Experiments on the difference between semantic similarity and relatedness

Peter Kolb
University of Potsdam

NODALIDA '09
Odense, 15th May 2009
overview

• semantic similarity vs. semantic relatedness
• relevance of the distinction for NLP applications
• distributional semantics
• DISCO
• evaluation data
• experiments
• conclusion
• future work
Semantic similarity

- numeric, continuous measure of similarity between pairs of words
- used for ontology learning, information retrieval, word sense disambiguation, recognition of textual entailment, and many more

Semantic similarity vs. semantic relatedness

- similarity: lexical items have similar meaning
- usually defined via synonymy and hyponymy
- similar words can be substituted for each other
- e.g.: palm - tree, doctor - surgeon, mention - remark
**semantic relatedness**

- broader concept: similar → related
- words connected by any kind of lexical or functional association
- meronymy, antonymy, is-a-way-of-doing, is-a-symptom-of, ...
- words from the same semantic field (doctor - hospital)
- words associated by common co-occurrence
- do not have to be substitutable for each other in context
- do not have to belong to the same part-of-speech category
relevance of the distinction

• some NLP applications work better with the either or the other type of similarity relation

• e.g. detecting term variants for search term expansion:
  - doctor: physician physican dotcor doctors medic
  - doctor: hospital nurse illness

• same for paraphrasing, thesaurus generation, ...

• for word sense disambiguation: palm - coconut is as useful as palm - tree
distributional semantics

- compute word similarities from distribution of words in text corpora
- words with similar meaning occur in similar contexts
- relation between words, not word senses: rock $\rightarrow$ jazz pop stone sand ...
- many different methods have been proposed
- LSA (Landauer & Dumais 1997): build term-document matrix, apply algebraic dimension reduction technique, compare terms (rows) of the reduced matrix
- PMI-IR (Turney 2001): similarity $:= \text{mutual information value of co-occurrence in web pages}$
distributional semantics (2)

- few evaluations regarding influence of generation method on the type of the resulting word space:
  - stricter (syntactic) contexts lead to “tighter” similarities
  - indirect co-occurrence results in tighter similarities than direct co-occurrence
  - severe dimensionality reduction (by Random Indexing or frequency cutoffs) leads to retrieval of more loosely related words
- what about other reduction techniques and parameters?
DISCO

- tokenize, eliminate function words
- count co-occurrences in ±3 word window, record position within window:

| -3 | -2 | -1 | +1 | +2 | +3 |
|----|----|----|----|----|----|
| the| nuts| provide| palm| oil| while| the |

- `<palm, -2, nuts>, <palm, +1, oil>, ...`
- co-occurrence matrix: words x features (v x f · r)
**DISCO (2)**

- transform co-occurrence frequencies into weights using formula based on mutual information
- compare all words via their feature sets using measure of vector similarity
  - distributionally similar words (indirect, second-order co-occurrences)
    - e.g.: *bread - bake, eat, crispy*
    - *cake - bake, eat, crispy*
  - *bread* and *cake* will be similar (even if they didn’t co-occur themselves)
DISCO (3)

• example: *palm* - *palms coconut olive pine citrus oak mango cocoa banana bananas trees fingers* ...

• now view this list of distributionally similar words as feature vector describing *palm*

• compare two words based on their sets of distributionally similar words

• makes use of higher-order co-occurrences

• this is also achieved in LSA by SVD [Kontostathis & Pottenger 2006]

  ➢ two DISCO measures: DISCO1 and DISCO2
evaluation data

- semantic relatedness [Finkelstein et al. 2001]: 353 word pairs with averaged relatedness rating by 16 subjects
- semantic similarity: WordNet::Similarity [Pedersen et al. 2004]
  - Perl module based on WordNet
  - implements three measures of semantic relatedness: Hirst-St.Onge (hso), Lesk, vector pairs (vp)
  - six measures of semantic similarity: Jiang and Conrath (jcn), Leacock and Chodorow (lch), Lin, path length, Resnik (res), Wu and Palmer (wup)
  - similarity measures exclusively based on WordNet’s IS-A noun hierarchy and synsets
exp1: relatedness

