# Semantic Similarity Knowledge and its Applications

Diana Inkpen

School of Information Technology and Engineering

University of Ottawa
Canada

**KEPT 2007**

# Semantic relatedness of words

- Semantic relatedness refers to the degree to which two concepts or words are related.
- Humans are able to easily judge if a pair of words are related in some way.
- Examples
  - apple orange
  - apple toothbrush

# Semantic similarity of words

Relatedness:

- Synonyms
- Is-a relations (hypernyms)
- Part-of relations (meronyms)
- Context, situation (e.g. restaurant, menu)
- Antonyms (!)
- etc.

- Semantic similarity is a subset of semantic relatedness.

# Methods for computing semantic similarity of words

- Several types of methods for computing the similarity of two words (two main directions):
  - dictionary-based methods (using WordNet, Roget’s thesaurus, or other resources)
  - corpus-based methods (using statistics)
  - hybrid (combining the first two)

# Dictionary-based methods WordNet example (path length = 3)

apple (sense 1)

=> edible fruit

=> produce, green goods, green groceries, garden truck

=> food

=> solid

=> substance, matter

=> object, physical object

=> entity

orange (sense 1)

=> citrus, citrus fruit

=> edible fruit

=> produce, green goods, green groceries, …

# WordNet::Similarity Software Package

http://www.d.umn.edu/~tpederse/similarity.html

- Leacock & Chodorow (1998)
- Jiang & Conrath (1997)
- Resnik (1995)
- Lin (1998)
- Hirst & St-Onge (1998)
- Wu & Palmer (1994)
- extended gloss overlap, Banerjee and Pedersen (2003)
- context vectors, Patwardhan (2003)

# Roget's Thesaurus

301 FOOD

n.

fruit, soft fruit, berry, gooseberry, strawberry, raspberry,

loganberry, blackberry, tayberry, bilberry, mulberry;

currant, redcurrant, blackcurrant, whitecurrant;

stone fruit, apricot, peach, nectarine, plum, greengage, damson, cherry;

apple, crab apple, pippin, russet, pear;

citrus fruit, orange, grapefruit, pomelo, lemon, lime, tangerine, clementine, mandarin;

banana, pineapple, grape;

rhubarb;

date, fig;

….

# Similarity using Roget's Thesaurus (Jarmasz and Szpakowicz, 2003)

Path length - Distance:

- Length 0: same semicolon group. journey's end – terminus
- Length 2: same paragraph. devotion – abnormal affection
- Length 4: same part of speech. popular misconception – glaring error
- Length 6: same head. individual – lonely
- Length 8: same head group. finance – apply for a loan
- Length 10: same sub-section. life expectancy – herbalize
- Length 12: same section. Creirwy (love) – inspired
- Length 14: same class. translucid – blind eye
- Length 16: in the Thesaurus. nag – like greased lightning

# Corpus-based methods

Use frequencies of co-occurrence in corpora

- Vector-space
  - cosine method, overlap, etc.
  - latent semantic analysis
- Probabilistic
  - information radius
  - mutual information

Examples of large corpora: BNC, TREC data, Waterloo Multitext, LDC Gigabyte corpus, the Web

# Corpus-based measures (Demo)

http://clg.wlv.ac.uk/demos/similarity/

- Cosine
- Jaccard coefficient
- Dice coefficient
- Overlap coefficient
- L1 distance (City block distance)
- Euclidean distance (L2 distance)
- Information Radius (Jensen-Shannon divergence)
- Skew divergence
- Lin's Dependency-based Similarity Measure
  http://www.cs.ualberta.ca/~lindek/demos.htm

