Introduction to Corpus-based Semantic Models

Marco Baroni, Stefan Evert and Alessandro Lenci

ESSLLI Distributional Semantics Workshop

Hamburg, August 4 2008

Outline

Introduction

- "You can tell a word by the company it keeps" (Firth)
- Corpus-based algorithms allow rapid collection of large scale semantic similarity matrices
- Words can be projected into a *semantic space* based on simple distributional information
- *Dogs* are more like *cats* than *cars*
- *Football* and *Manchester* are more "topically similar" than *football* and *Bush*
- Closely related to traditional work in Information Retrieval
  - Compute similarity of *query* to a set of documents

Examples

Nearest neighbours from English model trained on BNC

to sing

- song
- to dance
- sing
- music
- loud
- chorus
- choir
- hymn
- dance
- sound

ceasefire

- mujaheddin
- accord
- Croatia
- peace
- fighting
- Unita
- Djibouti
- PLO
- Iraqi
- Lebanon

Why?

- Lexicon/ontology/thesaurus development
- Language modeling (predict most likely next word in context: for speech recognition, machine translation. . .)
- Text analysis (hidden trends, semantic spaces across time and communities. . .)
- Modeling human semantic/conceptual knowledge and semantic/conceptual acquisition

Outline

Corpus-based Semantic Models (CSMs)

Lund and Burgess, 1998, Landauer et al. 1998, Schütze 1997, Sahlgren 2006...

- General-purpose Corpus-based Lexical Semantic Models
- Meaning of words defined by *set of contexts* in which word occurs
- Similarity of words represented as *geometric distance* among *context vectors*
  - (Alternatively: similarity of probability distributions, relative entropy...)

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark
park
owner
leash

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	+
park	
owner	
leash	

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	+
park	+
owner	
leash	

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	+
park	+
owner	+
leash	

Co-occurrence extraction for target word dog

The dog barked in the park.
The owner of the dog put him
on the leash since he barked.

bark	+
park	+
owner	+
leash	+

Co-occurrence extraction for target word dog

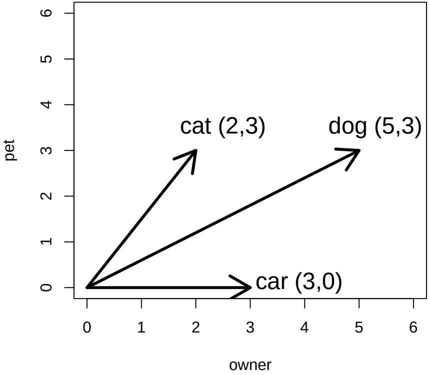
The dog barked in the park.
The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

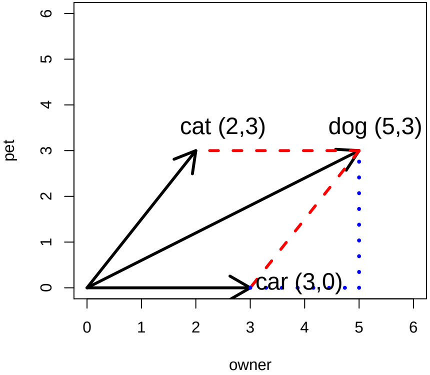
Meaning as co-occurrence

	leash	walk	run	owner	pet	bark
dog	3	5	2	5	3	2
cat	0	3	3	2	3	0
lion	0	3	2	0	1	0
light	0	0	0	0	0	0
bark	1	0	0	2	1	0
car	0	0	1	3	0	0

Similarity in space



Distributional semantics
Similarity in space



What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES…

What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES...
- A definition of context
  - Documents, all words in a fixed span, words in a fixed span minus stop words, words in certain syntactic configurations, words related by certain patterns...

What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES...
- A definition of context
  - Documents, all words in a fixed span, words in a fixed span minus stop words, words in certain syntactic configurations, words related by certain patterns...
- A way to measure co-occurrence in context
  - 0/1, raw frequency, Mutual Information, entropy; distance-based weighting...

What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES...
- A definition of context
  - Documents, all words in a fixed span, words in a fixed span minus stop words, words in certain syntactic configurations, words related by certain patterns...
- A way to measure co-occurrence in context
  - 0/1, raw frequency, Mutual Information, entropy; distance-based weighting...
- A way to construct the context matrix
  - Full co-occurrence matrix, matrix reduced with SVD, sums of random indices...

