Outline of today's lecture

Compositional distributional semantics (catching up)

Discourse structure

Coherence

Referring expressions and anaphora

Algorithms for anaphora resolution

Outline

Compositional distributional semantics

Extending distributional semantics to model the meaning of longer phrases and sentences.

Two kinds of models:

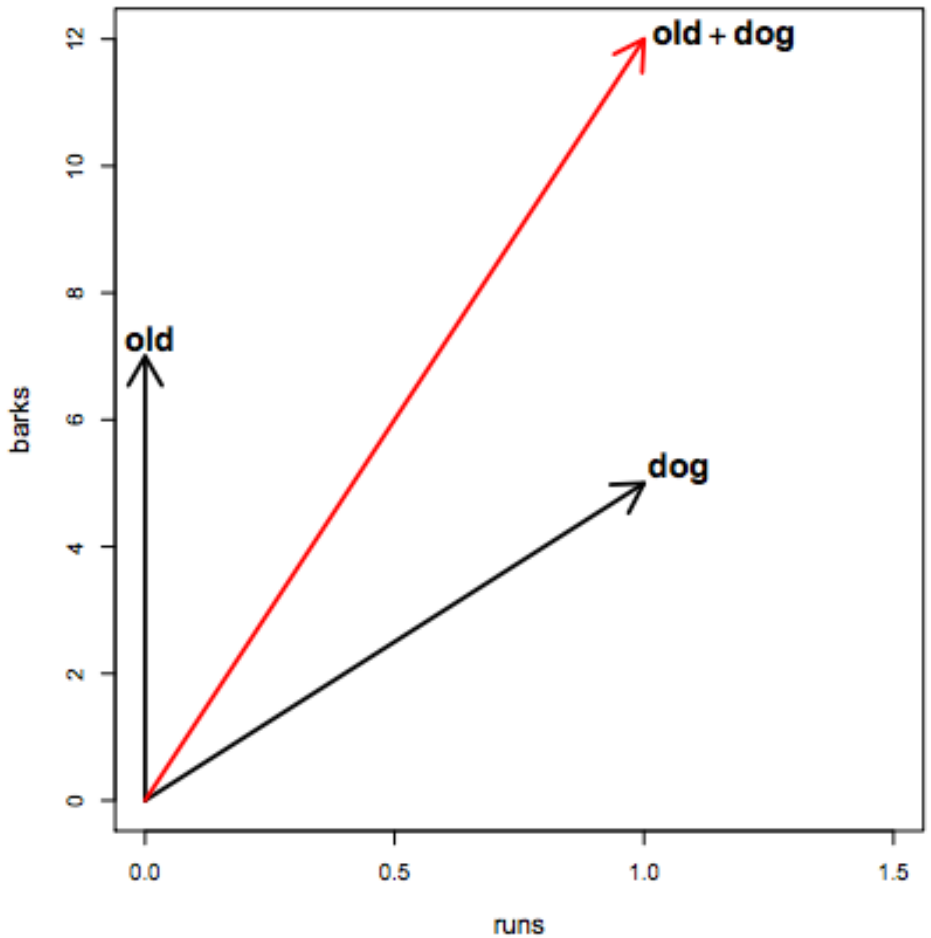
1. Vector mixture models
2. Lexical function models

1. Vector mixture models

Mitchell and Lapata, 2010. *Composition in Distributional Models of Semantics*

Models:

- Additive
- Multiplicative



Additive and multiplicative models

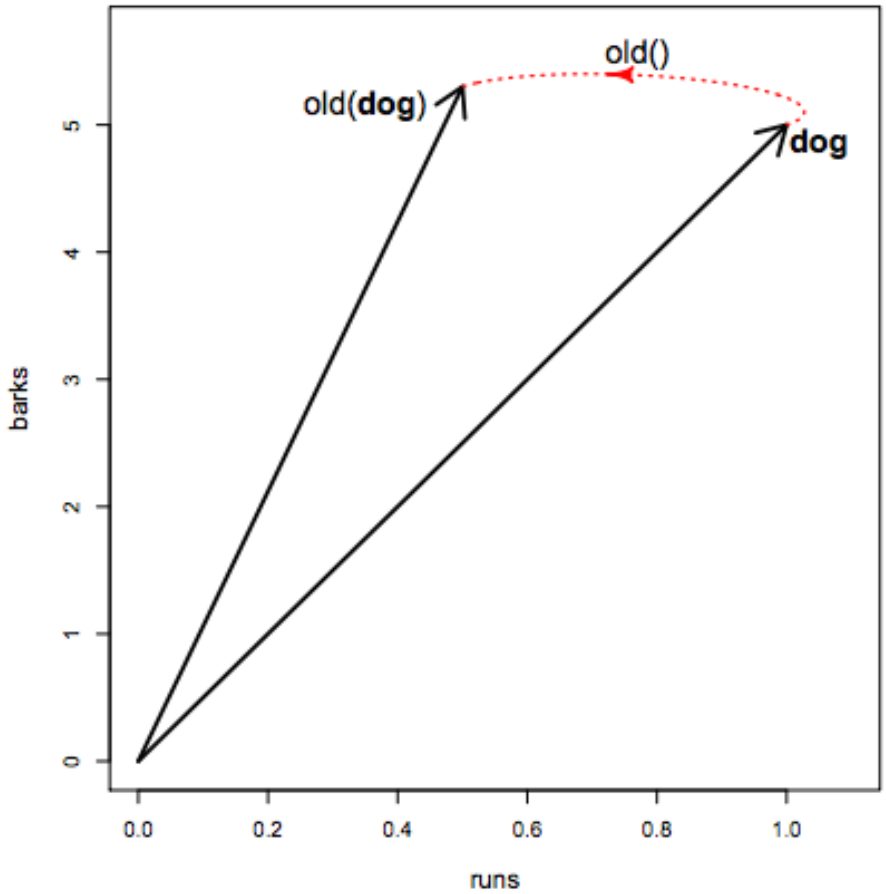
				additive		multiplicative	
	dog	**cat**	**old**	**old + dog**	**old + cat**	**old ⊙ dog**	**old ⊙ cat**
runs	1	4	0	1	4	0	0
barks	5	0	7	12	7	35	0

- correlate with human similarity judgments about adjective-noun, noun-noun, verb-noun and noun-verb pairs
- but... commutative, hence do not account for word order *John hit the ball = The ball hit John*!
- more suitable for modelling content words, would not port well to function words:
  e.g. *some dogs; lice and dogs; lice on dogs*

2. Lexical function models

Distinguish between:

- words whose meaning is directly determined by their distributional behaviour, e.g. nouns
- words that act as functions transforming the distributional profile of other words, e.g., verbs, adjectives and prepositions



Lexical function models

Baroni and Zamparelli, 2010. *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.*

Adjectives as lexical functions

$$old\ dog = F_{old}(dog)$$

- Adjectives are parameter matrices (Θ_{old} , Θ_{furry}, etc.).
- Nouns are vectors (house, dog, etc.).
- Composition is simply $old\ dog = \Theta_{old} \times dog$.

Learning adjective matrices

1. Obtain vector n_j for each noun n_j in lexicon.
2. Collect adjective noun pairs (a_i, n_j) from corpus.
3. Obtain vector h_{ij} of each bi-gram (a_i, n_j)
4. The set of tuples $\{(n_j, h_{ij})\}_j$ is a dataset D_i for adj. a_i
5. Learn matrix Θ_i from D_i using linear regression.

OLD	runs	barks
runs	0.5	0
barks	0.3	1

$\times$

	dog
runs	1
barks	5

$=$

	OLD(dog)
runs	$(0.5 \times 1) + (0 \times 5) = 0.5$
barks	$(0.3 \times 1) + (5 \times 1) = 5.3$

Outline

Document structure and discourse structure

- Most types of document are highly structured, implicitly or explicitly:
  - Scientific papers: conventional structure (differences between disciplines).
  - News stories: first sentence is a summary.
  - Blogs, etc etc
- Topics within documents.
- Relationships between sentences.

Rhetorical relations

Max fell. John pushed him.

can be interpreted as:

1. Max fell because John pushed him.
   EXPLANATION

or

2 Max fell and then John pushed him.
   NARRATION

Implicit relationship: discourse relation or rhetorical relation
because, *and then* are examples of cue phrases

Outline

Coherence

Discourses have to have connectivity to be coherent:

Kim got into her car. Sandy likes apples.

Can be OK in context:

Kim got into her car. Sandy likes apples, so Kim thought she'd go to the farm shop and see if she could get some.

Coherence

Discourses have to have connectivity to be coherent:

Kim got into her car. Sandy likes apples.

Can be OK in context:

Kim got into her car. Sandy likes apples, so Kim thought she'd go to the farm shop and see if she could get some.

Coherence in generation

Language generation needs to maintain coherence.