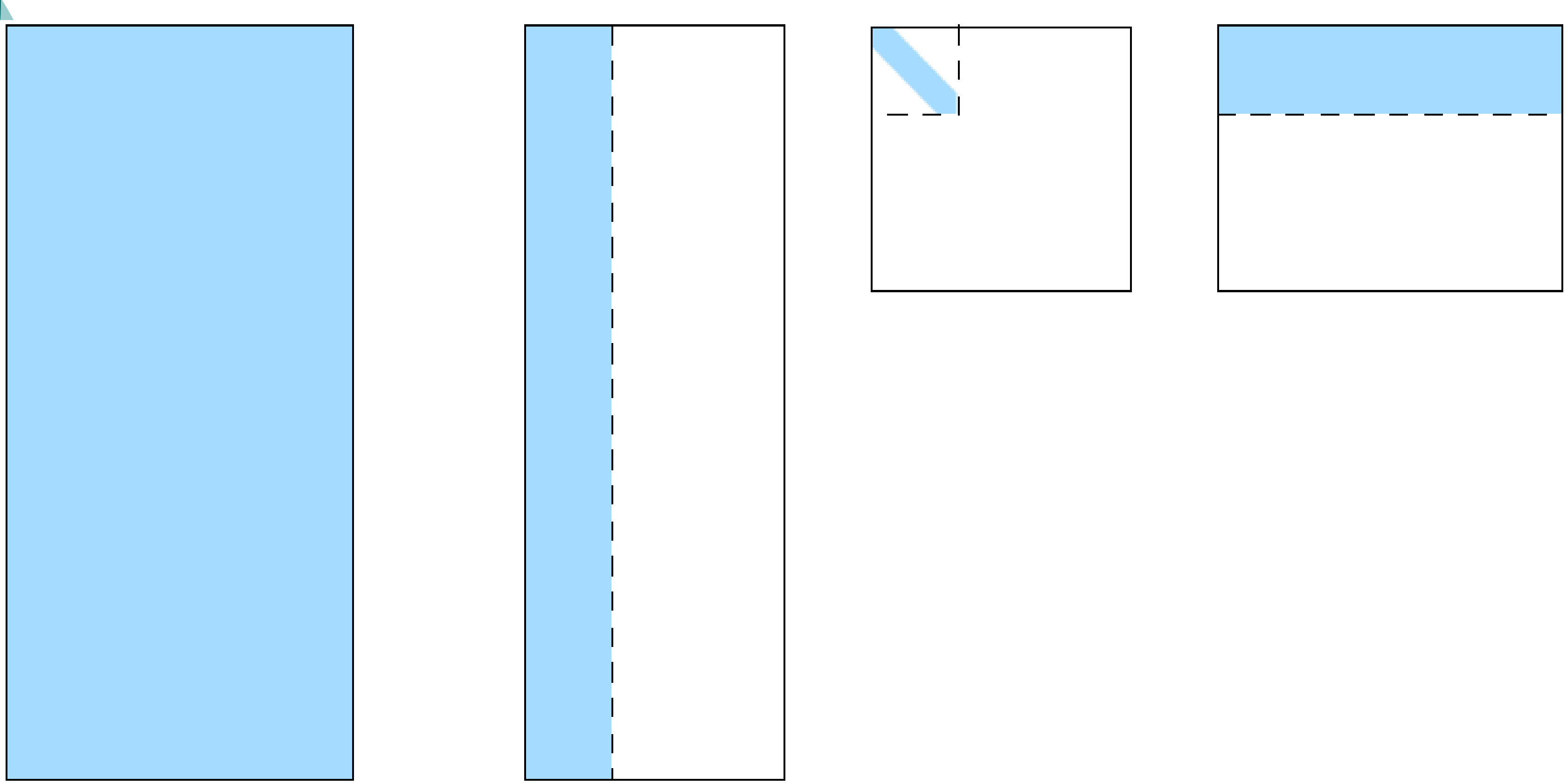
# Vector Space

$$\begin{pmatrix} & T_1 & T_2 & \ldots & T_t \\ D_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ D_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{pmatrix}$$

- Documents by words matrix
- Words by documents matrix
- Words by words matrix

$$A = \left(\begin{array}{l|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{cosmonaut} & 1 & 0 & 1 & 0 & 0 & 0 \\ \text{astronaut} & 0 & 1 & 0 & 0 & 0 & 0 \\ \text{moon} & 1 & 1 & 0 & 0 & 0 & 0 \\ \text{car} & 1 & 0 & 0 & 1 & 1 & 0 \\ \text{truck} & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right)$$

# Latent Semantic Analysis (LSA)
http://lsa.colorado.edu/ (Landauer & Dumais 1997)



$$\hat{A}_{t \times d} = T_{t \times k} \quad S_{k \times k} \quad {D_{d \times k}}^T$$

Produce a reduced matrix, fewer dimensions

# Pointwise Mutual Information

$$PMI(w_1, w_2) = \log\ P(w_1, w_2) / P(w_1)\ P(w_2)$$

$$PMI(w_1, w_2) = \log\ C(w_1, w_2)\ N / C(w_1)C(w_2)$$

N = number of words in the corpus

- use the Web as a corpus.
- use number of retrieved documents (hits returned by a search engine) to approximate word counts.