What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES...
- A definition of context
  - Documents, all words in a fixed span, words in a fixed span minus stop words, words in certain syntactic configurations, words related by certain patterns...
- A way to measure co-occurrence in context
  - 0/1, raw frequency, Mutual Information, entropy; distance-based weighting...
- A way to construct the context matrix
  - Full co-occurrence matrix, matrix reduced with SVD, sums of random indices...
- A way to measure distance/similarity among word vectors
  - cosine, Euclidean distance, Lin's measure...

Parameter Hell!

- At least for some “macro” parameter choices, large “micro” parametric variation
- E.g., if context is given by words in fixed span with stop word filtering:
  - How many words to left, to right?
  - Which stop words?

Parameter Hell!

- At least for some “macro” parameter choices, large “micro” parametric variation
- E.g., if context is given by words in fixed span with stop word filtering:
  - How many words to left, to right?
  - Which stop words?
- Interactions
  - E.g., Rapp 2003 finds that different weighting schemes are more/less suited to matrices with/without SVD

Parameter Hell!

- At least for some “macro” parameter choices, large “micro” parametric variation
- E.g., if context is given by words in fixed span with stop word filtering:
  - How many words to left, to right?
  - Which stop words?
- Interactions
  - E.g., Rapp 2003 finds that different weighting schemes are more/less suited to matrices with/without SVD
- See work by Bullinaria and Levy on the systematic exploration of the parameter space

What makes a semantic space model

- An input corpus
  - Academic American Encyclopedia, newsgroups, BNC, CHILDES...
- A definition of context
  - Documents, all words in a fixed span, words in a fixed span minus stop words, words in certain syntactic configurations, words related by certain patterns...
- A way to measure co-occurrence in context
  - 0/1, raw frequency, Mutual Information, entropy; distance-based weighting...
- A way to construct the context matrix
  - Full co-occurrence matrix, matrix reduced with SVD, sums of random indices...
- A way to measure distance/similarity among word vectors
  - cosine, Euclidean distance, Lin's measure...

Outline

Which context?

- Two words are similar if they tend to occur…
  - In the same documents
  - In paragraphs containing similar words
  - In sentences containing similar words
  - In meaningful syntactic relations with similar words
  - When connected by potentially interesting lexico-semantic patterns

Which context?

- Two words are similar if they tend to occur…
  - In the same documents
  - In paragraphs containing similar words
  - In sentences containing similar words
  - In meaningful syntactic relations with similar words
  - When connected by potentially interesting lexico-semantic patterns
- The wider the context, the more “topical” the relation; the narrower the context, the more “semantic” the relation

Wider and narrower contexts

Nearest neighbours of *dog*

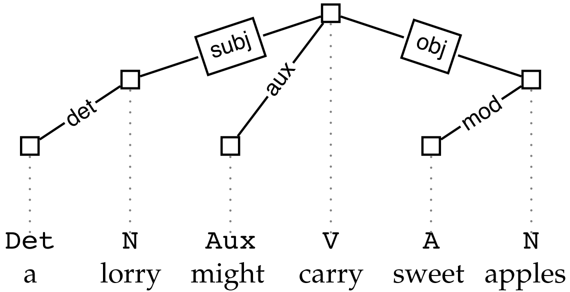
2-word window

- cat
- horse
- fox
- pet
- rabbit
- pig
- animal
- mongrel
- sheep
- pigeon

30-word window

- kennel
- puppy
- pet
- bitch
- terrier
- rottweiler
- canine
- cat
- to bark
- Alsatian

Syntax-based co-occurrences
From Padò and Lapata (2007)



a	Det det N	lorry
lorry	N subj V	carry
might	Aux aux V	carry
apples	N obj V	carry
sweet	A mod N	apples

Lexico-semantic patterns

Baroni and Lenci 2008, Baroni et al. *almost submitted*

- pets such as dogs
- lice in a number of dogs
- dogs and cats
- toys in the kennel of dogs

Outline

Dimensionality reduction

- From a $m \times n$ matrix to a $m \times k$ matrix, where $k << n$
- E.g., from a matrix of 20,000 target words by 10,000 contexts to a matrix of 20,000 target words by 300 “latent dimensions”

