Discourse

Putting sentences together (in text).

This lecture:

1. Relationships between sentences (and sentence-like clauses within sentences)
2. Coherence
3. Anaphora (pronouns etc)
4. An algorithm for anaphora resolution

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*

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

Strategic generation: constructing the logical form. Tactical generation: logical form to string.

Strategic 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%.

So far this has only been attempted for limited domains: e.g. tutorial dialogues.

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

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.

This can be exploited for text shortening:

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.

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.

referring expressions bits of language used to perform reference by a speaker.

antecedant the text evoking a referent.

anaphora the phenomenon of referring to an antecedant

Pronouns: a type of anaphor.
Pronoun resolution: only consider cases which refer to antecedant noun phrases.

- hard constraints (e.g., agreement)
- soft preferences / salience (depend on discourse structure)

Pronoun agreement

(25) A little girl is at the door — see what she wants, please?

(26) My dog has hurt his foot — he is in a lot of pain.

(27) * My dog has hurt his foot — it is in a lot of pain.

Complications:

(31) The team played really well, but now they are all very tired.

(32) Kim and Sandy are asleep: they are very tired.

(33) Kim is snoring and Sandy can’t keep her eyes open: they are both exhausted.

Reflexives

(34) John$_i$ cut himself$_i$ shaving. (himself = John, subscript notation used to indicate this)

(35) # 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.

Pleonastic pronouns

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

(36) It is snowing

(37) It is not easy to think of good examples.

(38) It is obvious that Kim snores.

(39) It bothers Sandy that Kim snores.

Salience (soft preferences)

Recency

(41) Kim has a fast car. Sandy has an even faster one. Lee likes to drive it.

Grammatical role Subjects > objects > everything else:

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

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:

(44) Bill went with Fred to the Grafton Centre. Kim went with him to Lion Yard.

Him=Fred

Coherence effects (mentioned above)

Lappin and Leass's algorithm

Discourse model: referring NPs in equivalence classes with global salience value.

For example:

N	*Niall Ferguson*, *him*	435
S	*Stephen Moss*	310
H	*the historian*	100
O	*Oxford study*	100

For each sentence:

1. Divide by two the global salience factors
2. Identify referring NPs
3. Calculate global salience factors for each NP (see below)
4. Update the discourse model with the referents and their global salience scores.

5. For each pronoun:
   (a) Collect potential referents
   (b) Filter referents
   (c) Calculate the per pronoun adjustments for each referent (see below).
   (d) Select the referent with the highest salience value for its equivalence class plus its per-pronoun adjustment.
   (e) Add the pronoun into the equivalence class for that referent, and increment the salience factor.

Weights

Global salience factors:

recency	100	(current sentence)
subject	80	
existential	70	*there is a cat*
direct object	50	
indirect object	40	*give Sandy a present*
oblique complement	40	*put the cat on a mat*
non-embedded noun	80	
non-adverbial	50	

(effectively, embedded -80 and adverbial -50
but no negative weights)

Per pronoun salience factors:

cataphora	-175	pronoun before NP
same role	35	e.g., pronoun and NP both subject

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.

Discourse referents:

N	*Niall Ferguson*, *him*	435
S	*Stephen Moss*	310

N has score 155 + 280 ((subject + non-embedded + non-adverbial + recency)/2 + (direct object + non-embedded + non-adverbial + recency))

S has score 310 (subject + non-embedded + non-adverbial + recency) + same role per-pronoun 35

Add *he* to the discourse referent equivalence class.

N	*Niall Ferguson*, *him*, *he*	515

Anaphora for everyone

Modification of Lappin and Leass that doesn't require a parser.

1. POS tag input text (Lingsoft tagger)
2. Regular expressions to identify NPs (NP chunking), mark expletive *it*
3. Regular expressions for grammatical role
4. Text segmentation: don't cross document boundaries etc.
5. Heuristics for reflexives
6. Otherwise much as Lappin and Leass

Evaluation

1. LL quoted 86% (computer manuals), KB 75% (mix genres)
2. much less standardized than POS tagging: datasets, metrics
3. results are genre-dependent
4. replication is difficult