# Second-order co-occurrences SOC-PMI (Islam and Inkpen, 2006)

- Sort lists of important neighbor words of the two target words, using PMI.
- Take the shared neighbors and aggregate their PMI values (from the opposite list)

$W_1$ = car

get $\beta_1$ semantic neighbors with highest PMI

$W_2$ = automobile

get $\beta_2$ semantic neighbors with highest PMI

$$Sim(W_1, W_2) = \frac{f^{\beta}(W_1)}{\beta_1} + \frac{f^{\beta}(W_2)}{\beta_2}$$

# Hybrid methods

- WordNet plus small sense-annotated corpus (Semcor)
  - Jiang & Conrath (1997)
  - Resnik (1995)
  - Lin (1998)

- More investigation needed in combining methods, using large corpora.

# Evaluation

- Miller and Charles 30 noun pairs
- Rubenstein and Goodenough 65 noun pairs
  - gem, jewel, 3.84
  - coast, shore, 3.70
  - asylum, madhouse, 3.61
  - magician, wizard, 3.50
  - shore,woodland,0.63
  - glass,magician,0.11
- Task-based evaluation
- Retrieval of semantic neighbors (Weeds *et al.* 2004)

# Correlation with human judges

| Method Name | Miller and Charles 30 Noun pairs | Rubenstein and Goodenough 65 Noun pairs |
|---|---|---|
| Cosine (BNC) | 0.406 | 0.472 |
| SOC-PMI (BNC) | 0.764 | 0.729 |
| PMI (Web) | 0.759 | 0.746 |
| Leacock & Chodorow (WN) | 0.821 | 0.852 |
| Roget | 0.878 | 0.818 |

# Applications of word similarity

- solving TOEFL-style synonym questions
- detecting words that do not fit into their context
  - real-word error correction (Budanitsky & Hirst 2006)
  - detecting speech recognition errors
- synonym choice in context, for writing aid tools
  - intelligent thesaurus

# TOEFL questions

- 80 synonym test questions from the Test of English as a Foreign Language (TOEFL)
- 50 synonym test questions from a collection of English as a Second Language (ESL)
- Example

  The Smiths decided to go to Scotland for a short .......... trip
  They have already booked return bus tickets.
  - (a) travel
  - (b) trip
  - (c) voyage
  - (d) move

# TOEFL questions results
## (Islam and Inkpen, 2006)

| Method Name | Number of Correct Test Answers | Question/answer words not found | Percentage of Correct Answers |
|---|---|---|---|
| Roget's Sim. | 63 | 26 | 78.75% |
| SOC-PMI | 61 | 4 | 76.25% |
| PMI-IR * | 59 | 0 | 73.75% |
| LSA ** | 51.5 | 0 | 64.37% |
| Lin | 32 | 42 | 40.00% |

People averaged 64.5%, adequate for admission to universities
* Turney (2001)
** Landauer and Dumais (1997)

# Results on the 50 ESL questions

| Method name | Number of correct test answers | Question or answer words not found | Percentage of correct answers |
|---|---|---|---|
| Roget | 41 | 2 | 82% |
| SOC-PMI | 34 | 0 | 68% |
| PMI-IR | 33 | 0 | 66% |
| Lin | 32 | 8 | 64% |

# Detecting Speech Recognition Errors
(Inkpen and Désilets, 2005)

Manual transcript: Time now for our geography quiz today. We're traveling down the Volga river to a city that, like many Russian cities, has had several names. But this one stands out as the scene of an epic battle in world war two in which the Nazis were annihilated.

