Combining distributional and compositional models of semantics.

Stephen Pulman

Dept of Computer Science, Oxford University

November 25, 2011

- Background: compositionality, formal semantics, notions of distribution
- Vector space models
- Survey of some recent approaches
- Criteria for evaluation

Compositionality

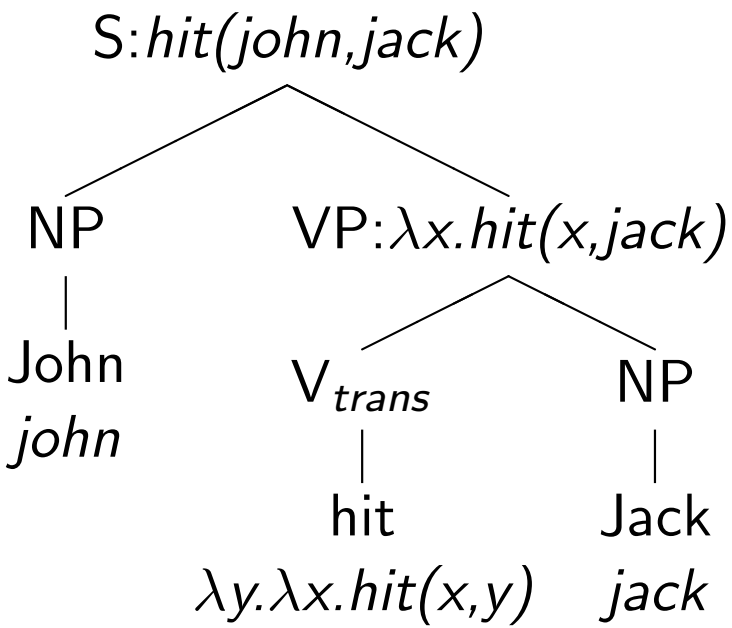
The meaning of a phrase (sentence etc.) is a function of the meanings (denotations) of its components.

- Some obvious exceptions - fixed phrases like 'in spite of', idioms, context-dependence...
- and some non-obvious ones: *stone lion*, *plastic gun*, etc. or Higginbotham et al. on conditionals.
- Compositionality usually regarded as a prerequisite for learnability (cf. Davidson)
- Some have argued it is not an empirical hypothesis...

Formal semantics in the Montague tradition

Usually plays out something like this: syntax driven semantic assembly, each constituent has its own denotation:

1. S $\rightarrow$ NP VP : VP(NP)
2. NP $\rightarrow$ Jack,John etc : jack, john etc.
3. VP $\rightarrow$ V_{trans} NP : V_{trans}(NP)
4. V_{trans} $\rightarrow$ hits : λy.λx.hit(x,y)



Distribution and distributional semantics

Two distinct notions of 'distribution' in the linguistics (and computational linguistics) literature:

Harris and American structuralists:

Verb = {X | X appears in frame 'be . . . -ing', 'something may . . . ' etc.}
NP = {X | X appears in frame '. . . is/are VP', 'it was . . . who/which...' etc.}

All or most linguistic units for a language can be identified by a set of such 'objective' tests - i.e. not appealing to judgements of meaning - and hierarchical structure emerges from a complete, often mutually recursive, set of such statements: e.g. the head of a phrase can appear alone everywhere the phrase can appear, a conjunction can appear where either of its conjuncts appear.

Collocations, and collocational meaning

J R Firth

"You shall know a word by the company it keeps."
"Collocations of a given word are statements of the habitual or customary places of that word."
Different word senses have different collocates:
racing/lecture circuit vs. short circuit vs. printed/integrated circuit board.
Collocation = very fine grained distribution? (cf. Maurice Gross's work)

Manning and Schütze:

Collocations include noun phrases like *strong tea* and *weapons of mass destruction*, phrasal verbs like *to make up*, and other stock phrases like *the rich and powerful*. Particularly interesting are the subtle and not-easily-explainable patterns of word usage that native speakers all know: why we say *a stiff breeze* but not *??a stiff wind* (while either *a strong breeze* or *a strong wind* is okay), or why we speak of *broad daylight* (but not *?bright daylight* or *??narrow darkness*). Collocations are characterized by limited compositionality...

