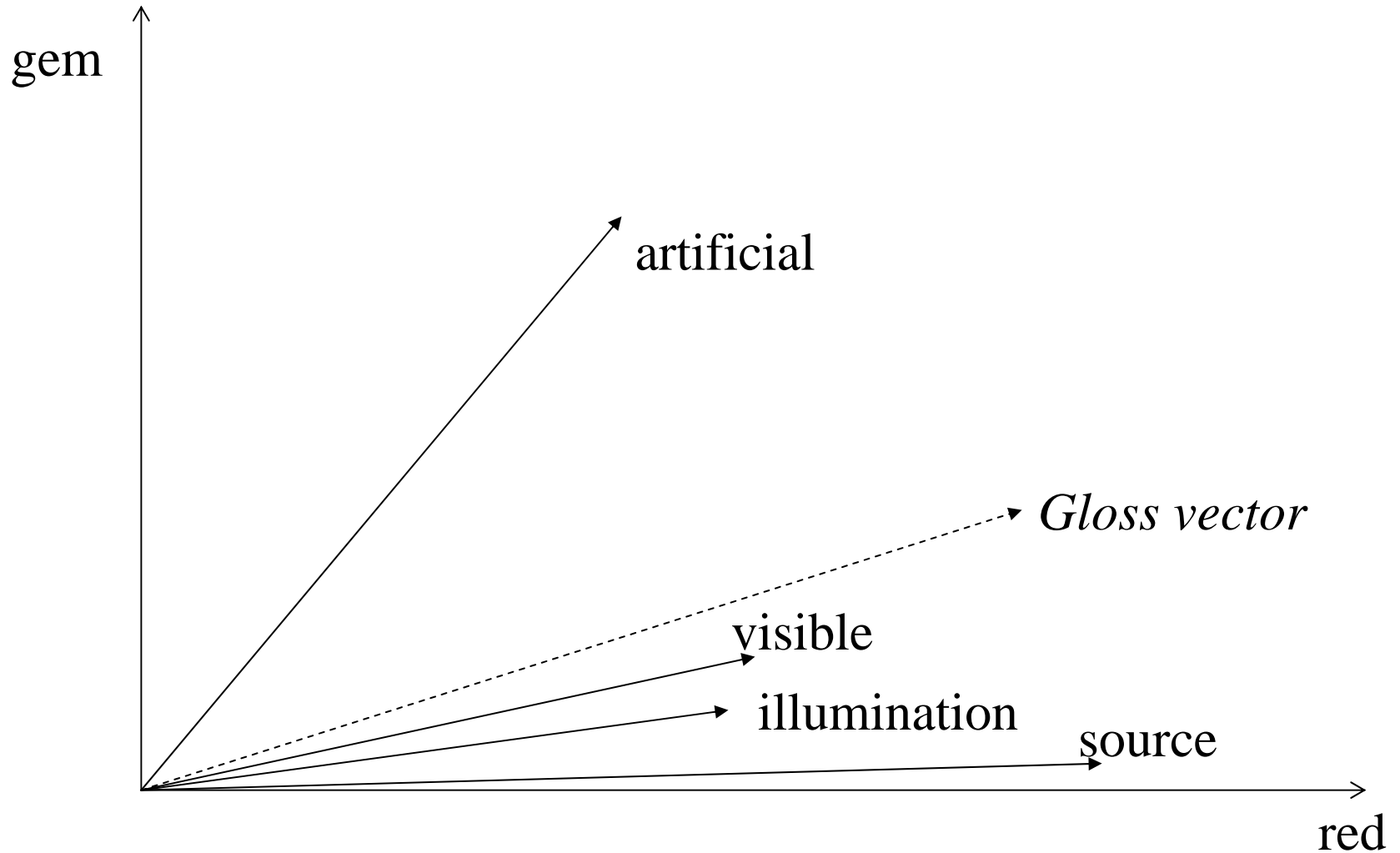
Gloss Vector – An Example

an artificial source of visible illumination

	eye	red	prison	gem	high
artificial	0	9	0	17	0
source	5	13	0	0	0
visible	23	14	1	3	0
illumination	18	10	0	4	2

Gloss vector	46	46	1	24	2
---------------------	----	----	---	----	---

Visualizing the Vectors



Vector Measure

$$\textit{Relatedness}(c_1, c_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{|\vec{v}_1| |\vec{v}_2|}$$