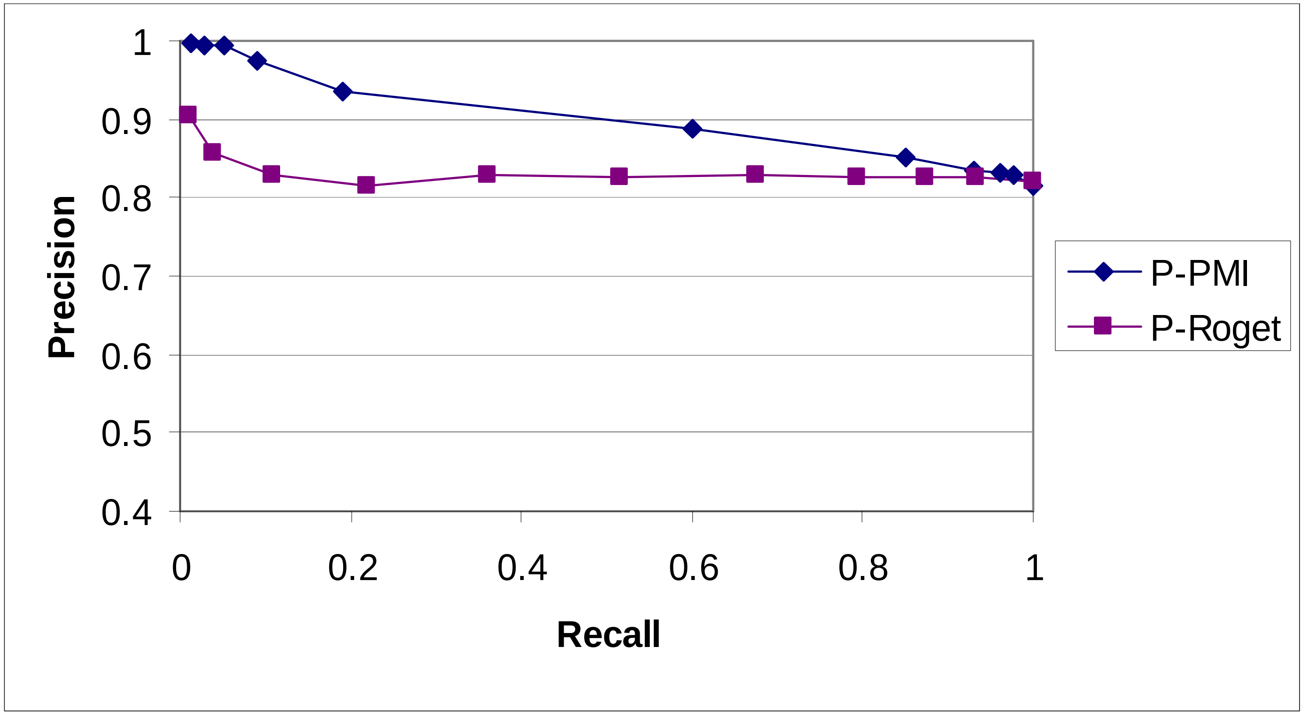
BBN transcript: time now for a geography was they were traveling down river to a city that like many russian cities has had several names but this one *stanza* is the scene of ethnic and national and world war two in which the nazis were nine *elated*

Detected outliers: *stanza, elated*

# Method - For each content word w in the automatic transcript:

1. Compute the **neighborhood** N(w), i.e. the set of content words that occur "close" to w in the transcript (include w).
2. Compute **pair-wise semantic similarity** scores $S(w_i,w_j)$ between all pairs of words $w_i \neq w_j$ in N(w), using a semantic similarity measure.
3. Compute the semantic coherence $SC(w_i)$ by "aggregating" the pair-wise semantic similarities $S(w_i, w_j)$ of $w_i$ with all its neighbors $w_j \neq w_i$ in N(w).
4. Let $SC_{avg}$ be the average of $SC(w_i)$ over all $w_i$ in the neighborhood N(w).
5. Label w as a recognition errors if $SC(w) < K\ SC_{avg}$.