Schütze: Vector space models of word meaning

Construct 'word space'

Select some words as 'features' or 'context words' and construct vectors for the target words representing how many times the feature words occurred within a 50 word window of the target words:

> **Bank** erosion and **stream** widening may occur with strong **water** flow.
>
> One way of **raising** this **finance** is to go to a **bank**.

	context	words			
target	river	stream	money	raise	finance
bank	10	15	25	20	13
water	28	25	2	15	0
cheque	0	0	30	20	25
etc.					

Use χ^2 test to make sure the co-occurrences are meaningful. Features can be 'local' i.e. those that occur in the contexts, or 'global' i.e. those that are most frequent in the corpus.

How do we deal with vectors?

The resulting rows are vectors in a multidimensional space, and can be compared for length and direction (we're not usually interested in length). Words with similar meanings should have vectors pointing in a similar direction (but antonyms do too!). We measure semantic similarity by 'cosine distance': 1 = identical, 0 = unrelated

Gory details

- 'dot product': $\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$.
- length $= \|\mathbf{a}\| = \sqrt{{a_1}^2 + {a_2}^2 + {a_3}^2} = \sqrt{\mathbf{a} \cdot \mathbf{a}}$.
- similarity $= \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$

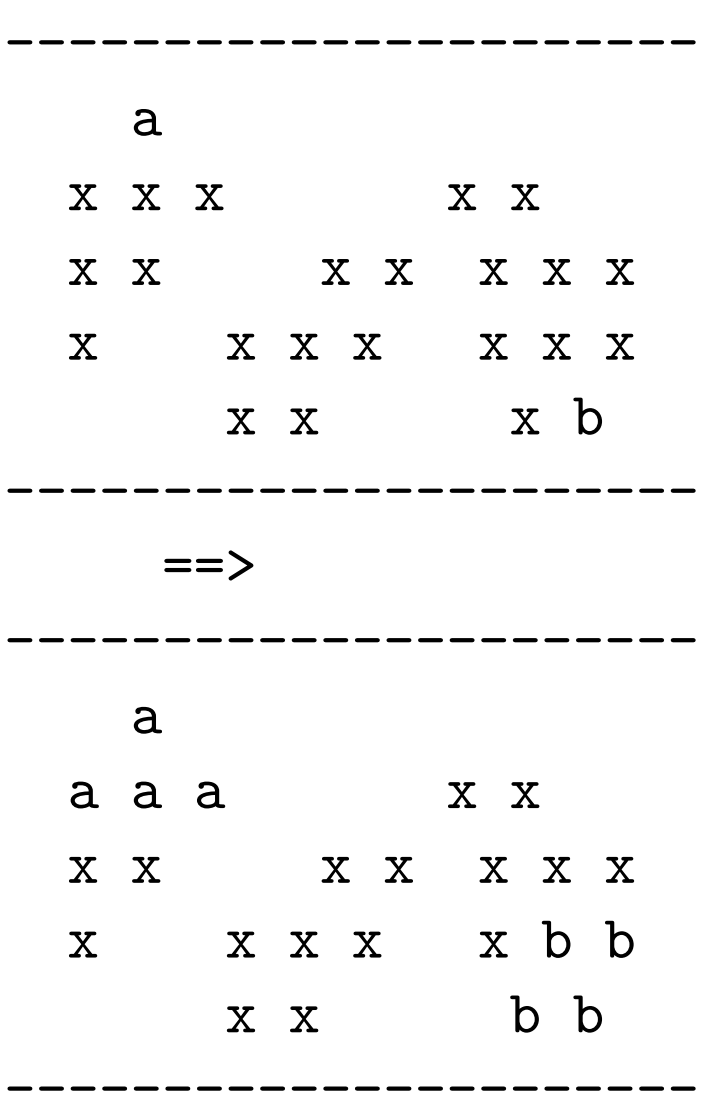
Distinguishing different word senses

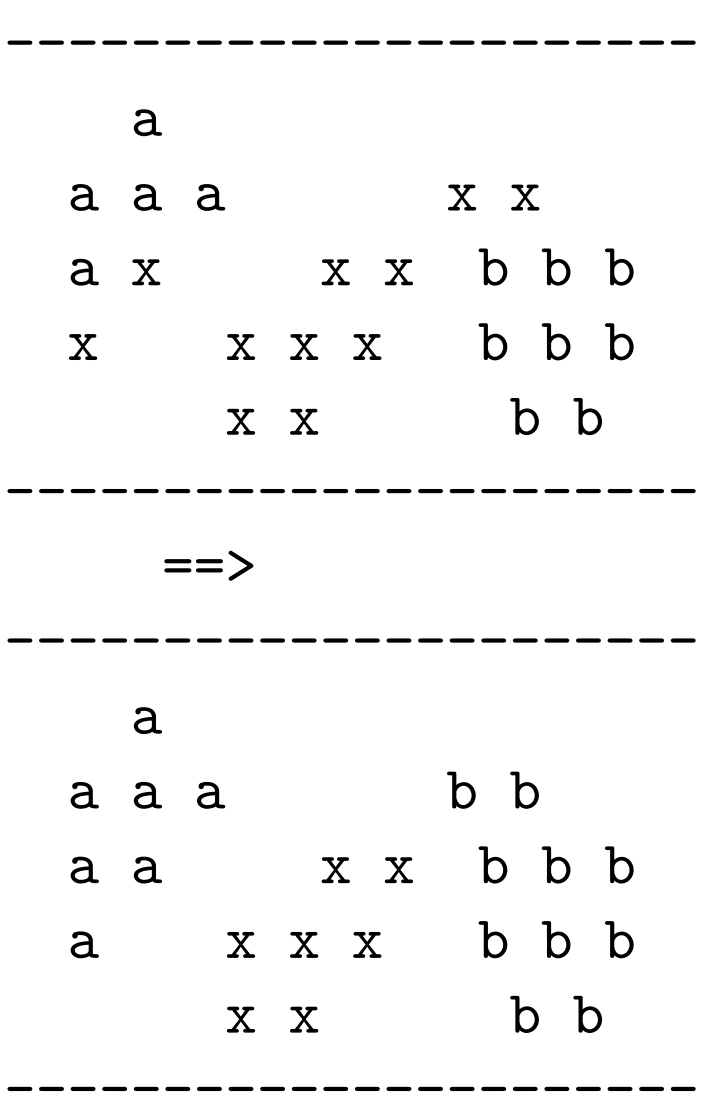
Notice that different senses of 'bank' are not differentiated, so:

Construct 'context space'
Go back to the corpus, and for each occurrence of a target word, construct the centroid (sum) of the vectors for each of the context words within the relevant window. The centroid 'averages' the direction of the set of vectors.
E.g in context 1 for **bank**, you would sum *stream+water*,
