ANLP 2016
Lecture 24: (Co-)reference

Henry S. Thompson
With input from Johanna Moore and, indirectly, Bonnie Webber and Adrian Brasoveanu
10 November 2016

1. Introducing coreference

Consider the following narrative:

> Violent strikes rocked Happyland. A spokesman for the country's Department of Peace said they would meet with the strikers tomorrow. Another spokesman said that this was intended to demonstrate the country's commitment to resolving the dispute.

It contains something like 12 noun phrases

- But there are only 7 entities being referred to

So there must be some shared referents

2. Some terminology

Referring expression
A part of an utterance used to identify or introduce an entity
- in(to) a discourse (in real language use)
- in(to) a discourse **model** (in a theory or implementation)

Referents
are such entities
- in a **model**
- or (imagined to be) in the world

Reference
is the relation between a **referring expression** and a **referent**

Coreference
When more than one **referring expression** is used to refer to the same entity

Anaphora
Reference to, or depending on, a previously introduced entity

3. Definite reference

> Violent strikes rocked Happyland. A spokesman for the country's Department of Peace said they would meet with the strikers tomorrow. Another spokesman said that this was intended to demonstrate the country's commitment to resolving the dispute.

NP's in red above are **definite** referring expressions

- as are the country and the dispute

Their use presupposes the existence of a unique (and uniquely identifiable) referent

- If a hearer does not already know of such a referent
    - Or, we will sometimes say, have such a referent in their **discourse model**
- They will usually accept the assumption that there *is* one
    - Or, we might say, add one to their model
- provided it's not inconsistent to do so
- This is known as **accommodation**

Referring expressions can be embedded in other **referring expressions**

- A spokesman for the country's Department of Peace
- the country's Department of Peace

4. Anaphora

> Violent strikes rocked Happyland. A spokesman for the country's Department of Peace said they would meet with the strikers tomorrow. Another spokesman said that this was intended to demonstrate the country's commitment to resolving the dispute.

The country (twice), they, the strikers, another spokesman, this, the dispute are **anaphoric expressions**

- They rely on the previous discourse for their interpretation.
- So the country, the strikers and the dispute are *both* definite referring expressions and anaphoric expressions.

5. Indefinite Reference

Indefinite NPs usually introduce new referents to the discourse:

> A spokesman for the Department of Peace said he would meet with the strikers.

> Another spokesman said that this . . .

When an indefinite NP is in the scope of **propositional attitude** verbs

- e.g. *want*, *need*, *worry about*
- there is an ambiguity about its referent

Willard wants a sloop

- This might be a specific sloop that the speaker and/or Willard has in mind
- or an arbitrary, as yet unidentified, sloop
    - As Quine puts it, he is seeking "mere relief from slooplessness"

6. Coreference

Referring expressions with the same referent are said to **corefer**.

Indefinite NPs can set up referents for subsequent coreferential anaphoric expressions:

There are fairies$_i$ in my garden. The fairies$_i$/They$_i$ are having a ball

(Here and below we use subscripts to denote distinct *referents*)

- So the *same* subscript signals an instance of **coreference**

Not *all* fairies - just the ones in my garden

Not all subsequent anaphoric expressions corefer with their antecendent:

There are fairies$_i$ in my garden$_m$. Other fairies$_j$ live elsewhere$_k$

"other fairies" ≡ fairies other than fairies$_i$

- i.e., the ones in my garden$_m$

"elsewhere" ≡ places other than my garden$_m$

7. Coreference and Pronouns

Pronouns serve as **anaphoric expressions** when they rely on the previous discourse for their interpretation

Definite pronouns
He, *she*, *it*, *they* etc.

Indefinite pronouns
One, *some*, *elsewhere*, *others* etc.

- As in "Some survived the fall, but one broke"

Some pronouns have other roles as well:

- periphrastic *it*: "It is raining", "It is surprising that you ate a banana"
- generic *they* and *one*: "They'll get you for that", "One doesn't do that sort of thing in public"

And some determiners have an anaphoric role:

- *some*, *another*, *other* etc.
- as in *other fairies*

The expression from the previous discourse used in interpreting a pronoun used anaphorically is called its **antecedent**.