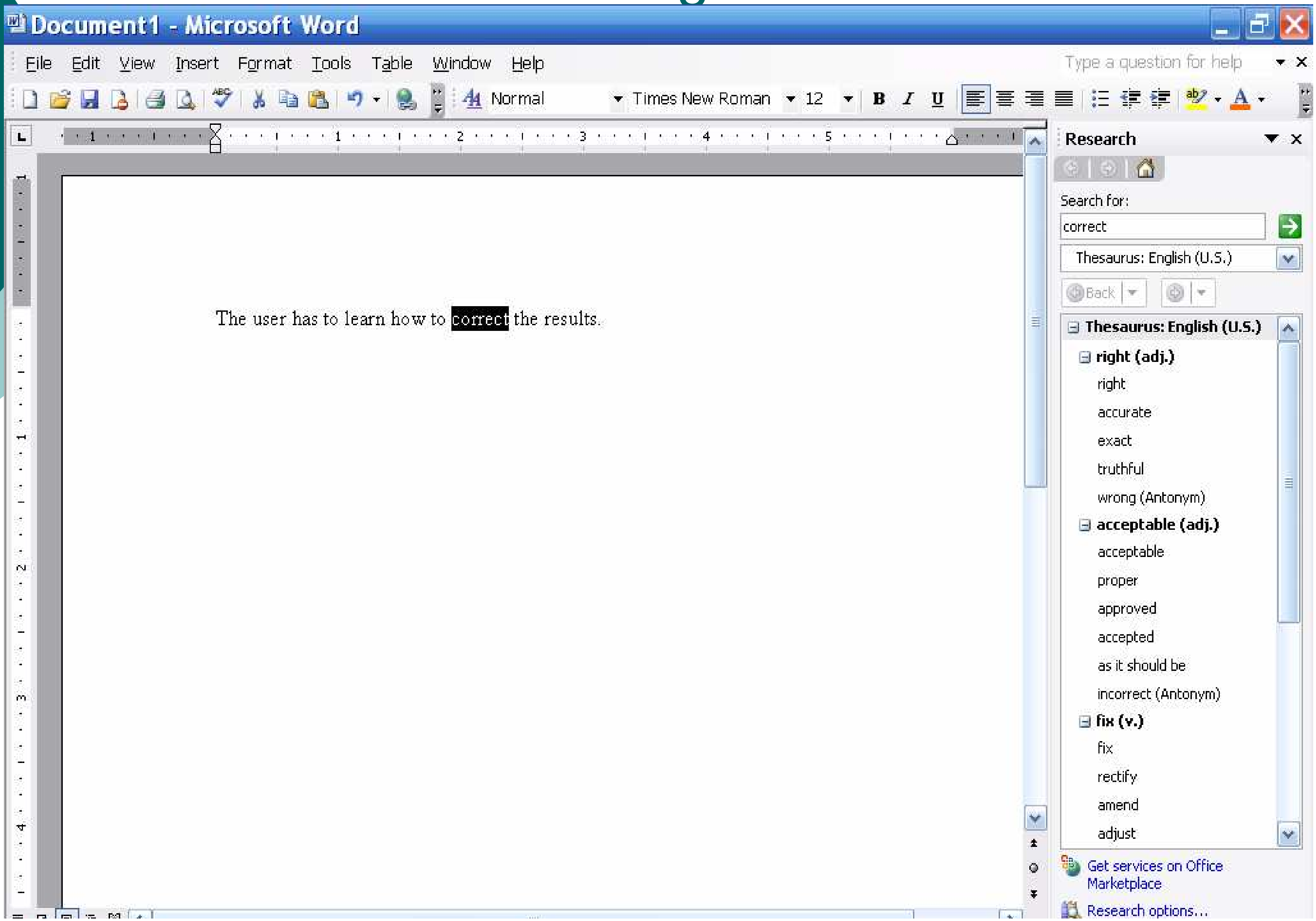
# Detecting Speech Recognition Errors (Roget vs. PMI)



