Natural Language Processing: Part II (Paper 10)
Overview of Natural Language Processing (L90): Part III/ACS

2018, 12 Lectures, Michaelmas Term (these notes, lecture 10)
October 6, 2018
Lectured by Simone Teufel (sht25@cl.cam.ac.uk), Paula Buttery (pjb48@cl.cam.ac.uk), Ann Copestake (aac@cl.cam.ac.uk)
http://www.cl.cam.ac.uk/users/{sht25,pjb48,aac10}/

10 Lecture 10: Discourse

The techniques we have seen in lectures 2–9 relate to the interpretation of words and individual sentences, but utterances are always understood in a particular context. Context-dependent situations include:

1. Referring expressions: pronouns, definite expressions etc.
2. Universe of discourse: *every dog barked*, doesn't mean every dog in the world but only every dog in some explicit or implicit contextual set.
3. Responses to questions, etc: only make sense in a context: *Who came to the party? Not Sandy.*
4. Implicit relationships between events: *Max fell. John pushed him* — the second sentence is (usually) understood as providing a causal explanation.

In the first part of this lecture, I give a brief overview of *rhetorical relations* which can be seen as structuring text at a level above the sentence.[1] I'll then go on to talk about one particular case of context-dependent interpretation — anaphor resolution.

10.1 Rhetorical relations and coherence

Consider the following discourse:

(1) Max fell. John pushed him.

This discourse can be interpreted in at least two ways:

(2) Max fell because John pushed him.

(3) Max fell and then John pushed him.

This is yet another form of ambiguity: there are two different interpretations for (1) but there is no syntactic or semantic ambiguity in the interpretation of the two individual sentences in it. There seems to be an implicit relationship between the two sentences in (1): a *discourse relation* or *rhetorical relation*. (I will use the terms interchangeably here, though different theories use different terminology, and rhetorical relation tends to refer to a more surfacy concept than discourse relation.) In (2) the link between the second and first part of the sentence is explicitly an explanation, while (3) is an explicit narration: *because* and *and then* are said to be *cue phrases*. Theories of discourse/rhetorical relations try to reify this intuition using link types such as *Explanation* and *Narration*.

[1] A related, but somewhat different notion, is used in modelling dialogues to link utterances together.

10.2 Coherence

Discourses have to have connectivity to be coherent:

(4) Kim got into her car. Sandy likes apples.

Both of these sentences make perfect sense in isolation, but taken together they are incoherent. Adding context can restore coherence:

(5) 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.

The second sentence can be interpreted as an explanation of the first. In many cases, this will also work if the context is known, even if it isn't expressed.

Language generation requires a way of implementing coherence. For example, consider a system that reports share prices. This might generate:

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

This is much less acceptable than a connected discourse:

> 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%.

Here *but* indicates a Contrast. Not much actual information has been added (assuming we know what sort of company Dell, HP and Safeway are), but the discourse is easier to follow.

Discourse coherence assumptions can affect interpretation:

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

If we interpret this as Explanation, then 'he' is most likely Bill. But if it is Justification (i.e., the speaker is providing evidence to justify the first sentence), then 'he' is John.

10.3 Factors influencing discourse interpretation

1. Cue phrases. These are sometimes unambiguous, but not usually. e.g. *and* is a cue phrase when used in sentential or VP conjunction.
2. Punctuation (or the way the sentence is said — intonation etc) and text structure. For instance, parenthetical information cannot be related to a main clause by Narration (it is generally Explanation), but a list is often interpreted as Narration:

   > Max fell (John pushed him) and Kim laughed.
   > Max fell, John pushed him and Kim laughed.

   Similarly, enumerated lists can indicate a form of narration.
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.

It should be clear that it is potentially very hard to identify rhetorical relations. In fact, recent research that simply uses cue phrases and punctuation is quite promising. This can be done by hand-coding a series of finite-state patterns, or by supervised learning.

10.4 Discourse structure and summarization

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. (The main phrase is sometimes called the *nucleus* and the subsidiary one is the *satellite*.) This can be exploited in summarization.

For instance, suppose we remove the satellites in the first three sentences of this subsection:

> 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.

Other relationships, such as Narration, give equal weight to both elements, so don't give any clues for summarization.

Rather than trying to find rhetorical relations for arbitrary text, genre-specific cues can be exploited, for instance for scientific texts. This allows more detailed summaries to be constructed. In the next lecture, I'll give an overview of an approach to summarization which exploits discourse structure and coherence in a somewhat different way.

