Lecture 6: Semantic Spaces and Similarity

Lexical Semantics and Discourse Processing
MPhil in Advanced Computer Science

Simone Teufel

Natural Language and Information Processing (NLIP) Group
UNIVERSITY OF CAMBRIDGE
Simone.Teufel@cl.cam.ac.uk

February 4, 2011

Reading:

- Jurafsky and Martin, chapters 20.7 (Word Similarity: Distributional Methods);
- Dekang Lin (1998), Automatic Retrieval and Clustering of Similar Words, ACL-98.

Measuring Similarity between Words

- Automatically determine how "similar" two words are. But how to define similarity?
- Generally accepted that there are at least two dimensions:
  - Word **Relatedness**: includes relations such as antonymy (*car–petrol*)
  - Word **Similarity**: near-synonyms; substitutable in context (*car–bicycle*)
- Human intuitions about word-pairs and how similar they are exist and are replicable:
  - Rubenstein and Goodenough (1965) – 65 word pairs
  - Miller and Charles (1991) – 30 word pairs
- Apart from the Distributional Measures treated here, there are also Thesaurus-based Methods (cf. JM chapter 20.6)

Vector Space

- A word is represented as a bag-of-word feature vector.
- The features represent the N words of a lexicon that occur in a window context.
- In IR, the "context" is always exactly one document.
- Each word is a point high-dimensional vector space
- We can now compare words with each other in vector space, but also words with sets of words (e.g., documents vs. queries in IR).

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Representing the meaning of a word in VS

In a realistic situation:

1. Choose context window size (or use documents)
2. Choose dimensionality of vector: what counts as a term?
3. Choose a type of feature: co-occurrence, lexicalised grammatical relation ...
4. Choose how each cell in the vector is to be weighted:
   - presence or absence (binary)
   - term frequency in contexts or document
   - TF*IDF (cf. lecture 3)
   - **association measures**
5. Choose a **proximity measure**

Feature type: lexicalised grammatical relations (Lin 1998)

subj-of, absorb	1
subj-of, adapt	1
subj-of, behave	1
...	
pobj-of, inside	16
pobj-of, into	30
...	
nmod-of, abnormality	3
nmod-of, anemia	8
nmod-of, architecture	1
...	
obj-of, attack	6
obj-of, call	11
obj-of, come from	3
obj-of, decorate	2
...	
nmod, bacteria	3
nmod, body	2
nmod, bone marrow	2

Context word: **cell**; frequency counts from 64-Million word corpus.

Association measures: weighting co-occurrences

How surprised should we be to see this feature associated with the target word?

- **Pointwise Mutual Information** (Fano, 1961):

$$assoc_{PMI}(w, f) = log_2 \frac{P(w, f)}{P(w)P(f)}$$

- **Lin Association Measure**:

$$assoc_{Lin}(w, f) = log_2 \frac{P(w, f)}{P(w)P(r|w)P(w'|w)}$$

  r: grammatical function; w': grammatically related word.
- **t-test**:

$$assoc_{t-test}(w, f) = \frac{P(w, f) - P(w)P(f)}{\sqrt{P(f)P(w)}}$$

Distance metrics

- **Manhattan Distance**: (Levenshtein Distance, L1 norm)

$$distance_{manhattan}(\vec{x}, \vec{y}) = \sum_{i=1}^{N} |x_i - y_i|$$

- **Euclidean Distance**: (L2 norm)

$$distance_{euclidean}(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

Similarity Metrics

- **Cosine**: (normalisation by vector lengths)

$$sim_{cosine}(\vec{x},\vec{y}) = \frac{\vec{x}\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

- **Jaccard** (Grefenstette, 1994):

$$sim_{jacc}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} max(x_i, y_i)}$$

- **Dice Coefficient** (Curran, 2003):

$$sim_{dice}(\vec{x},\vec{y}) = \frac{2\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} (x_i + y_i)}$$

Information-Theoretic Association Measures

How similar two words are depends on how much their distributions diverge from each other.