In trading yesterday: Dell was up 4.2%, Safeway was down 3.2%, HP was up 3.1%.

Better:

Computer manufacturers gained in trading yesterday: Dell was up 4.2% and HP was up 3.1%. But retail stocks suffered: Safeway was down 3.2%.

More about generation in the next lecture.

Coherence in interpretation

Discourse coherence assumptions can affect interpretation:

Kim's bike got a puncture. She phoned the AA.

Assumption of coherence (and knowledge about the AA) leads to *bike* interpreted as motorbike rather than pedal cycle.

John likes Bill. He gave him an expensive Christmas present.

If EXPLANATION - 'he' is probably Bill.
If JUSTIFICATION (supplying evidence for first sentence), 'he' is John.

Factors influencing discourse interpretation

1. Cue phrases (e.g. *because, and*)
2. Punctuation (also prosody) and text structure.

   Max fell (John pushed him) and Kim laughed.
   Max fell, John pushed him and Kim laughed.
3. Real world content:

   Max fell. John pushed him as he lay on the ground.
4. Tense and aspect.

   Max fell. John had pushed him.
   Max was falling. John pushed him.

Hard problem, but 'surfacy techniques' (punctuation and cue phrases) work to some extent.

Rhetorical relations and summarisation

Analysis of text with rhetorical relations generally gives a binary branching structure:

- nucleus and satellite: e.g., EXPLANATION, JUSTIFICATION

  Max fell because John pushed him.

- equal weight: e.g., NARRATION

Rhetorical relations and summarisation

Analysis of text with rhetorical relations generally gives a binary branching structure:

- nucleus and satellite: e.g., EXPLANATION, JUSTIFICATION

  Max fell because John pushed him.

- equal weight: e.g., NARRATION

Summarisation by satellite removal

If we consider a discourse relation as a relationship between two phrases, we get a binary branching tree structure for the discourse. In many relationships, such as Explanation, one phrase depends on the other: e.g., the phrase being explained is the main one and the other is subsidiary. In fact we can get rid of the subsidiary phrases and still have a reasonably coherent discourse.

Summarisation by satellite removal

If we consider a discourse relation as a relationship between two phrases, we get a binary branching tree structure for the discourse. In many relationships, such as Explanation, one phrase depends on the other: e.g., the phrase being explained is the main one and the other is subsidiary. In fact we can get rid of the subsidiary phrases and still have a reasonably coherent discourse.

Summarisation by satellite removal

If we consider a discourse relation as a relationship between two phrases, we get a binary branching tree structure for the discourse. In many relationships, such as Explanation, one phrase depends on the other: e.g., the phrase being explained is the main one and the other is subsidiary. In fact we can get rid of the subsidiary phrases and still have a reasonably coherent discourse.

We get a binary branching tree structure for the discourse. In many relationships one phrase depends on the other. In fact we can get rid of the subsidiary phrases and still have a reasonably coherent discourse.

Outline

Referring expressions

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

referent a real world entity that some piece of text (or speech) refers to. the actual Prof. Ferguson

referring expressions bits of language used to perform reference by a speaker. 'Niall Ferguson', 'he', 'him'

antecedent the text initially evoking a referent. 'Niall Ferguson'

anaphora the phenomenon of referring to an antecedent.

cataphora pronouns appear before the referent (rare)

What about *a snappy dresser*?

Pronoun resolution

- Identifying the referents of pronouns
- Anaphora resolution: generally only consider cases which refer to antecedent noun phrases.

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Pronoun resolution

- Identifying the referents of pronouns
- Anaphora resolution: generally only consider cases which refer to antecedent noun phrases.

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Pronoun resolution

- Identifying the referents of pronouns
- Anaphora resolution: generally only consider cases which refer to antecedent noun phrases.

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Outline

Anaphora resolution as supervised classification

- assign class to data points (also called instances)
- instances: potential pronoun/antecedent pairings
- class is TRUE/FALSE
- training data labelled with class and features
- derive class for test data based on features
- candidate antecedents are all NPs in current sentence and preceeding 5 sentences (excluding pleonastic pronouns)

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Hard constraints: Pronoun agreement

- A little girl is at the door — see what she wants, please?
- My dog has hurt his foot — he is in a lot of pain.
- * My dog has hurt his foot — it is in a lot of pain.

Complications:

- The team played really well, but now they are all very tired.
- Kim and Sandy are asleep: they are very tired.
- Kim is snoring and Sandy can't keep her eyes open: they are both exhausted.