10.5 Referring expressions

I'll now move on to talking about another form of discourse structure, specifically the link between referring expressions. The following example will be used to illustrate referring expressions and anaphora resolution:

> 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. (quote taken from the Guardian)

Some terminology:

referent a real world entity that some piece of text (or speech) refers to. e.g., the two people who are mentioned in this quote.

referring expressions bits of language used to perform reference by a speaker. In, the paragraph above, *Niall Ferguson*, *him* and *the historian* are all being used to refer to the same person (they *corefer*).

antecedent the text initially evoking a referent. *Niall Ferguson* is the antecedent of *him* and *the historian*

anaphora the phenomenon of referring to an antecedent: *him* and *the historian* are *anaphoric* because they refer to a previously introduced entity.

What about *a snappy dresser*? Traditionally, this would be described as predicative: that is, it is a property of some entity (similar to adjectival behaviour) rather than being a referring expression itself.

Generally, entities are introduced in a discourse (technically, *evoked*) by indefinite noun phrases or proper names. Demonstratives (e.g., *this*) and pronouns are generally anaphoric. Definite noun phrases are often anaphoric (as above), but often used to bring a mutually known and uniquely identifiable entity into the current discourse. e.g., *the president of the US*.

Sometimes, pronouns appear before their referents are introduced by a proper name or definite description: this is *cataphora*. E.g., at the start of a discourse:

> Although she couldn't see any dogs, Kim was sure she'd heard barking.

both cases of *she* refer to Kim - the first is a *cataphor*.

10.6 Pronoun agreement

Pronouns generally have to agree in number and gender with their antecedents. In cases where there's a choice of pronoun, such as *he*/*she*/*they* or *it* for an animal (or a baby, in some dialects), then the choice has to be consistent.

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

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

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

Things to consider include: *they* when referring to a single individual (including 'gender-neutral they'), use of *they* with *everybody*, group nouns, conjunctions and discontinuous sets:

(9) Somebody's at the door — see what they want, will you?

(10) I don't know who the new teacher will be, but I'm sure they'll make changes to the course.[2]

(11) Everybody's coming to the party, aren't they?

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

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

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

10.7 Reflexives

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

(16) # John$_i$ cut him$_j$ shaving. (i $\neq$ j — a very odd sentence)

The informal and not fully adequate generalisation is that reflexive pronouns must be co-referential with a preceding argument of the same verb (i.e., something it subcategorises for), while non-reflexive pronouns cannot be. In linguistics, the study of inter-sentential anaphora is known as *binding theory*:

10.8 Pleonastic pronouns

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

(17) It is snowing

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

(19) It is obvious that Kim snores.

(20) It bothers Sandy that Kim snores.

Note also:

(21) They are digging up the street again

This is an (informal) use of *they* which, though probably not technically pleonastic, doesn't apparently refer in the standard way (they = 'the authorities'??).

[2]This is now standard usage: the use of the masculine pronoun (*he* etc) with indefinite reference is no longer generally acceptable.

10.9 Salience

There are a number of effects related to the structure of the discourse which cause particular pronoun antecedents to be preferred, after all the hard constraints discussed above are taken into consideration.

Recency More recent antecedents are preferred. Only relatively recently referred to entities are accessible.

(22) Kim has a big car. Sandy has a small one. Lee likes to drive it.

it preferentially refers to Sandy's car, rather than Kim's.

Grammatical role Subjects > objects > everything else:

(23) Fred went to the Grafton Centre with Bill. He bought a hat.

he is more likely to be interpreted as Fred than as Bill.

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

(24) Fred was getting bored. He decided to go shopping. Bill went to the Grafton Centre with Fred. He bought a hat.

He=Fred (maybe) despite the general preference for subjects.

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

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

Him=Fred, because the parallel interpretation is preferred.

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

(26) Bill likes Fred. He has a great sense of humour.

He = Fred preferentially, possibly because the second sentence is interpreted as an explanation of the first, and having a sense of humour is seen as a reason to like someone.

10.10 Lexical semantics and world knowledge effects

The made-up examples above were chosen so that the meaning of the utterance did not determine the way the pronoun was resolved. In real examples, world knowledge may override salience effects. For instance (from Radio 5):

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

Here *they* has to refer to Australia, despite the general preference for subjects as antecedents. The analysis required to work this out is actually non-trivial: you might like to try writing down some plausible meaning postulates which would block the inference that *they* refers to England. (Note also the plural pronoun with singular antecedent, which is normal for sports teams, in British English at least.)

Note, however, that violation of salience effects can easily lead to an odd discourse:

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

Systems which output natural language discourses, such as summarization systems, have to keep track of anaphora to avoid such problems.

