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 (catching up)

Discourse structure

Coherence

Referring expressions and anaphora

Algorithms for anaphora resolution

Compositional distributional semantics

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

Two kinds of models:

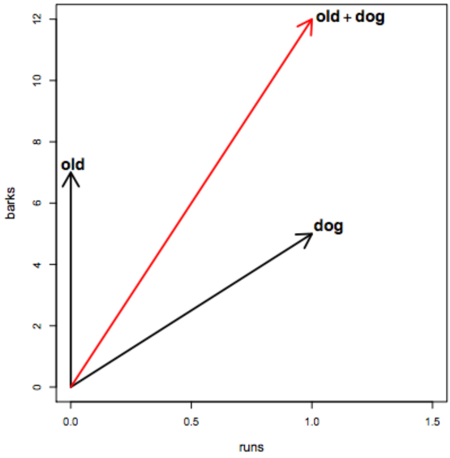
1. Vector mixture models
2. Lexical function models

1. Vector mixture models

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

Models:

- Additive
- Multiplicative



Additive and multiplicative models

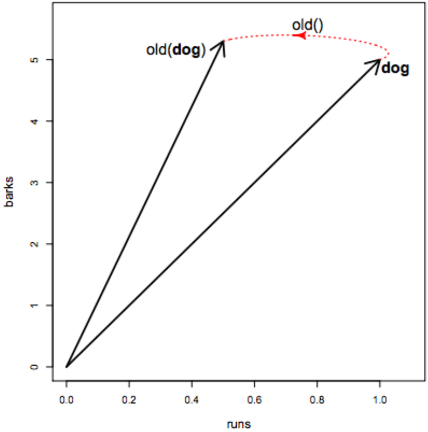
				additive		multiplicative	
	dog	**cat**	**old**	**old + dog**	**old + cat**	**old $\odot$ dog**	**old $\odot$ cat**
runs	1	4	0	1	4	0	0
barks	5	0	7	12	7	35	0

- 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

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

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

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	$(0.3 \times 1) + (5 \times 1)$ $= 5.3$

Learning adjective matrices

1. Obtain a distributional vector $\mathbf{n}_j$ for each noun n_j in the lexicon.
2. Collect adjective noun pairs (a_i, n_j) from the corpus.
3. Obtain a distributional vector $\mathbf{p}_{ij}$ of each pair (a_i, n_j) from the same corpus using a conventional DSM.
4. The set of tuples $\{(\mathbf{n}_j, \mathbf{p}_{ij})\}_j$ represents a dataset $\mathcal{D}(a_i)$ for the adjective a_i.
5. Learn matrix $\mathbf{A}_i$ from $\mathcal{D}(a_i)$ using linear regression.

Minimize the squared error loss:

$$L(\mathbf{A}_i) = \sum_{j \in \mathcal{D}(a_i)} \|\mathbf{p}_{ij} - \mathbf{A}_i \mathbf{n}_j\|^2$$

Polysemy in lexical function models

Generally:

- use single representation for all senses
- assume that ambiguity can be handled as long as contextual information is available

Exceptions:

- Kartsaklis and Sadrzadeh (2013): homonymy poses problems and is better handled with prior disambiguation
- Gutierrez et al (2016): literal and metaphorical senses better handled by separate models
- However, this is still an open research question.

Outline

- Compositional distributional semantics (catching up)
- Discourse structure
- Coherence
- Referring expressions and anaphora
- Algorithms for anaphora resolution

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

Rhetorical relations

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

- nucleus (the main phrase) and satellite (the subsidiary phrase: e.g., EXPLANATION, JUSTIFICATION

  Max fell because John pushed him.

- equal weight: e.g., NARRATION

  Max fell and Kim kept running.

Rhetorical relations

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

- nucleus (the main phrase) and satellite (the subsidiary phrase: e.g., EXPLANATION, JUSTIFICATION

  Max fell because John pushed him.

- equal weight: e.g., NARRATION

  Max fell and Kim kept running.

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 interpretation

Discourse coherence assumptions can affect interpretation:

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.

Discourse parsing: hard problem, but ‘surfacy techniques’ (punctuation and cue phrases) work to some extent.

Outline

- Compositional distributional semantics (catching up)
- Discourse structure
- Coherence
- Referring expressions and anaphora
- Algorithms for anaphora resolution

Co-reference and 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.

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:

- I don't know who the new teacher will be, but I'm sure they'll make changes to the course.
- The team played really well, but now they are all very tired.
- Kim and Sandy are asleep: they are very tired.

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

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.

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 $\vec{f}$:

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

Apply Bayes Theorem:

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

Constant denominator:

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

Independent feature assumption ('naive'):

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

Problems with simple classification model

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

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

Evaluation

- 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.