but in context 2, it would be *raise+finance*.

Construct 'senses' by clustering the context vectors. There are many clustering algorithms. Schütze used a form of 'agglomerative clustering', where each vector initially forms its own cluster, and clusters are repeatedly merged based on some criterion until the target number of clusters (2-10 here) is arrived at. Each cluster should correspond to a distinct sense, which can be represented by the centroid of the vectors in the cluster.

Clustering




a
a a a x x
a x x x b b b
x x x x b b b
x x b b
==>
a
a a a b b
a a x x b b b
a x x x b b b
x x b b

To disambiguate a word in a context

1. construct the vector for that context, as above
2. compare the sense vectors for the word to the context vector
3. choose the sense whose vector is closest to the context vector

Evaluation: use ambiguous words if you have the data, otherwise use 'pseudo-words'. Choose two distinct words from the corpus, for example *computer* and *banana*, and replace all occurrences of them by the pseudo-word *bananacomputer*. We can evaluate how well the algorithm does on disambiguating *bananacomputer* by looking at the original form of the corpus. In fact, since single words are often ambiguous, it is better to create pseudo-words from pairs: e.g.
wide range + consulting firm

Results:

For naturally ambiguous words: **interest, space, plant, ...** and pseudo-words: **'wide range'+'consulting firm', 'league baseball'+'square feet',...** average accuracy is:

Natural:	2 clusters	10 clusters
local	76.0%	84.4%
global	80.8%	83.1%
Artificial:		
local	89.9%	92.2%
global	98.6%	98.0%

But is this a general theory of **word meaning**?

Philosophical asides... two oppositions

Empiricism vs. Nativism

Empiricism: you learn by experience with minimal *a priori* knowledge; nativism: you have rich *a priori* knowledge, much less experience needed.