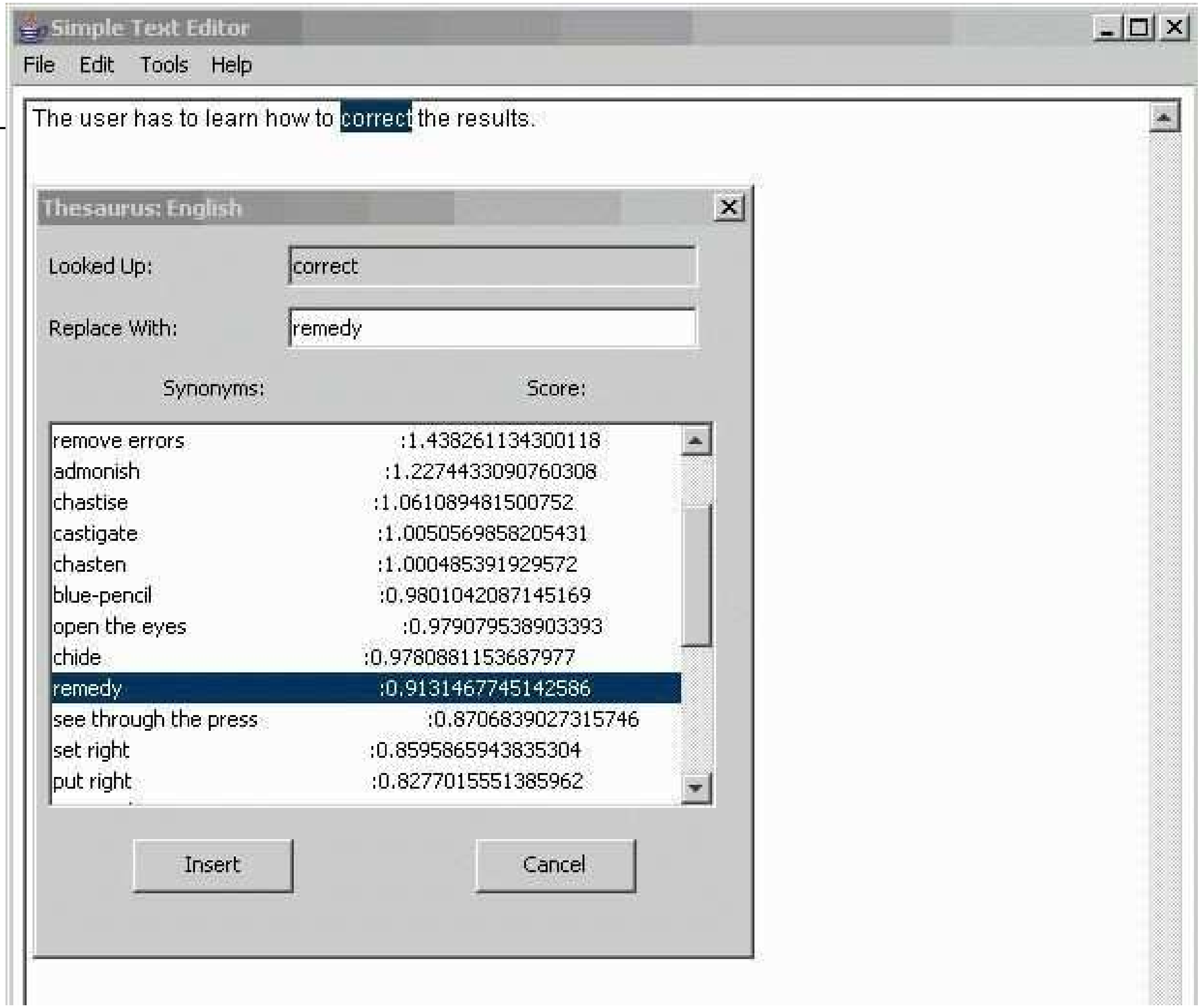

Data: 100 stories from TDT, plus manual transcripts.

Variation of threshold k determines confidence level for identifying errors.

# Thesaurus as Writing Aid



# Intelligent Thesaurus



# Intelligent Thesaurus (Inkpen, 2007) Training and Test Data

Sentence: This could be improved by more detailed consideration of the processes of ......... propagation inherent in digitizing procedures.

error

Solution set: mistake, blooper, blunder, boner, contretemps, error, faux pas, goof, slip, solecism

Sentence: The effort required has had an unhappy effect upon his prose, on his ability to make the discriminations the complex …………. demands.

job

Solution set: job, task, chore

# Semantic coherence of a word with its context

- PMI, using as corpus 1 terabyte of Web data - the Waterloo Multitext system (Clarke and Terra 2003).
- Window of k words before the gap and k words after the gap (best k=2)
- Counts of two words in window of size q in the corpus (best q = 3)
- Number of word pairs or number of documents (words vs. docs)

$s = \ldots w_1 \ldots w_k$ *Gap* $w_{k+1} \ldots w_{2k} \ldots$

$\mathrm{Score}(NS_i, s) = \boldsymbol{\Sigma}_{j=1,k}\, \mathrm{PMI}(NS_i,w_j) + \boldsymbol{\Sigma}_{j=k+1,2k}\, \mathrm{PMI}(NS_i,w_j)$

# Results for the intelligent thesaurus

| Test set | Baseline most freq. syn. | Edmonds' method, 1997 | Accuracy first choice | Accuracy first two choices |
|---|---|---|---|---|
| Data set 1 (7gr)<br>Syns: WordNet<br>Sentences: WSJ | 44.8% | 55% | 66.0% | 88.5% |
| Data set 2 (11gr)<br>Syns: CTRW<br>Sentences: BNC | 57.0% | – | 76.5% | 87.5% |

# Similarity of two short texts

- A method for computing the similarity of two texts, based on the similarities of their words.
- Applications of text similarity knowledge:
  - designing exercises for second language-learning
  - acquisition of domain-specific corpora
  - information retrieval
  - text categorization

# Text similarity method (Islam and Inkpen, 2007 subm.)

- Use corpus-based similarity for two words (SOC-PMI)
- Use string similarity (longest common subsequence)
- Select a word from S1 and a word from S2 that have highest similarity, iterate for the rest of the texts, aggregate scores.