10.11 Algorithms for resolving anaphora

NLP researchers are interested in all types of coreference, but most work has gone into the problem of finding antecedents for pronouns. As well as discourse understanding, this is often important in MT. For instance, English *it* usually has to be resolved to produce a high-quality translation into German because German has grammatical gender (although if all the candidate antecedents have the same gender, we don't need to do any further resolution). I will outline an approach to anaphora resolution using a statistical classifier, but there are many other approaches.

We can formulate pronoun resolution as a classification problem, which can be implemented using one of the standard machine learning approaches to supervised classification (examples of approaches include Naive Bayes, perceptron, k-nearest neighbour), assuming that we have a suitable set of training data. For each pairing of a (non-pleonastic) pronoun and a candidate antecedent, the classifier has to make a binary decision as to whether the candidate is an actual antecedent, based on some features associated with the pairing. For simplicity, we can assume that the candidate antecedents for a pronoun are all the noun phrases within a window of the surrounding text consisting of the current sentence and the preceding 5 sentences (excluding pleonastic pronouns). For 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.
>
> Pronoun *he*, candidate antecedents: *Niall Ferguson*, *a snappy dresser*, *Stephen Moss*, *him*, *an hour*, *the historian*, *the historian's Oxford study*.

Notice that this simple approach leads to *a snappy dresser* being included as a candidate antecedent and that a choice had to be made as to how to treat the possessive. I've included the possibility of cataphors, although these are sufficiently rare that they are often excluded.

For each such pairing, we build a *feature vector*[3] using features corresponding to some of the factors discussed in the previous sections. For instance (using t/f rather than 1/0 for binary features for readability):

Cataphoric Binary: t if the pronoun occurs before the candidate antecedent.

Number agreement Binary: t if the pronoun agrees in number with the candidate antecedent.

Gender agreement Binary: t if the pronoun agrees in gender with the candidate antecedent.

Same verb Binary: t if the pronoun and the candidate antecedent are arguments of the same verb (for binding theory).

Sentence distance Discrete: { 0, 1, 2 ... } The number of sentences between pronoun and candidate.

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 } This indicates something about the syntax of the potential antecedent noun phrase.

Taking some pairings from the example above:

pronoun	antecedent	cataphoric	num	gen	same	distance	role	parallel	form
him	*Niall Ferguson*	f	t	t	f	1	subj	f	prop
him	*Stephen Moss*	f	t	t	t	0	subj	f	prop
him	*he*	t	t	t	f	0	subj	f	pron
he	*Niall Ferguson*	f	t	t	f	1	subj	t	prop
he	*Stephen Moss*	f	t	t	f	0	subj	t	prop
he	*him*	f	t	t	f	0	obj	f	pron

[3]The term 'instance' is sometimes used in AI, but I prefer 'feature vector', because we're mainly interested in the nature of the features.

Notice that with this set of features, we cannot model the "repeated mention" effect mentioned in §10.9. It would be possible to model it with a classifier-based system, but it requires that we keep track of the coreferences that have been assigned and thus that we maintain a model of the discourse as individual pronouns are resolved. I will return to the issue of discourse models below. Coherence effects are very complex to model and world knowledge effects are indefinitely difficult (AI-complete in the limit), so both of these are excluded from this simple feature set. Realistic systems use many more features and values than shown here and can approximate some partial world knowledge via classification of named entities, for instance.

To implement the classifier, we require some knowledge of syntactic structure, but not necessarily full parsing. We could approximately determine noun phrases and grammatical role by means of a series of regular expressions over POS-tagged data instead of using a full parser. Even if a full syntactic parser is available, it may be necessary to augment it with special purpose rules to detect pleonastic pronouns.

The training data for this task is produced from a corpus which is marked up by humans with pairings between pronouns and antecedent phrases. The classifier uses the marked-up pairings as positive examples (class TRUE), and all other possible pairings between the pronoun and candidate antecendant as negative examples (class FALSE). For instance, if the pairings above were used as training data, we would have:

class	cataphoric	num	gen	same	distance	role	parallel	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

Note the pre-lecture exercise which suggests that you participate in an online experiment to collect training data. If you do this, you will discover a number of complexities that I have ignored in this account.

In very general terms, a supervised classifier uses the training data to determine an appropriate mapping (i.e., *hypothesis* in the terminology used in the Part 1B AI course) from feature vectors to classes. This mapping is then used when classifying the test data. To make this more concrete, if we are using a probabilistic approach, we want to choose the class c out of the set of classes C ($\{$ TRUE, FALSE $\}$ here) which is most probable given a feature vector $\vec{f}$:

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

(See lecture 3 for the explanation of argmax and $\hat{c}$.) As with the POS tagging problem, for a realistic feature space, we will be unable to model this directly. The Naive Bayes classifier is based on the assumption that we rewrite this formula using Bayes Theorem and then treat the features as conditionally independent (the independence assumption is the "naive" part). That is:

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

As with the models discussed in Lecture 3, we can ignore the denominator because it is constant, hence:

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

Treating the features as independent means taking the product of the probabilities of the individual features in $\vec{f}$ for the class:

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

In practice, the Naive Bayes model is often found to perform well even with a set of features that are clearly not independent.

