Discourse – Coreference Resolution

COMP-550

J&M Ch. 18.1 (1st); J&M Ch. 21.3–21.8 (2nd); J&M Ch. 23, 24 (3rd)

Outline

Discourse

- Coherence vs. cohesion

Coreference and anaphora

- Types of coreference phenomena

Hobbs algorithm

Machine learning for coreference resolution

Discourse

Language does not occur one sentence or utterance at a time.

Types of discourse:

Monologue – one-directional flow of communication

Dialogue – multiple participants

- Turn taking
- More varied communicative acts: asking and answering questions, making corrections, disagreements, etc.

Coherence

A property of a discourse that “makes sense” – there is some logical structure or meaning in the discourse that causes it to hang together.

Coherent:

Indoor climbing is a good form of exercise.

It gives you a whole-body workout.

Incoherent:

Indoor climbing is a good form of exercise.

Rabbits are cute and fluffy.

Cohesion

The use of linguistic devices to tie together text units

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Lexical Cohesion

Related words in a passage

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Coreference Chains

Anaphoric devices

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must also inform a citizen that a stop is voluntary and they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Discourse Markers

Cue words mark discourse relations

Ontario's Liberal government is proposing new regulations that would ban the random stopping of citizens by police.

The new rules say police officers cannot arbitrarily or randomly stop and question citizens.

Officers must **also** inform a citizen that a stop is voluntary **and** they have the right to walk away.

http://www.cbc.ca/news/canada/toronto/carding-regulations-ontario-1.3292277

Reference and Coreference

that cat
Whiskers
something furry
it

Referring expressions **Referent**

(a.k.a., mentions)

“*That cat*”, “*Whiskers*”, “*it*”, and any other expression that point to the same referent are said to **corefer**.

Anaphora and Antecedents

In a passage:

> **Maru** is a male Scottish Straight cat in Japan who has become popular on YouTube. …
>
> His owner is rarely seen in the videos. The videos include title cards in English and Japanese setting up and describing the events ...
>
> https://en.wikipedia.org/wiki/Maru_(cat)

An **anaphor** points to a *previous* linguistic expression, which is its **antecedent**.

Cataphora

Cataphors are anaphors that point to cats.

When <u>he</u>'s grumpy, ***Whiskers*** *refuses to eat.*

Just kidding! Actually, a **cataphor** points to an antecedent that *follows* it.

Types of Referring Expressions

Proper names

McGill University

Whiskers

Montreal

Pronouns

I

you

it

their

ours

herself

More Types of Refering Expressions

Noun phrases

Indefinite

Some water

A deer

This random dude (Note that *this* is ambiguous)

Definite

The cat

The election

Demonstratives (They point to something)

This/That hotdog

These problems

Cross-linguistically Speaking

Zero anaphora

- Many languages omit pronouns in certain contexts.
- Often called **pro-drop** languages
- Computational task: *detect* and *resolve* them

Sometimes, you can tell what pronoun is missing:

- e.g., Spanish:

No	*habl-o*	*español.*
NOT	Speak-1Sg	Spanish

(I) don’t speak Spanish.

Languages like this: Spanish, Italian, Russian, and many others

Omitting Pronouns

Other times, you really have to tell from the surrounding context.

- e.g., Japanese:

  ai shi -te- -ru.

  Love PROG PRES

  (I) love (you).

  But could also be *(He) loves (her). (They) love (me).*

  Languages like this: Japanese, Korean, Chinese varieties

This occasionally happens in informal English, usually in the first person.

Attended a dope COMP-550 class today.

Other Kinds of Reference

Bridging reference

Reference to entities that can be inferred from a previously mentioned entity

I like my office. <u>The windows</u> are large and <u>the table</u> is made of mahogany.

You should get a cactus. <u>They</u> are easy to care for.

Non-Referential Pronouns

Pleonastic pronouns

It is raining.

Snap out of it!

Clefting

It is COMP-550 which is giving me headaches.

- Used to put the focus on some point
- Seems marginally referential?

Event Coreference Resolution

s1: Hewlett-Packard is negotiating to **[buy]** technology services provider Electronic Data Systems.

s8: With a market value of about $115 billion, HP could easily use its own stock to finance the **[purchase]**.

s9: If the **[deal]** is completed, it would be HP's biggest **[acquisition]** since it *[bought]* Compaq Computer Corp. for $19 billion in 2002.

(Bejan and Harabagiu, 2010)

Event Coreference Resolution

What does it mean for events to corefer?

- Same causes and effects (Davidson, 1969)
- **Happen in same time and place** (Quine, 1985)

Cues for event coreference (Bejan and Harabagiu, 2010):

- Share same event properties
- Share same participants

Algorithms for Coreference Resolution

Let's focus on **pronominal anaphora resolution** (i.e., determining antecedents of pronouns).

