# Natural Language Processing: Part II Overview of Natural Language Processing (L90): ACS

## Lecture 10: Discourse

Simone Teufel (Materials by Ann Copestake)

Computer Laboratory
University of Cambridge

October 2018

# Outline of today's lecture

Putting sentences together (in text).

Coherence

Anaphora (pronouns etc)

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

# Lecture 10: Discourse

Coherence

Anaphora (pronouns etc)

Algorithms for anaphora resolution

# 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.
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 summarization

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

- nucleus and satellite: e.g., EXPLANATION, JUSTIFICATION
- equal weight: e.g., NARRATION

Max fell because John pushed him.

# Rhetorical relations and summarization

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

- nucleus and satellite: e.g., EXPLANATION, JUSTIFICATION
- equal weight: e.g., NARRATION

Max fell because John pushed him.

# 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.

# Lecture 10: Discourse

Coherence

Anaphora (pronouns etc)

Algorithms for anaphora resolution

# 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'

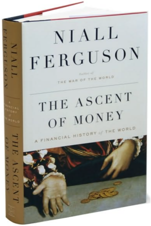
**antecedent** the text initially evoking a referent. 'Niall Ferguson'

**anaphora** the phenomenon of referring to an antecedent.

# Niall Ferguson and Stephen Moss. . .

**Niall Ferguson** is a British historian and conservative political commentator. He is a senior research fellow at Jesus College, Oxford. He is the bestselling author of several books, including The Ascent of Money.


**Stephen Moss** is a feature writer at the Guardian.




**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**

**Guardian Unlimited Education**

I think I may hate Niall Ferguson. He is 36, extremely good-looking, has just been given a chair in history at Oxford, and today publishes a voluminous history of money and power over three centuries, less than three years after his equally substantial revisionist history of the first world war. Ferguson received a £500,000 advance for a three-book deal with Penguin, yet this new book isn't even part of that. He had a year's sabbatical at the Bank of England and did the research there. Has he never heard of writer's block? Frankly, his apparent lack of suffering is insufferable.

He greets me warmly in his immaculate room at Jesus College, Oxford, and abandons his humming laptop to make me coffee. He is beautifully dressed in perfectly tailored suit and mauve shirt. I can't help notice his matching mauve cufflinks. Now, I'm sorry, but it is a well-known fact that dons must wear baggy pullovers and corduroy jackets, preferably patched at the elbow. They should give tutorials while in the bath or perched on the branch of a tree, and publish occasional, esoteric articles on the prevalence of the Black Death in Bolton or Henry VIII's flute sonatas. How on earth did this fellow, with his gorgeous clothes, vast books, frequent forays into journalism and outrageous productivity, ever get in?

The worst thing is, he is extremely likeable. I have come armed with a

# Pronoun resolution

Pronouns: a type of anaphor.
Pronoun 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

Pronouns: a type of anaphor.
Pronoun 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.

# Hard constraints: Pronoun agreement

Pronouns must agree with their antecedents in number and gender.
BUT:

- 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** 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 hat.

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

**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** (mentioned above)

# World knowledge

Sometimes inference will override soft preferences:

Andrew Strauss again blamed the batting after England lost to Australia last night. They now lead the series three-nil.

*they* is Australia.

But a discourse can be odd if strong salience effects are violated:

The England football team won last night. Scotland lost.
? They have qualified for the World Cup with a 100% record.

# World knowledge

Sometimes inference will override soft preferences:

Andrew Strauss again blamed the batting after England lost to Australia last night. They now lead the series three-nil.

*they* is Australia.
But a discourse can be odd if strong salience effects are violated:

The England football team won last night. Scotland lost.
? They have qualified for the World Cup with a 100% record.

# Lecture 10: Discourse

Coherence

Anaphora (pronouns etc)

Algorithms for anaphora resolution

# Anaphora resolution as supervised classification

- Classification: training data labelled with class and features, derive class for test data based on features.
- For potential pronoun/antecedent pairings, class is TRUE/FALSE.
- Assume candidate antecedents are all NPs in current sentence and preceeding 5 sentences (excluding pleonastic pronouns)

# Example

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

Issues: detecting pleonastic pronouns and predicative NPs, deciding on treatment of possessives (*the historian* and *the historian's Oxford study*), named entities (e.g., *Stephen Moss*, not *Stephen* and *Moss*), allowing for cataphora, . . .

# Example

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.

Issues: detecting pleonastic pronouns and predicative NPs, deciding on treatment of possessives (*the historian* and *the historian's Oxford study*), named entities (e.g., *Stephen Moss*, not *Stephen* and *Moss*), allowing for cataphora, . . .

# 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 $\vec{f}$:

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

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:

  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. Assume the data is previously marked-up with pronouns and possible antecedents, each pronoun is linked to an antecedent, measure percentage correct. 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.