A definite pronoun corefers with its antecedent.

The antecedent of an indefinite pronoun contributes in a more oblique way.

8. Reference resolution

Reference resolution is the process of determining the **referent** of a **referring expression**

- Whether by humans or machines

Context obviously plays a crucial role in reference resolution

Situational
: The real-world surroundings (physical and temporal) for the discourse

Mental
: The knowledge/beliefs of the participants

Discourse
: What has been communicated so far

9. Discourse context—discourse model

For people we assume

- and in a computational system we construct

a **discourse model**

- That is, a set of "representations of the entities that have been referred to in the discourse and the relationships in which they participate" (J&M 21.3)

To produce and interpret referring expressions, a system must have methods for

- constructing a discourse model that evolves dynamically
- mapping between referring expressions and referents
    - Strictly speaking, via a *representation* (in a model) of their (real word) referents

In other words, for each referring expression, it must be able to determine when to

- *Add* a new entity the model to serve as the expression's referent
    - J&M call this **evoking**
- *Find* an existing entity ditto
    - J&M call this **accessing**

10. Implementing reference resolution

Most approaches to implementing reference resolution distinguish two stages:

1. Filter the set of possible referents by appeal to linguistic constraints
2. Rank the resulting candidates based on some set of heuristics

11. Constraints on pronouns: Feature agreement

English pronouns agree with the **number** and/or **gender** of the *referent* of their antecedent.

- *Robin has a new car. It/*She/*They is red*
- *Robin has a sister. *It/She/*They/*We is well-read*
- *Robin has three cars. *It/*She/They/*We are all red*

As well as the **person** (but **case** is determined locally):

- *Robin and I/*me were late. *Me/*They/We/I missed the show*
- *Robin and I/*me were late. The usher wouldn't let *we/*I/us/me in*

French pronouns agree with the number and gender of the *form* of their antecedent

- *Voici une pomme. Je me demande si elle/*il/*elles est mûre* [feminine form]
- *Here's an apple. I wonder if *she/it/*they is ripe* [inanimate/neuter referent]

12. Constraints on pronouns: Syntax

All anaphors, including pronouns, rely on the previous text for all or part of their interpretation.

When the text is in the same sentence, pronominal coreference is subject to **binding conditions**

- *John likes him* vs. *John likes himself*
- *John thinks Mary likes him/herself* vs. **John thinks Mary likes himself*
- *Her brother admires Mary* ⇒ Whose brother?
    - One one reading, an example of **cataphora**

And, sometimes, to **selectional restrictions** based on the verb that **governs** it

John parked his car in the garage. He had driven it around for hours

- it = the car, it ≠ garage

I picked up the book and sat in a chair. It broke

- it = chair, it ≠ book

We will see how automated approaches to anaphora resolution exploit such constraints

13. Constraints aren't enough

The kind of strong constraints we've just seen are not always enough to reduce the candidate set for resolution to a single entity

- *John punched Bill. He broke his jaw/hand*
- *Obama hates her husband, but Clinton worked for/stays married to him anyway*

14. Heuristics for pronoun interpretation

Many different features influence how a listener will resolve a definite pronoun (i.e., what they will take to be its antecedent):

Recency

The most recently introduced entity is a better candidate

- *First Robin bought a phone, and then a tablet. Kim is always borrowing it*

Grammatical role

- some grammatical roles (e.g. **SUBJECT**) are felt to be more **salient** than others (e.g., **OBJECT**)

- *Bill went to the pub with John. He bought the first round*
- "John" is more recent, but "Bill" is more salient.

15. More heuristics

Repeated mention

A repeatedly-mentioned entity is likely to be mentioned again

- John needed portable web access for his new job. He decided he wanted something classy. Bill went to the Apple store with him. He bought an iPad.

- "Bill" is the previous subject, but "John"'s repeated mentions tips the balance.