Heuristics based on syntactic parses

- **Hobbs Algorithm (1978)**
- Lappin and Leass (1994)
- Approaches based on Centering Theory (Grosz et al., 1995)

Machine learning approaches

- **Handcrafted features (Soon et al., 2001)**
- **Neural models (Lee et al., 2017)**

Cues for Anaphora Resolution

Maru is a male Scottish Straight cat in Japan who has become popular on YouTube. …

His owner is rarely seen in the videos. The videos include title cards in English and Japanese setting up and describing the events ...

Relevant cues?

- Number and gender
- Recency
- Syntactic information (grammatical role information)

Syntactic Heuristics

For example:

The students taught themselves. *[themselves = the students]*

The students taught them. *[them ≠ the students]*

- Reflexives must be *bound* by a subject in a certain syntactic relationship in the same sentence.
- Personal pronouns *must not* be bound in this way.
- Formalized in a theory called **Binding Theory** (Chomsky, 1981)

Hobbs Algorithm (1978)

A traversal algorithm which requires:

- Constituent parse tree
- Morphological analysis of number and gender

Overall steps:

1. Search the current sentence right-to-left, starting at the pronoun
2. If no antecedent found, search previous sentence(s) left-to-right

Steps in Hobbs Algorithm In Detail

1. Begin at the NP node immediately dominating the pronoun.
2. Go up to the first NP or S above it. Call this node X and the path to it p.
3. Do a left-to-right breadth-first traversal of all branches below X to the left of p . Propose as antecedent any NP node encountered that has an NP or S between it and X.
4. If X is the highest S in the sentence, consider the parse trees of previous sentences in recency order, and traverse each in turn in left-to-right breadth-first order. When an NP is encountered, propose it as an antecedent. If X is not the highest S, continue to step 5
5. From X, go up to the first NP or S above it. Call this new node X and the path to it p.
6. If X is an NP and p doesn't pass through the Nominal that X immediately dominates, propose X as an antecedent.
7. Do a left-to-right breadth-first traversal of all branches below X to the left of p. Propose any NP encountered as the antecedent.
8. If X is an S, do a left-to-right breadth-first traversal of all branches below X to the right of p, but don't go below any NP or S encountered. Propose any NP encountered as the antecedent.
9. Go to step 4.

Example of Hobbs Algorithm

Alice saw a beautiful cupcake in the patisserie window.

She showed it to Bob.

She devoured it.

Assume a standard parse of the sentences of the type we have been drawing in this class.

Assume a perfect gender/entity type checker

Coreference Resolution by ML

Subtasks:

Mention detection

- Decide which spans of text are mentions/referring expressions and are anaphoric
- May be a separate step, or integrated into an end-to-end system

Coreference resolution

- Determine coreference links in passage

Before considering existing models, how would you design a coreference resolution system using the ML approaches we have discussed?

Handcrafted Features

Soon et al. (2001) defined 12 features for NP coreference resolution (not just pronominal):

Feature Type	Feature	Description
Lexical	SOON_STR	C if, after discarding determiners, the string denoting NP_i matches that of NP_j; else I.
Grammatical	PRONOUN_1*	Y if NP_i is a pronoun; else N.
	PRONOUN_2*	Y if NP_j is a pronoun; else N.
	DEFINITE_2	Y if NP_j starts with the word "the;" else N.
	DEMONSTRATIVE_2	Y if NP_j starts with a demonstrative such as "this," "that," "these," or "those;" else N.
	NUMBER*	C if the NP pair agree in number; I if they disagree; NA if number information for one or both NPs cannot be determined.
	GENDER*	C if the NP pair agree in gender; I if they disagree; NA if gender information for one or both NPs cannot be determined.
	BOTH_PROPER_NOUNS*	C if both NPs are proper names; NA if exactly one NP is a proper name; else I.
	APPOSITIVE*	C if the NPs are in an appositive relationship; else I.
Semantic	WNCLASS*	C if the NPs have the same WordNet semantic class; I if they don't; NA if the semantic class information for one or both NPs cannot be determined.
	ALIAS*	C if one NP is an alias of the other; else I.
Positional	SENTNUM*	Distance between the NPs in terms of the number of sentences.

Table from (Ng and Cardie, 2002)

Soon et al., 2001

They trained a supervised decision tree classifier using these features.

Results on MUC-6 data set:

58.6/67.3/62.6 in terms of R/P/F1

Ng and Cardie, (2002) extended the feature set.

62.4/73.5/67.5

Durrett and Klein (2013) incorporated many features into a log-linear model (~3M).

Word-level features + simple recency, syntax, and gender/number features actually work very well.