Dimensionality reduction

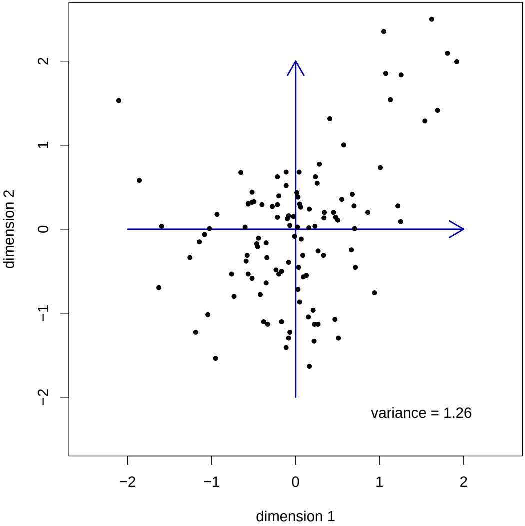
- From a $m \times n$ matrix to a $m \times k$ matrix, where $k << n$
- E.g., from a matrix of 20,000 target words by 10,000 contexts to a matrix of 20,000 target words by 300 "latent dimensions"
- Why?
  - Efficiency/space
  - Hope that latent dimensions will capture "deeper" patterns of correlation

Outline

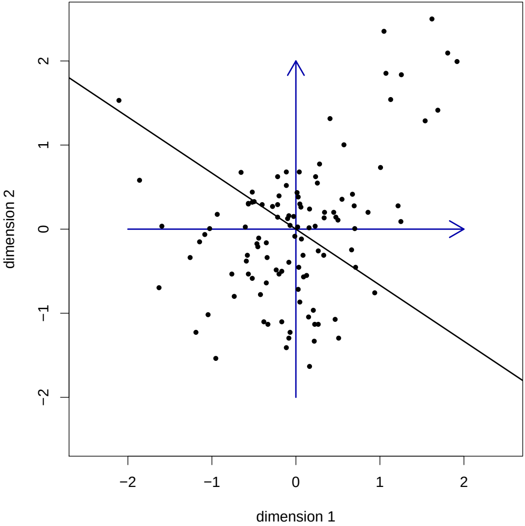
Principal component analysis (PCA)

- Find a set of orthogonal dimensions such that the first dimension "accounts" for the most *variance* in the original data-set, the second dimension accounts for as much as possible of the remaining variance, etc.
- The top k dimensions (principal components) are the best sub-set of k dimensions to approximate the spread in the original data-set