Harris's distributional hypothesis can be cast as an empiricist 'learning procedure' for language.
Chomsky concluded that some properties of language could never be learned in this way, and turned to a form of nativism.

Analytic vs. synthetic truths

Sentences true by virtue of meaning, or true by virtue of the facts?

If there are analytic truths, they cannot be derived distributionally. So how are they learned? But if there are no analytic truths and word meanings are learned by distributional means, this opens the door to relativism, of all kinds: how do I know you understand words the same way I do?

Word Senses Within Semantic Theories

Fodor and Lepore(1999): any theory of semantic content (word meaning) should support:

- assignment of truth/satisfaction conditions to sentences
- **compositionality**
- translation: 'Pre-theoretic intuition has it that meaning is what good translations preserve. A semantic theory should provide a notion of meaning according to which this turns out true.'
- intentional explanation: 'A semantic theory should reconstruct a notion of content that is adequate to the purposes of intentional (e.g., belief/desire) explanation.'

Atomism, Molecularism and Holism

What, if any, are the meaning-constitutive relations between words? Three (coarsely drawn) positions:

- Atomism: no words are connected to any others by meaning constitutive relations, only by beliefs. (Fodor, 1998 on). Implicitly the position held by formal semanticists.
- Molecularism: at least some meanings are related to/reducible to others by meaning-constitutive relations:

  'persuade X that P' $\rightarrow$ 'X believes that P'

  'X killed Y' $\rightarrow$ 'Y is dead'; etc.

  (Almost all linguists). In formal semantics these meaning relations are by stipulation (meaning postulates)
- Holism: all meanings are related (and thus defined by) relations to all (some) other meanings/beliefs: an interconnected web. There is no difference between meaning-constitutive and belief-constitutive relations. (Quine, Davidson...). This is the view that vector space models implement.

The challenge:

We know that vector based models of word meaning can perform well on tasks like disambiguation (Schütze) and detection of synonyms, especially when the bases include syntactic information: Padó and Lapata (2008):

- Semantic similarity: 65 pairs of nouns, some nearly synonymous (gem,jewel), others not (noon,string) with human judgements.
- Result: A Pearson coefficient of around 0.6 between the vector space model and human judgements.
- The TOEFL tests: you will find the office at the main **intersection**: (a) place (b) crossroads (c) roundabout (d) building
- Result: Vector space model gets it right about 73% of the time. (Non-native speakers get about 64%!)

But can we show that vector-based models of meaning can support compositionality? Can we get vectors for sentences such that 'dogs chase cats' differs from 'cats chase dogs'?

Beginnings of compositionality? (Widdows 2003)

Two vectors are orthogonal if their dot product is 0. The orthogonal subspace of a vector a is the set of vectors orthogonal to it. We can capture the effect of conjunction and 'negation' $a \wedge \neg b$ by projecting vector a onto the orthogonal subspace of b

suit		suit NOT	lawsuit
suit	1.000000	pants	0.810573
lawsuit	0.868791	shirt	0.807780
suits	0.807798	jacket	0.795674
plaintiff	0.717156	silk	0.781623
sued	0.706158	dress	0.778841
plaintiffs	0.697506	trousers	0.771312
suing	0.674661	sweater	0.765677
lawsuits	0.664649	wearing	0.764283
etc			

Clark and Pulman 2007

- Obvious vector operations like addition or multiplication are not compositional: 'man bites dog' will mean the same as 'dog bites man'.
- We tried to overcome this by treating composition as a tensor product operation ($\otimes$) combining vectors both from a word space and a grammatical relations space:
- `[a,b,c]`$\otimes$`[d,e] = [ad,ae,bd,be,cd,ce]`
- John drinks strong beer = John$\otimes$subj$\otimes$drinks$\otimes$obj$\otimes$ (beer$\otimes$adj$\otimes$strong)
- Good: man$\otimes$subj$\otimes$bites$\otimes$dog$\otimes$obj $\neq$ dog$\otimes$subj$\otimes$bites$\otimes$man$\otimes$obj
- Bad: Dimension(A$\otimes$B) = Dimension(A)$\times$Dimension(B), so can only compare sentences with isomorphic structures, because otherwise tensor products have different lengths.

Mitchell and Lapata 2008

'Vector-based Models of Semantic Composition'
Observations from the psycholinguistics literature: (Kintsch).