Hard constraints: Reflexives

- John$_i$ cut himself$_i$ shaving. (himself = John, subscript notation used to indicate this)
- # John$_i$ cut him$_j$ shaving. (i $\neq$ j — a very odd sentence)

Reflexive pronouns must be coreferential with a preceeding argument of the same verb, non-reflexive pronouns cannot be.

Hard constraints: Pleonastic pronouns

Pleonastic pronouns are semantically empty, and don't refer:

- It is snowing
- It is not easy to think of good examples.
- It is obvious that Kim snores.
- It bothers Sandy that Kim snores.

Soft preferences: Salience

- Recency: More recent antecedents are preferred. They are more accessible.

  Kim has a big car. Sandy has a smaller one. Lee likes to drive it.

- Grammatical role: Subjects $>$ objects $>$ everything else:

  Fred went to the Grafton Centre with Bill. He bought a CD.

- Repeated mention: Entities that have been mentioned more frequently are preferred.

Soft preferences: Salience

- Parallelism Entities which share the same role as the pronoun in the same sort of sentence are preferred:

  Bill went with Fred to the Grafton Centre. Kim went with him to Lion Yard. Him=Fred

- Coherence effects: The pronoun resolution may depend on the rhetorical / discourse relation that is inferred.

  Bill likes Fred. He has a great sense of humour.

Features

Cataphoric Binary: t if pronoun before antecedent.

Number agreement Binary: t if pronoun compatible with antecedent.

Gender agreement Binary: t if gender agreement.

Same verb Binary: t if the pronoun and the candidate antecedent are arguments of the same verb.

Sentence distance Discrete: { 0, 1, 2 . . . }

Grammatical role Discrete: { subject, object, other } The role of the potential antecedent.

Parallel Binary: t if the potential antecedent and the pronoun share the same grammatical role.

Linguistic form Discrete: { proper, definite, indefinite, pronoun }

Feature vectors

pron	ante	cat	num	gen	same	dist	role	par	form
him	*Niall F.*	f	t	t	f	1	subj	f	prop
him	*Ste. M.*	f	t	t	t	0	subj	f	prop
him	*he*	t	t	t	f	0	subj	f	pron
he	*Niall F.*	f	t	t	f	1	subj	t	prop
he	*Ste. M.*	f	t	t	f	0	subj	t	prop
he	*him*	f	t	t	f	0	obj	f	pron

Training data, from human annotation

class	cata	num	gen	same	dist	role	par	form
TRUE	f	t	t	f	1	subj	f	prop
FALSE	f	t	t	t	0	subj	f	prop
FALSE	t	t	t	f	0	subj	f	pron
FALSE	f	t	t	f	1	subj	t	prop
TRUE	f	t	t	f	0	subj	t	prop
FALSE	f	t	t	f	0	obj	f	pron

Naive Bayes Classifier

Choose most probable class given a feature vector $\vec{f}$:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|\vec{f})$$

Apply Bayes Theorem:

$$P(c|\vec{f}) = \frac{P(\vec{f}|c)P(c)}{P(\vec{f})}$$

Constant denominator:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(\vec{f}|c)P(c)$$

Independent feature assumption ('naive'):

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i=1}^{n} P(f_i|c)$$

Problems with simple classification model

- Cannot implement ‘repeated mention’ effect.
- Cannot use information from previous links:

  Sturt think they can perform better in Twenty20 cricket. It requires additional skills compared with older forms of the limited over game.

  it should refer to Twenty20 cricket, but looked at in isolation could get resolved to *Sturt*. If linkage between *they* and *Sturt*, then number agreement is pl.

Not really pairwise: really need discourse model with real world entities corresponding to clusters of referring expressions.

Evaluation

- simple approach is link accuracy, i.e. percentage of correct links.

But:

- Identification of non-pleonastic pronouns and antecendent NPs should be part of the evaluation.
- Binary linkages don't allow for chains:

  Sally met Andrew in town and took him to the new restaurant. He was impressed.

Multiple evaluation metrics exist because of such problems.

Classification in NLP

- Also sentiment classification, word sense disambiguation and many others. POS tagging (sequences).
- Feature sets vary in complexity and processing needed to obtain features.
- Statistical classifier allows some robustness to imperfect feature determination.
- Acquiring training data is expensive.
- Few hard rules for selecting a classifier: e.g., Naive Bayes often works even when independence assumption is clearly wrong (as with pronouns).
- Experimentation, e.g., with WEKA toolkit.