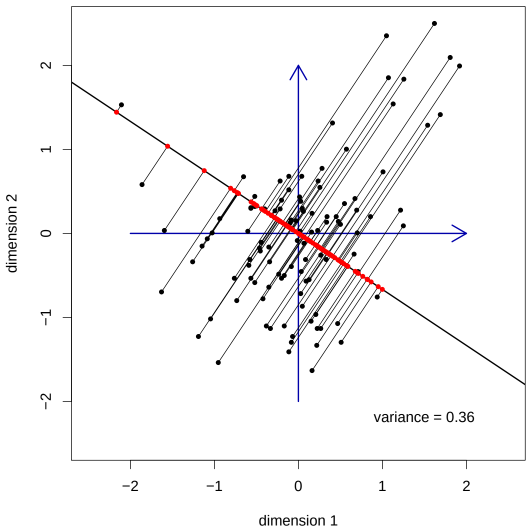
Preserved variance: examples



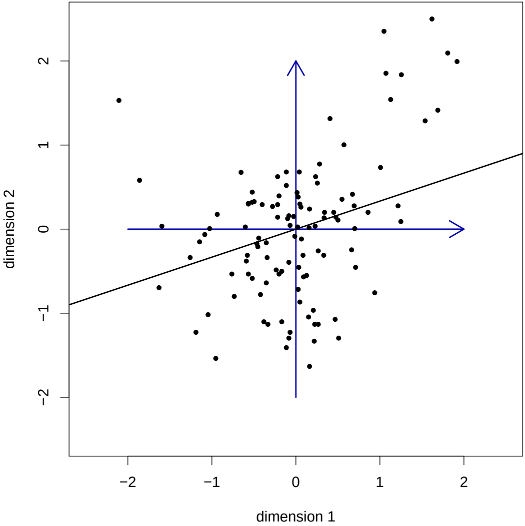
Preserved variance: examples



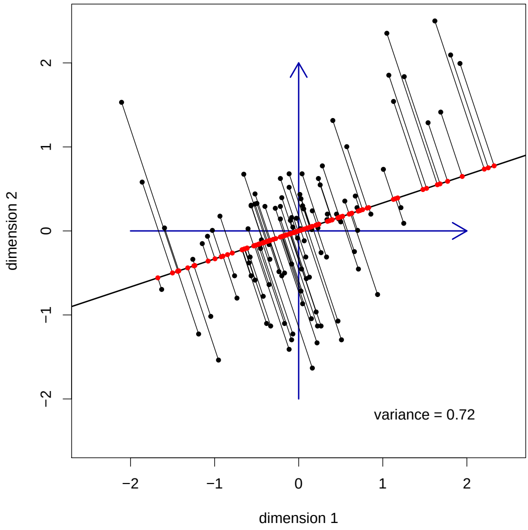
Preserved variance: examples



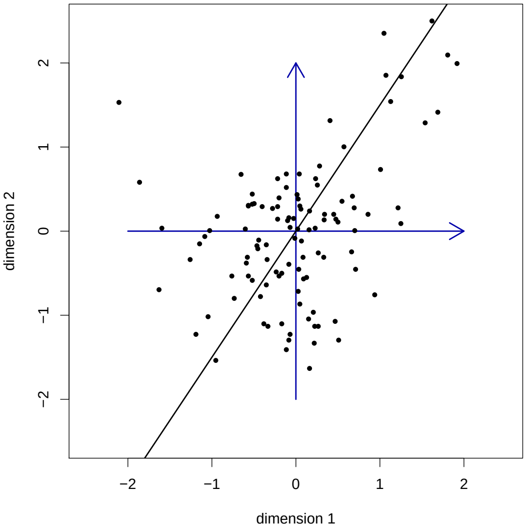
Preserved variance: examples



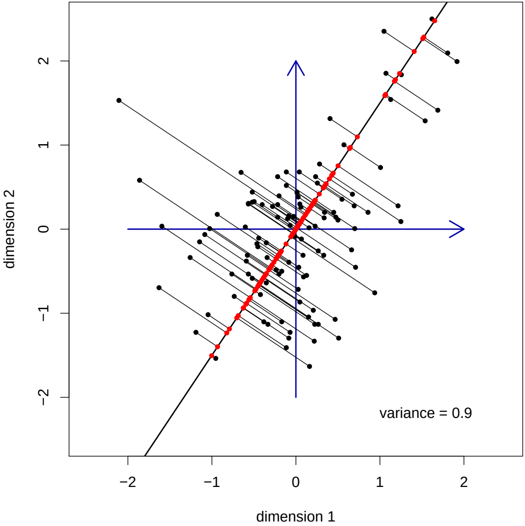
Preserved variance: examples



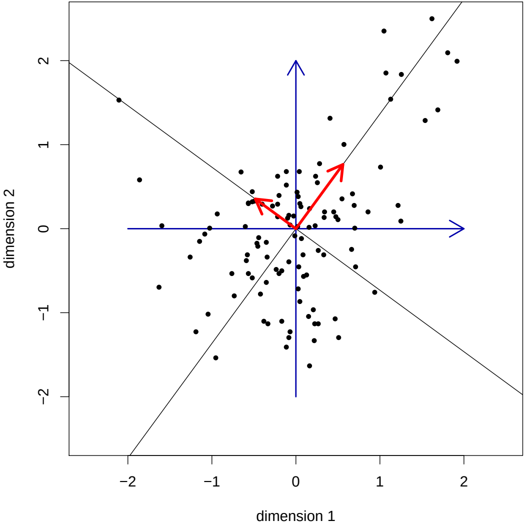
Preserved variance: examples



Preserved variance: examples



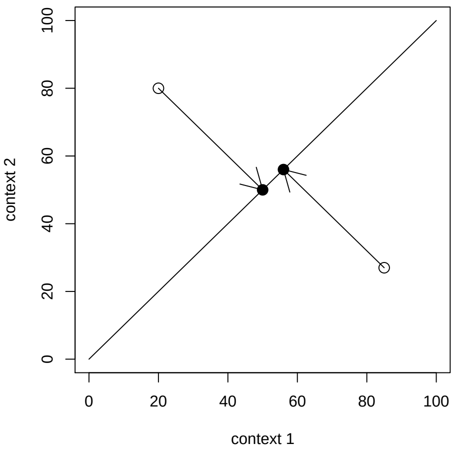
Adding an orthogonal dimension



Dimensionality reduction as generalization

- Contexts with similar co-occurrence patterns likely to be collapsed onto same dimension in reduced space
- Accounts for “synonymic contexts”
- E.g., occurring near *spaceman* or near *astronaut* should count as essentially the same thing

Dimension reduction as generalization



PCA and SVD

- In CSM tradition, principal components are extracted using technique called Singular Value Decomposition
- Essentially, SVD extracts principal components directly from word-by-word (or word-by-document) matrix, instead of building co-variance matrix
- Given co-occurrence matrix M, SVD decomposes M into:

$$M = U\Sigma V^T$$

- First k columns of $U\Sigma$ give projections of target words into reduced space
- Choosing k is an empirical matter; it is often in the 150-300 range

Outline

The low-cost alternative: Random Indexing
Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

The low-cost alternative: Random Indexing
Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

dog	0	0	0	0

The low-cost alternative: Random Indexing

Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

dog	0	0	0	0

dog is a pet

The low-cost alternative: Random Indexing

Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0

The low-cost alternative: Random Indexing
Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0
owner of the dog						

The low-cost alternative: Random Indexing

Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0
owner of the dog	–>	dog	1	-1	0	0

The low-cost alternative: Random Indexing

Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0
owner of the dog	–>	dog	1	-1	0	0
dog on the leash						

The low-cost alternative: Random Indexing
Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0
owner of the dog	–>	dog	1	-1	0	0
dog on the leash	–>	dog	0	-1	-1	0

The low-cost alternative: Random Indexing

Sahlgren 2005

- Represent each context element with a (low-dimensional) index of randomly assigned 1, -1 and (mostly) 0:

pet	0	-1	0	0
owner	1	0	0	0
leash	-1	0	-1	0

- As you go through corpus, add random index corresponding to each context to target word contextual vector:

		dog	0	0	0	0
dog is a pet	–>	dog	0	-1	0	0
owner of the dog	–>	dog	1	-1	0	0
dog on the leash	–>	dog	0	-1	-1	0

- Cosine similarity (or other similarity measure) computed on resulting contextual vectors

Pros and cons

- Pros:
  - Very efficient: low dimensionality from the beginning to the end
  - Implementation trivial (assign random values to vector, sum vectors)
  - Incremental: at any stage, target vectors constitute low-dimensional semantic space

Pros and cons

- Pros:
  - Very efficient: low dimensionality from the beginning to the end
  - Implementation trivial (assign random values to vector, sum vectors)
  - Incremental: at any stage, target vectors constitute low-dimensional semantic space
- Cons:
  - No latent semantic space effect: contexts are “squashed” randomly
  - Lower accuracy, at least on some tasks (Gorman and Curran 2006)

Outline

Topic Models

Hofmann 2001, Blei et al. 2003, Griffiths et al. 2007

- Hierarchical generative probabilistic model
  - pick a distribution over topics (document)
  - pick words from the topic distribution
- Latent "topics" as a form of dimensionality reduction

Topic Models

- Pros:
  - Full-fledged probabilistic model, theoretically easy to integrate in a larger probabilistic picture
  - Handles polysemy/word sense disambiguation well:
    - *bank* might be likely under two different topics, but in context with *money* financial topic prevails
    - no “triangle inequality” issues of geometric models (high probability of *bank* after *river*, *money* does not imply that *river* and *money* are also close)

Topic Models

- Pros:
  - Full-fledged probabilistic model, theoretically easy to integrate in a larger probabilistic picture
  - Handles polysemy/word sense disambiguation well:
    - *bank* might be likely under two different topics, but in context with *money* financial topic prevails
    - no “triangle inequality” issues of geometric models (high probability of *bank* after *river*, *money* does not imply that *river* and *money* are also close)
- Cons:
  - AFAIK, current estimation (and testing) procedures do not scale up well
  - Current Topic Models are document-based, good for finding the “gist” of a text, application to more fine-grained lexical semantics phenomena to be investigated