- Many verbs like 'run' are ambiguous. Their subject will 'select' the appropriate sense: *The horse ran* vs. *The colour ran.*
- The resulting meaning should be closer to a relevant related word or 'landmark' like *gallop* vs. *dissolve*.
- M and L collected human judgements on 15 verbs x 4 noun combinations with 2 landmarks each.
- E.g. 'his shoulders slumped' is closer to 'slouch' than to 'decrease', whereas 'his shares slumped' is closer to 'decrease' than to 'slouch'.

Experiments

- Vectors were built on a lemmatised version of the BNC
- context window of +/- 5 words, 2000 context words.
- vector components = $\frac{P(context|target)}{P(context)}$
- Various ways of combining vectors were tried: addition, weighted addition, multiplication, or a weighted combination of addition and multiplication.
- The best results (i.e. closest to human judgements) involved **pointwise multiplication**, i.e. `[a,b,c]*[d,e,f] = [ad,be,cf]`

Erk and Padó 2008

Learn a 'structured vector' space consisting of triples of form w = $\langle$v,R,$R^{-1}\rangle$ where:

- *v* is a word space vector of the usual kind
- R is a set of vectors representing possible occupants of grammatical or dependency relations of the type 'has-subj', 'has-obj', etc.
- R^{-1} is a set of vectors representing relations like 'is-subj-of', 'is-modified-by': i.e.
- (R are the relations that the word selects, and R^{-1} are relations that the word is selected by.)

These are formed as a weighted centroid of the individual words. So, for a word like 'catch' we might have vectors like: $\langle$catch,{has-subj,has-obj, ...},{is-comp-of,is-modified-by,....}$\rangle$ where 'has-subj' would be a vector formed from observed subjects of 'catch' and 'is-comp-of' would be a vector formed from verbs observed to take 'catch' as a complement.

Compositional operations

So when the following words are combined in the object relation, 'catch ball'

$\langle$catch,{has-obj,α},{is-comp-of,...}$\rangle$
$\langle$ball, {has-mod,...},{is-obj-of,β}$\rangle$

the result is to replace the original meanings by a pair of structured vectors:

'catch in the context of ball':
$\langle$catch$\times$is-obj-of,α,{is-comp-of,...}$\rangle$
'ball in the context of catch':
$\langle$ball$\times$has-obj,{has-mod,...},$\beta\rangle$

As with M and L, they tried different vector operations, but **component-wise multiplication** was best.

In principle, each of these words could be further combined, so that, for example, the subject of 'catch' would indirectly be influenced by the properties of 'ball'.

E and P replicate M and L's experiment with slightly - but not significantly - better results, and show improved performance on a lexical substitution task.

Marco Baroni and Roberto Zamparelli 2010

'Nouns are vectors, adjectives are matrices.'
With a large enough corpus, you can learn a distributional model for frequently occurring Adj+Noun combinations, as well as for Adj and Noun separately. We can then test empirically whether the vectors for the Adj+Noun combination are a compositional function of the Adj and Noun vectors.
B and Z tested several different models:

- Addition: Adj+Noun = Adj vector + Noun vector
- Multiplication: Adj+Noun = pointwise multiplication of Adj and Noun
- Adjective is a matrix, not a vector: Adj+Noun = Adj*Noun

Multiply vector by matrix

```
[A B C][P]   [AP + BQ + CR]
[D E F][Q] = [DP + EQ + FR]
[G H I][R]   [GP + HQ + IR]
```

Learning matrices

B and Z have the vector for the Noun, and the vector for the Adj+Noun, and they solve, for each Adj, the problem:

```
[? ? ?]     [n1]
[? ? ?]  *  [n2]  =  [an1,an2,an3]
[? ? ?]     [n3]
```

Dimension of Noun and Adj+Noun is x, so matrix is x*x. The intuition is that the j weights in the i-th row of the matrix predict the values of the i-th dimension of the Adj+Noun vector as a linear combination of the j dimensions of the component noun.

Results

Train matrices for 36 adjectives. Test by, among other things, making 'new' Adj and Noun combinations i.e. not those used to learn the Adj matrices) and seeing whether the actual vector for that Adj+Noun combination is close in cosine distance. Rank out of 26k candidates:

	25%	50%	75%
matrix	17	170	1k+
add	27	252	1k+
mult	279	1k+	1k+

Can also look at nearest generated Adj+Noun:

	observed	predicted
recent request	recent enquiry	recent enquiry
difficult partner	difficult organisation	difficult department
special something	little animal	special thing

Is this really compositional?