# Evaluation of text similarity

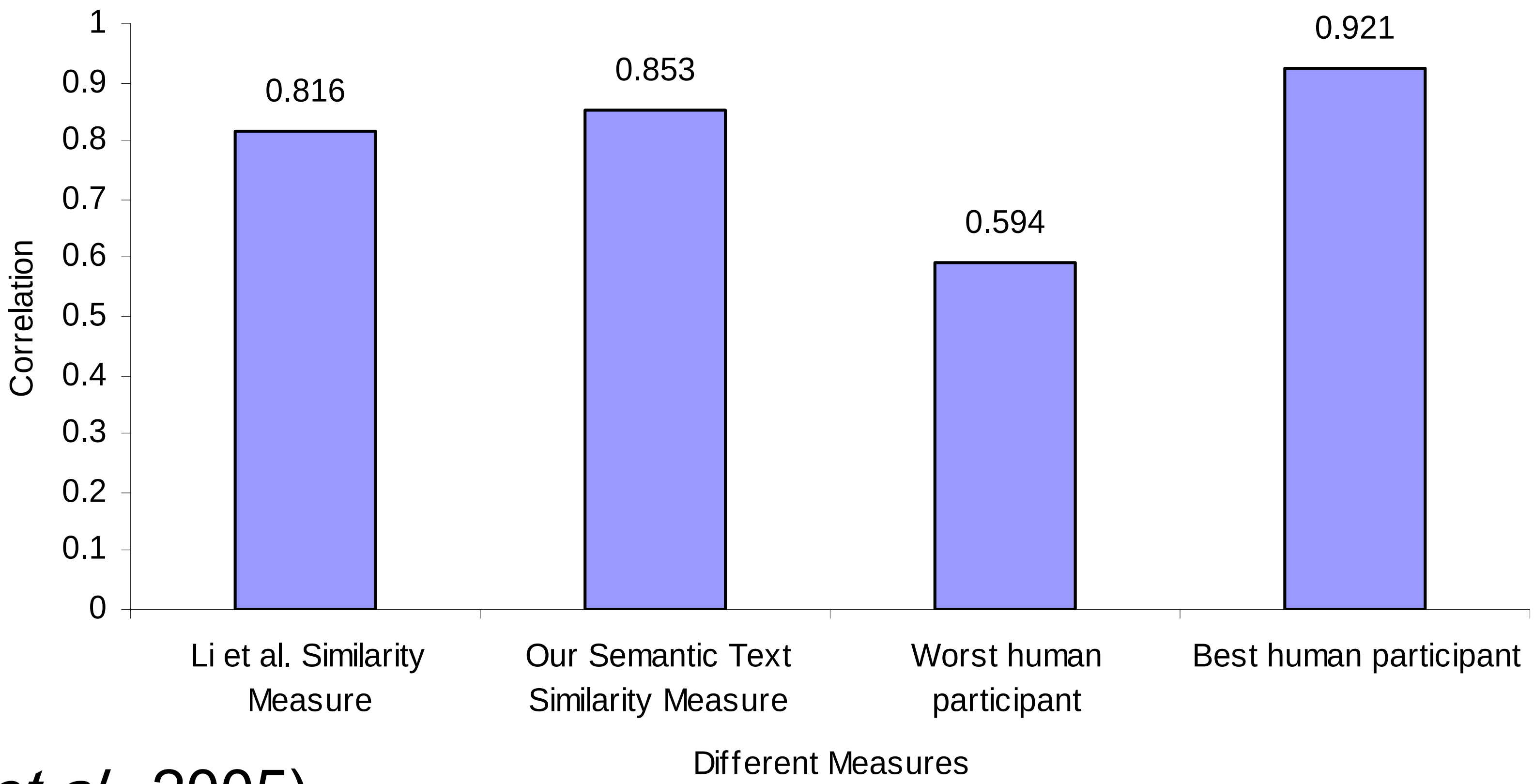
Test data:

- 30 sentence pairs (Li *et al.*, 2005)
- Microsoft paraphrase corpus

Example:

- Fighting erupted after four North Korean journalists confronted a dozen South Korean activists protesting human rights abuses in the North outside the main media centre.
- Trouble flared when at least four North Korean reporters rushed from the Taegu media centre to confront a dozen activists protesting against human rights abuses in the North.