Parallelism

Parallel syntactic constructs can create an expectation of coreference in parallel positions

- *Susan went with Alice to the cinema. Carol went with her to the pub*

16. Heuristics, concluded

Verb semantics

A verb may serve to foreground one of its argument positions for subsequent reference because of its semantics

*John criticised **Bill** after **he** broke **his** promise*

vs.

***John** telephoned Bill after **he** broke **his** promise*

***Louise** apologised to Sandra for **her** response*

vs.

> *Louise praised **Sandra** for **her** response*

World knowledge

At the end of the day, sometimes only one reading *makes sense*:

> ***The city council** denied the demonstrators a permit because **they** feared violence*

vs.

> *The city council denied **the demonstrators** a permit because **they** advocated violence*

17. Coreference: A more general case

Anaphoric pronoun resolution is a specific instance of the more general problem of coreference resolution

- Definite expressions *other* than pronouns are also candidates for reference resolution

Some of the heuristics enumerated above are relevant for the general case

- Particularly for more generic phrases such as "the guy" or "your man"

But other relations also come in to play

Some are semantic

Hyponomy

That is, subclassing

- ***Megabuck PLC** announces its 3rd quarter results today. **The company** is expected . . .*

Meronymy

That is, part-whole relations

- *I had to take **my car** into the shop today. **The brakes** were squeeking really badly*

And some more superficial

- As simple as textual similarity
  - ***Victoria Chen** opened the meeting. **Dr. Chen** is . . .*
- Or requiring some specialised processing
  - ***2016** results are up 15%. Trading conditions **this year** . . .*

18. Automatic methods

There is a rich history of automatic definite reference and pronoun resolution systems

- Initially rule-based
- More recently using machine learning

One particular type of approach fits well with what we've already looked at in this course

- Feature-based systems using logistic regression

The problem is viewed as a simple binomial classification task

- For every pair of referring expressions
- Are they coreferential, or not?

As indicated above, different features are appropriate, or at least will be differently weighted, for the general coreference case and the more specific pronominal case

19. Using a gold standard

Given an corpus annotated for coreference, to train a model we simply

- Given an NP_i that is known to co-refer with NP_j where NP_j is the closest such (typically preceding) NP, create a *positive* training instance **(NP_i, NP_j)**
  - To be understood as identifying NP_j as the antecedent of NP_i
- For all NPs *between* NP_j and NP_i, create a *negative* training instance**(NP_i, NP_{j+1})**, **(NP_i, NP_{j+2})**, etc.

Tabulate the value of likely candidate features

- the nature of NP_i and NP_j: pronouns, definite NPs, demonstrative NPs (*this/that/these/those* X), proper names;
- distance betwen NP_i and NP_j: 0 if same sentence, 1 if adjacent sentence, etc.;
- whether NP_i and NP_j agree in number;
- whether NP_i and NP_j agree in gender;
- whether their semantic classes are in agreement;
- Edit distance between NP_i and NP_j;
- Etc.

Use logistic regression (see Lecture 20) to train weights for the positive and negative cases.

Note: Following J&M, the subscripts simply index the *position* of NPs

- This is a change from their use earlier in the lecture
- Where the identify/difference of subscript was used to notate coreference or its absence.

20. Using the model

Compute feature vectors for all possible referring expressions

For pronominal anaphora, we can just choose the most-positively scored (or largest positive vs. negative difference) antecedant

- Allowing for the possibility of no in-the-discourse coreferent at all

For definite referring expressions, choosing among available candidates versus not-in-the-discourse is a bit trickier

- And may require a separate model in its own right

21. Conclusion

There are usable "Off the shelf" coreference resolvers for English

- **emPronoun**, from Brown University, described in Charniak & Elsner, downloadable from `http://bllip.cs.brown.edu/papers/ec-eacl09.pdf`

- BART, from Johns Hopkins
- Deterministic Coreference Resolution System, from Stanford

There is still room for improvement in both coreference and anaphor resolution methods.

Knowing what expressions corefer and how other expressions relate to their antecedents can improve performance of Language Technology systems.