- measure correlation of systems with Finkelstein’s relatedness gold standard:
  - DISCO1 and DISCO2 based on 267 million tokens from Wikipedia
  - PMI-IR using AltaVista
  - LSA using http://lsa.colorado.edu
  - WordNet::Similarity
exp1: relatedness

| vectors | LSA  | PMI-IR | DISCO1 | DISCO2 |
|---------|------|--------|--------|--------|
| 0.41    | 0.56 | 0.63   | 0.39   | 0.51   |

| hso    | lesk | vp    | jcn   | lch   | lin   | path  | res   | wup   |
|--------|------|-------|-------|-------|-------|-------|-------|-------|
| 0.35   | 0.21 | 0.39  | 0.23  | 0.35  | 0.30  | 0.38  | 0.36  | 0.30  |

- DISCO2 significantly better than DISCO1 (Fisher’s z-score transformation, $\alpha = 0.05$)
- higher-order co-occurrences can partly substitute for SVD
- PMI-IR significantly better than DISCO2
- WordNet-based measures perform quite poorly
**exp2: similarity**

- correlation of systems with WordNet::Similarity

|       | jcn | lch | lin | path | res | wup | avg. |
|-------|-----|-----|-----|------|-----|-----|------|
| PMI-IR| 0.14| 0.12| 0.06| 0.15 | 0.22| 0.11| 0.13 |
| LSA   | 0.16| 0.26| 0.21| 0.29 | 0.28| 0.22| 0.24 |
| DISCO1| 0.38| 0.39| 0.33| 0.45 | 0.43| 0.33| 0.38 |
| DISCO2| 0.15| 0.40| 0.39| 0.35 | 0.44| 0.40| 0.36 |
exp.2: similarity

- DISCO2 not better than DISCO1
- LSA worse than DISCO
- higher-order co-occurrences/SVD do not help in computing semantic similarity?
- PMI-IR based on direct co-occurrence: bad for computing similarity
exp.3: DISCO1 parameters

- BNC instead of Wikipedia (only 40% of size)
- correlation of DISCO1 with relatedness gold standard (finkel353) drops from 0.39 to 0.34
- correlation with similarity gold standard (res) still at 0.43
  - measures of relatedness profit from more input data
  - (difference is not significant at $\alpha = 0.05$)
exp. 3-2: window position

- DISCO1 with window position features vs. pure bag-of-words window

|          | finkel353 | res |
|----------|-----------|-----|
| DISCO1 WPT | 0.34      | 0.43|
| DISCO1 bag-of-words | 0.32      | 0.12|
exp.3-3: lemmatization

- lemmatize the corpus with the Tree Tagger

|                | finkel353 | res |
|----------------|-----------|-----|
| DISCO1 WPT     | 0.34      | 0.43|
| DISCO1 WPT     | 0.36      | 0.41|
| lemmatized     |           |     |
exp.3-4: dependency triples

- dependency-parse the corpus with Minipar
- syntactic dependency triples

|                | finkel353 | res   |
|----------------|-----------|-------|
| DISCO1 WPT     | 0.34      | 0.43  |
| DISCO1 WPT     | 0.36      | 0.39  |
| dependency     |           |       |
exp.3-5: frequency cutoff

- matrix size $\nu \times f \cdot r$
- reduce size of $f$ by frequency

|       | f    | finkel | res  |
|-------|------|--------|------|
| 101,000 | 0.34 | 0.43   |
| 50,000  | 0.37 | 0.43   |
| 20,000  | 0.40 | 0.45   |
| 10,000  | 0.41 | 0.46   |
| 5,000   | 0.40 | 0.43   |
| 1,000   | 0.38 | 0.43   |
| 500     | 0.36 | 0.33   |
exp.3-6: SVD

- apply SVD to reduced DISCO1-10K matrix
- reduce from $v \times 10,000 \cdot r$ to $v \times 300$

|        | finkel | jcn | lch | lin | path | res | wup |
|--------|--------|-----|-----|-----|------|-----|-----|
| DISCO1-10K | 0.41   | 0.62| 0.52| 0.50| 0.52 | 0.46| 0.47|
| DISCO1-10K-SVD | 0.55   | 0.46| 0.37| 0.41| 0.39 | 0.38| 0.35|
conclusion

- there's a difference between semantic similarity and relatedness
- difference has practical consequences for NLP applications
- techniques for word space construction produce word spaces that are biased towards the one or the other type of similarity
Future work

- Which kind of word space should be applied to a given problem?
- For which applications is the similarity-relatedness distinction relevant?
- How to improve semantic similarity?