Outline

Evaluation

- Tricky: performance heavily task-dependent
- Distinguish at least tasks that require recognition of topical similarity and “true” semantic similarity

Evaluation

- Tricky: performance heavily task-dependent
- Distinguish at least tasks that require recognition of topical similarity and “true” semantic similarity
- General trend seems to be in favour of:
  - large-ish corpora (as long as linguistic pre-processing is robust to noise)
  - some linguistic pre-processing (lemmatization, function word filtering)
  - applying SVD

The TOEFL synonym match task

- 80 items

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed, believed, requested, correlated*

The TOEFL synonym match task

- 80 items
- Target: *levied*
  Candidates: *imposed*, *believed*, *requested*, *correlated*

Human performance on the synonym match task

- Average foreign test taker: 64.5%

Human performance on the synonym match task

- Average foreign test taker: 64.5%
- Macquarie University staff (Rapp 2004):
  - Average of 5 non-natives: 86.75%
  - Average of 5 natives: 97.75%

TOEFL results

- Humans:
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%
- Machines:
  - Classic LSA: 64.4%
  - PL's dependency-based model: 73%
  - Rapp's 2003 SVD-based model trained on lemmatized BNC: 92.5%

TOEFL results

- Humans:
  - Foreign test takers: 64.5%
  - Macquarie non-natives: 86.75%
  - Macquarie natives: 97.75%
- Machines:
  - Classic LSA: 64.4%
  - PL's dependency-based model: 73%
  - Rapp's 2003 SVD-based model trained on lemmatized BNC: 92.5%
- (Classic LSA and Rapp's model implicitly tuned on test task)

Outline

Homonymy and polysemy

Nearest neighbours from English model trained on BNC

apple

- Microsystems
- tandem
- inc
- NCR
- corp
- IBM
- inc
- Novell
- Univel
- Oracle

chicken

- bread
- soup
- meat
- pudding
- cake
- sausage
- fried
- tomato
- chocolate
- carrot

“Typing” similarity

Nearest neighbours of *motorcycle* from English model trained on BNC

- motor → component
- car → co-hyponym
- diesel → component?
- to race → proper function
- van → co-hyponym
- BMW → hyponym
- to park → proper function
- vehicle → hypernym
- engine → component
- to steal → frame?

Compositionality

- The following sentences will be indistinguishable to most current CSMs:
  - Pandas eat bamboo
  - Bamboos eat pandas

Some references 1

M. Baroni and A. Lenci (to appear). Concepts and properties in word spaces. In A. Lenci (ed.), *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science*.

D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3**, 993-1022.

J.A. Bullinaria, J.A. and J.P. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* **39**, 510-526.

J. Gorman and J. Curran (2006). Scaling distributional similarity to large corpora. *ACL 2006*, 361-368.

T. Griffiths, M. Steyvers and J. Tenenbaum (2007). Topics in semantic representation. *Psychological Review* **114**, 211-244.

M. Hearst (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING 1992*.

Th. Hofmann (2001). Unsupervised learning by Probabilistic Latent Semantic Analysis. *Machine Learning* **42**, 177-196.

P. Koehn and K. Knight (2002). Learning a translation lexicon from monolingual corpora. *ACL-SIGLEX 2002*.

Some references 2

T. Landauer, P. Foltz, and D. Laham (1998). An introduction to Latent Semantic Analysis. *Discourse Processes* **25**, 259-284.

K. Lund and C. Burgess (1996) Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments & Computers* **28**, 203-208.

S. Padó and M. Lapata (2007). Dependency-based construction of semantic space models. *Computational Linguistics* **33**, 161-199.

R. Rapp (2003). Word sense discovery based on sense descriptor dissimilarity. *Proceedings of the Ninth Machine Translation Summit*.

R. Rapp (2004). A freely available automatically generated thesaurus of related words. *Proceedings of LREC 2004*.

M. Sahlgren (2005). An introduction to Random Indexing. *TKE 2005*.

M. Sahlgren (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm University.

H. Schütze (1997). *Ambiguity resolution in language learning.*, CSLI Publications.