- Many other recent attempts: Guevara, Daoud Clarke, Oxford, Saarbrücken... all using more complex bases for the relevant vector spaces, usually involving dependency relations.
- But all the word vectors are 'first order' in that all the different senses of a word are in the same vector (unlike Schütze's 'second order' vectors).
- And they all get the best results using some form of multiplication.
- It's not difficult to see why. The tasks usually amount to a three (or more) way similarity judgement, which depends on a disambiguation.
- **The effect of pointwise multiplication is to disambiguate, not to compose!**

Multiplication disambiguates

Assume that

sim(horse,gallop) > sim(horse, dissolve)
sim(colour,dissolve) > sim(colour,gallop)

The first order vector for 'run' will have components for collocation with 'horse'-related words and 'colour'-related words. The vectors for 'horse' will have lower values for the context words associated with 'colour' and vice-versa. So when we multiply 'horse' by 'run' the effect will be to reduce the values of the 'colour' components of the vector:

context:	pasture	feed	gallop	ride	spectrum	bright	durable
colour	1	2	1	1	5	6	8
horse	6	8	7	8	1	3	1
run	5	5	5	5	5	5	5
colour×run	5	10	5	5	25	30	40
horse×run	30	40	35	40	5	15	5

Semantic compositionality

When we combine words into phrases, at least two things happen:

- irrelevant meanings of ambiguous words are largely[1] filtered out
- we get a meaning representing the combination

Current vector space models only achieve the first. They are doing **disambiguation, not composition**. It's arguable that this kind of disambiguation is not even dependent on compositionality:

John sat on the bank. He was fishing.

Semantic compositionality of the kind illustrated by formal semantics in fact assumes that word sense disambiguation has already been done, with the word mapped to the appropriate logical constant!

[1] Mary painted a nice watercolour of a horse near a tree. But she spilled water on it and the horse began to run into the tree.

What next?

Need to develop different measures that will separate disambiguation from true composition. Maybe use Schütze-type second order vectors?
Suggested tasks:

- definitions: Is 'John is a carnivore' closer to 'John eats meat' than 'John eats vegetables'?
- syntactic variation: Is 'the cat chased the mouse' closer to 'the mouse was chased by the cat' than to 'the mouse chased the cat', and so on for other similar constructions? (But you could probably do this with a good parser?)

(Maybe) prototype structure:

- 'pet' closer to 'parrot' than 'sparrow'
- 'bird' closer to 'sparrow' than to 'parrot'
- 'pet bird' closer to 'parrot' than to 'sparrow' and different from 'bird pet'

References

K. Erk and S. Padó 2008 A Structured Vector Space Model for Word Meaning in Context (details). Proceedings of EMNLP 2008.
Z. Harris "From Morpheme to Utterance". Language 22:3.161183.
J. R. Firth Papers in Linguistics 1934-1951 (1957) London: Oxford University Press.
Marco Baroni and Roberto Zamparelli 2010 20 Nouns are vectors, adjectives are matrices, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), East Stroudsburg PA: ACL, 1183-1193
S. Clark and S Pulman 2007 Combining Symbolic and Distributional Models of Meaning Proceedings of the AAAI Spring Symposium on Quantum Interaction, pp.52-55, Stanford, CA, 2007
J. Mitchell and M Lapata 2008 Vector-based Models of Semantic Composition In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 236-244. Columbus, OH.
Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke and Stephen Pulman, 2011, Concrete Sentence Spaces for Compositional Distributional Models of Meaning Proceedings of the 9th International Conference on Computational Semantics (IWCS-11), pp.125-134, Oxford, UK, 2011
Jerry Fodor, 1998, "Concepts: Where Cognitive Science Went Wrong", Clarendon Press, Oxford.
Jerry Fodor and Ernie Lepore, 1999, "All At Sea in Semantic Space: Churchland on Meaning Similarity", the Journal of Philosophy, v 96, pp. 381-403
Hinrich Schütze, 1998, "Automatic word sense discrimination", Journal of Computational Linguistics, 24(1):97–123.