# Correlation with human judges on the 30 sentence pairs




(Li *et al.*, 2005)

Method based on a lexical co-occurrence network

# Results on the MS Paraphrase corpus

| Metric | Accuracy | Precision | Recall | F-measure |
|---|---|---|---|---|
| Random | 51.3 | 68.3 | 50.0 | 57.8 |
| Vector-based | 65.4 | 71.6 | 79.5 | 75.3 |
| J & C | 69.3 | 72.2 | 87.1 | 79.0 |
| L & C | 69.5 | 72.4 | 87.0 | 79.0 |
| Lesk | 69.3 | 72.4 | 86.6 | 78.9 |
| Lin | 69.3 | 71.6 | 88.7 | 79.2 |
| W & P | 69.0 | 70.2 | 92.1 | 80.0 |
| Resnik | 69.0 | 69.0 | 96.4 | 80.4 |
| Combined(S) * | 71.5 | 72.3 | 92.5 | 81.2 |
| Combined(U) * | 70.3 | 69.6 | **97.7** | **81.3** |
| PMI-IR | 69.9 | 70.2 | 95.2 | 81.0 |
| LSA | 68.4 | 69.7 | 95.2 | 80.5 |
| **STS** | **72.6** | **74.7** | 89.1 | **81.3** |

* Mihalcea *et al.* (2006)

# Cross-language similarity

- Cross-language similarity of two words:
  - take maximum between $W_2$ and all possible translations of $W_1$

| Example | French | English |
| --- | --- | --- |
| | pomme = apple | orange |
| | = potato | |
| | = head | |

- Cross-language similarity of two texts – based on similarity between words.

# Conclusion

- Methods for word similarity
- Evaluation
- Applications
- Methods for text similarity

# Future work

- Combine word similarity methods
- Second-order co-occurrences in Web corpora (Google 5-gram corpus)
- Cross-language similarity

# References

Banerjee S. and Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. IJCAI 2003

Budanitsky A. and Hirst G. Evaluating WordNet-based measures of semantic distance. Computational Linguistics, 32(1), 2006.

Edmonds P. Choosing the word most typical in context using a lexical co-occurrence network, ACL 1997

Hirst G. and St-Onge D. Lexical Chains as representations of context for the detection and correction of malapropisms. In WordNet An electronic Database, 1998

Inkpen D. Near-synonym choice in an Intelligent Thesaurus, HLT-NAACL 2007

Inkpen D. and Désilets A. Semantic similarity for detecting recognition errors in automatic speech transcripts. EMNLP 2005

Islam A. and Inkpen D. Semantic similarity of short texts, submitted 2007

Islam A. and Inkpen D. Second order co-occurrence PMI for determining the semantic similarity of words, LREC 2006

Jarmasz M. and Szpakowicz S. Roget's thesaurus and semantic similarity, RANLP 2003

Jiang J. and Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. COLING 1997

# References


Landauer T.K. and Dumais S.T. A Solution to Plato’s problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 1997

Leacock C. and Chodorow M. Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database, 1998

Li Y., McLean D., Bandar Z., O’Shea J., and Crockett K. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowledge and Data Eng. 18:8, 2006

Lin D. An information-theoretic definition of similarity. ICML 1998

Mihalcea R., Corley, C. Strapparava, C. Corpus-based and knowledge-based measures of text semantic similarity. AAAI 2006

Patwardhan S. Incorporating dictionary and corpus information into a vector measure of semantic relatedness. MSc Thesis, 2003.

Resnik P. Semantic similarity in a taxonomy: An information-based ,easure and its applications to problems of ambiguity in natural language. JAIR 11, 1999

Weeds J., Weir D. and McCarthy D. Characterising measures of lexical distributional similarity. COLING 2004

Wu Z. and Palmer M. Verb semantics and lexical selection. ACL 1994

Turney P.D. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. ECML 2001