There are fundamental limitations on performance caused by treating the problem as classification of individual pronoun-antecedent pairs rather than as building a discourse model including all the coreferences. Inability to implement 'repeated mention' is one such limitation, another is the inability to use information gained from one linkage in resolving further pronouns. Consider yet another 'team' example:

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

A classifier which treats each pronoun entirely separately might well end up resolving the *it* at the start of the second sentence to *Sturt* rather than the correct *Twenty20 cricket*. However, if we already know that *they* corefers with *Sturt*, coreference with *it* will be dispreferred because number agreement does not match (recall from §10.6 that pronoun agreement has to be consistent). This type of effect is especially relevant when general coreference resolution is considered. One approach is to run a simple classifier initially to acquire probabilities of links and to use those results as the input to a second system which clusters the entities to find an optimal solution. I will not discuss this further here, however.

10.12 Evaluation of pronoun resolution

At first sight it seems that we could require that every (non-pleonastic) pronoun is linked to an antecedent, and just measure the accuracy of the links found compared to the test data. One issue which complicates this concerns the identification of the pronouns (some may be pleonastic, others may refer to concepts which aren't expressed in the text as noun phrases) and also identification of the target noun phrases, with embedded noun phrases being a particular issue. We could treat this as a separate problem and assume we're given data with the non-pleonastic pronouns and the candidate antecedents identified, but this isn't fully realistic.

A further range of problems arise essentially because we are using the identification of some piece of text as an antecedent for the pronoun as a surrogate for the real problem, which is identification of references to real world entities. For instance, suppose that, in the example below, our algorithm links *him* to *Andrew* and also links *he* to *Andrew*, but the training data has linked *him* to *Andrew* and *he* to *him*.

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

Our algorithm has successfully linked the coreferring expressions, but if we consider the evaluation approach of comparing the individual links to the test material, it will be penalised. Of course it is trivial to take the transitive closure of the links, but it is not easy to develop an evaluation metric that correctly allows for this and does not, for example, unfairly reward algorithms that link all the pronouns together into one cluster. As a consequence of this sort of issue, it has been difficult to develop agreed metrics for evaluation.

10.13 Statistical classification in language processing

Many problems in natural language can be treated as classification problems: besides pronoun resolution, we have seen sentiment classification and word sense disambiguation, which are straightforward examples of classification. POS-tagging is also a form of classification, but there we take the tag sequence of highest probability rather than considering each tag separately. As we have seen above, we actually need to consider relationships between coreferences to model some discourse effects.

Pronoun resolution has a more complex feature set than the previous examples of classification that we've seen and determination of some of the features requires considerable processing, which is itself error prone. A statistical classifier is somewhat robust to this, assuming that the training data features have been assigned by the same mechanism as used in the test system. For example, if the grammatical role assignment is unreliable, the weight assigned to that feature might be less than if it were perfect.

One serious disadvantage of supervised classification is reliance on training data, which is often expensive and difficult to obtain and may not generalise across domains. Research on unsupervised methods is therefore popular.

There are no hard and fast rules for choosing which statistical approach to classification to use on a given task. Many NLP researchers are only interested in classifiers as tools for investigating problems: they may either simply use the same classifier that previous researchers have tried or experiment with a range of classifiers using a toolkit such as WEKA.[4]

[4] `http://www.cs.waikato.ac.nz/ml/weka/` Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Performance considerations may involve speed as well as accuracy: if a lot of training data is available, then a classifier with faster performance in the training phase may enable one to use more of the available data. The research issues in developing a classifier-based algorithm for an NLP problem generally center around specification of the problem, development of the labelling scheme and determination of the feature set to be used.

10.14 Further reading

J&M discuss the most popular approach to rhetorical relations, *rhetorical structure theory* or RST (section 21.2.1). I haven't discussed it in detail here, partly because I find the theory very unclear: attempts to annotate text using RST approaches tend not to yield good interannotator agreement (see comments on evaluation in lecture 3), although to be fair, this is a problem with all approaches to rhetorical relations. The discussion of the factors influencing anaphora resolution and the description of the classifier approach that I've given here are partly based on J&M's account in Chapter 21: they discuss a log-linear classifier there, but Naive Bayes is described in 20.2.2 and I have followed that description.