- **The values of this measure ranges between 0 and 1.**
- **The measure combines dictionary and corpus information to measure semantic relatedness.**

Comparison of the Measures to Human Relatedness

- We use 30 word pairs from Miller and Charles' experiment.**
- These are a subset of 65 pairs used by Rubenstein and Goodenough (1965) in a similar experiment.**
- We find the correlation of the ranking by the measures with human ranking of the 30 pairs.**

The Word Pairs

<i>Car</i>	<i>Magician</i>	<i>Tool</i>	<i>Cemetery</i>	<i>Coast</i>
<i>Automobile</i>	<i>Wizard</i>	<i>Implement</i>	<i>Woodland</i>	<i>Forest</i>
<i>Gem</i>	<i>Midday</i>	<i>Brother</i>	<i>Food</i>	<i>Lad</i>
<i>Jewel</i>	<i>Noon</i>	<i>Monk</i>	<i>Rooster</i>	<i>Wizard</i>
<i>Journey</i>	<i>Furnace</i>	<i>Lad</i>	<i>Coast</i>	<i>Chord</i>
<i>Voyage</i>	<i>Stove</i>	<i>Brother</i>	<i>Hill</i>	<i>Smile</i>
<i>Boy</i>	<i>Food</i>	<i>Crane</i>	<i>Forest</i>	<i>Glass</i>
<i>Lad</i>	<i>Fruit</i>	<i>Implement</i>	<i>Graveyard</i>	<i>Magician</i>
<i>Coast</i>	<i>Bird</i>	<i>Journey</i>	<i>Shore</i>	<i>Rooster</i>
<i>Shore</i>	<i>Cock</i>	<i>Car</i>	<i>Woodland</i>	<i>Voyage</i>
<i>Asylum</i>	<i>Bird</i>	<i>Monk</i>	<i>Monk</i>	<i>Noon</i>
<i>Madhouse</i>	<i>Crane</i>	<i>Oracle</i>	<i>Slave</i>	<i>String</i>

Human Relatedness Study

Measure	M&C	R&G
Vector	0.88	0.85
Jiang-Conrath	0.83	0.87
Extended Gloss Overlaps	0.81	0.81
Hirst-St.Onge	0.78	0.81
Resnik	0.77	0.78
Lin	0.76	0.80
Leacock-Chodorow	0.72	0.75

Variations in the Vector Measure

Word Vector Dimensions	Relations	
	All	Gloss
No frequency cutoffs	0.71	0.57
20,000 most frequent words	0.72	0.52
Words with frequencies 5 to 1,000	0.88	0.62

Variations on the Information Content based Measures

Source	Res	Lin	Jcn
SemCor (200,000)	0.71	0.70	0.72
Brown (1,000,000)	0.73	0.74	0.80
Treebank (1,000,000)	0.75	0.75	0.83
BNC (100,000,000)	0.75	0.75	0.81

Effect of Smoothing and Counting

Information Content from BNC	Res	Lin	Jcn
Our counting, no smoothing	0.75	0.75	0.81
Our counting, smoothing	0.75	0.75	0.81
Resnik counting, no smoothing	0.77	0.75	0.79
Resnik counting, smoothing	0.77	0.76	0.79

Application-oriented Comparison

- **Using Adapted Lesk as test-bed, determine accuracy on SENSEVAL-2 data, using the various measures.**
- **Window of context of 5 was used.**

WSD Results

Measure	Nouns Only	All POS
Jiang-Conrath	0.46	n/a
Ex. Gloss Overlaps	0.43	0.34
Lin	0.39	n/a
Vector	0.33	0.29
Hirst-St.Onge	0.33	0.23
Resnik	0.29	n/a
Leacock Chodorow	0.28	n/a

Conclusions

- **Modified the Adapted Lesk algorithm, to use any measure of relatedness for WSD.**
- **Introduced Gloss Overlaps as a measure of semantic relatedness.**
- **Created a new measure of relatedness, based on context vectors.**
- **Compared these to five other measures of semantic relatedness.**

Conclusions

- **Comparison was done with respect to human perception of relatedness.**
- **An application-oriented comparison of the measures was also done.**

Future Work

- **Determining ways to get better word vectors (frequency cutoffs).**
- **Dimensionality reduction techniques, such as Singular Value Decomposition (SVD).**
- **Use of semantic relatedness in Medical Informatics (on-going).**
- **Principled approach to context selection (WSD).**

WordNet::Similarity

<http://search.cpan.org/dist/WordNet-Similarity>