- **Kuhlback-Leibler Divergence**

$$D(P||Q) = \sum_x P(x) log \frac{P(x)}{Q(x)}$$

Unfortunately, KL is undefined when $Q(x) = 0$ and $P(x) \neq 0$, which is frequent. Therefore:

- **Jensen-Shannon Divergence**

$$sim_{JS}(\vec{x}||\vec{y}) = D(\vec{x}|\frac{\vec{x}+\vec{y}}{2}) + D(\vec{y}|\frac{\vec{x}+\vec{y}}{2})$$

Example: Lin's Online Similarity Tool

hope (N)		hope (V)		brief (A)		brief (N)	
optimism	0.141	would like	0.158	lengthy	0.256	legal brief	0.139
chance	0.137	wish	0.140	hour-long	0.191	affidavit	0.103
expectation	0.137	plan	0.139	short	0.174	filing	0.0983
prospect	0.126	say	0.137	extended	0.163	petition	0.0865
dream	0.119	believe	0.135	frequent	0.163	document	0.0835
desire	0.118	think	0.133	recent	0.158	argument	0.0832
fear	0.116	agree	0.130	short-lived	0.155	letter	0.0786
effort	0.111	wonder	0.130	prolonged	0.149	rebuttal	0.0778
confidence	0.109	try	0.127	week-long	0.149	memo	0.0768
promise	0.108	decide	0.125	occasional	0.146	article	0.0758

all MINIPAR relations used; $assoc_{Lin}$ used; similarity metric from Lin(98) used.

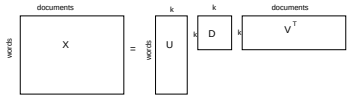
Evaluating Distributional Similarity Metrics

- Intrinsic: Compare to human association norms
- Intrinsic: Compare to thesaurus(es), using precision and recall (e.g., Curran(03) found that Dice and Jaccard and t-test association metric worked best)
- Extrinsic: Use as part of end-to-end applications such as:
  - detection of malapropism (contextual misspellings): "It is minus 15, and then there is the **windscreen** factor on top of that." (Jones and Martin 1997)
  - WSD (Schuetze 1998) and WS ranking (McCarthy et al. 2004)
  - text segmentation (Choi, Wiemer-Hastings and Moore, 2001)
  - automatic thesaurus extraction (Grefenstette 1994, Lin 1998)
  - Information retrieval (Salton, Wang and Yang 1975)
  - essay and exam (multiple choice) grading
  - text comprehension (Landauer and Dumais 1997)
  - semantic priming (Lund and Burgess 1996)

LSA

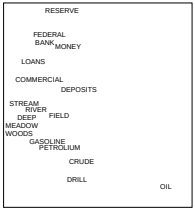
- Vectors in standard vector space are very sparse
- Orthogonal dimensions clearly wrong for near-synonyms *canine–dog*
- Different word senses are conflated into the same dimension
- One way to solve this: **dimensionality reduction**
- Hypothesis for LSA (Latent Semantic Analysis; Landauer): true semantic space has fewer dimensions than number of words observed. Extra dimensions are noise.

Singular Value Decomposition



Similarity between words is measured using matrix U.

Example: first 2 dimensions



from Griffiths, Steyvers, Tenenbaum (2007)

LSA as a cognitive model

- TOEFL test: which of 4 multiple choices is correct synonym of a test word
- LSA: 64.5% correct; real applicants: 64.5%
- Can also explain human learning rate.
  - 40K-100K words known by age 20: 7-15 new words each day; one new word is learned in each paragraph.
  - But: experiments show only 5-10% successful learning of novel words
  - L&D hypothesize that reading provides knowledge about other words not present in immediate text.
  - Simulations show: direct learning gains 0.0007 words per word encountered. Indirect learning gains 0.15 words per article → 10 new words per day

Pado and Lapata 2007

- Investigate dependency-based semantic spaces in detail, on three NLP tasks (WSD, TOEFL-testing, and semantic priming)
- Quantify the degree to which words are attested in similar semantic environments
- Weight the relative importance of different syntactic structures.

Further Reading

- Pado and Lapata (2007). Dependency-based Construction of Semantic Spaces. *Computational Linguistics*.
- Griffiths, Steyvers, Tenenbaum (2007). Topics in Semantic Representation. *Psychological Review*, 114(2):211.
- Landauer and Dumais (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211.