Neural Coreference Resolution

Model of Lee et al., (2017)

- End-to-end system (E2E)
- Two functions to learn:

  $s_m(i)$ score of span i being a mention

  $s_a(i, j)$ score of span j being antecedent of span i
- Need to score all possible span pairs
- Discard spans of text that are unlikely to be mentions, to make inference tractable (**pruning**)

Mention Detection

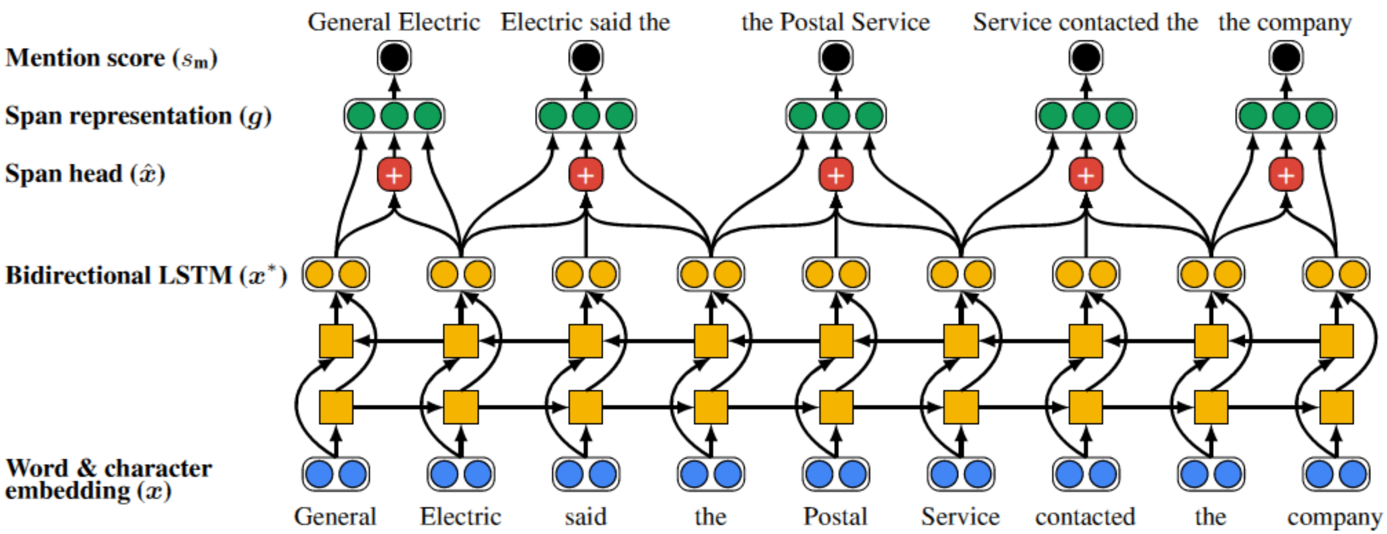

Figure 1: First step of the end-to-end coreference resolution model, which computes embedding representations of spans for scoring potential entity mentions. Low-scoring spans are pruned, so that only a manageable number of spans is considered for coreference decisions. In general, the model considers all possible spans up to a maximum width, but we depict here only a small subset.

Coreference Resolution

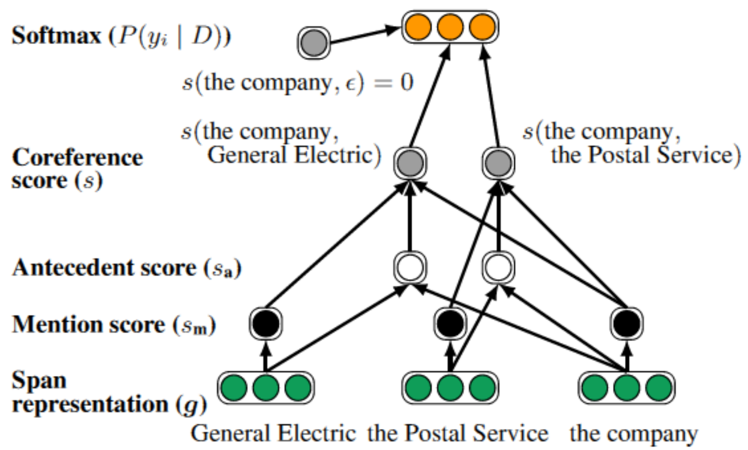

Figure 2: Second step of our model. Antecedent scores are computed from pairs of span representations. The final coreference score of a pair of spans is computed by summing the mention scores of both spans and their pairwise antecedent score.

Transformer-Based Systems

Modern systems are based on pre-trained LLMs. Two general classes:

- **Encoder models:** follows in the footsteps of E2E and frames coreference resolution as a classification problem
- **Decoder models**: frames coreference resolution as a sequence-to-sequence task

Transformer-Based Systems

Modern systems are based on pre-trained LLMs. Two general classes:

- **Encoder models:** follows in the footsteps of E2E and frames coreference resolution as a classification problem
- **Decoder models**: frames coreference resolution as a sequence-to-sequence task

We find that encoder models tend to be more efficient and perform well if LLM size and model is controlled for (Porada et al., 2024).

Link-Append Model

(Bohnet et al., 2023)

Input: *Speaker-A* I still have n't gone to that fresh French restaurant by your house
Prediction: SHIFT: next sentence

Input: *Speaker-A* I_2 still have n't gone to that fresh French restaurant by your house *Speaker-A* I_{17} 'm like dying to go there
Prediction:

A $\mathrm{I}_{17} \rightarrow \mathrm{I}_2$
B SHIFT: next sentence

Input: *Speaker-A* [1 I] still have n't gone to that fresh French restaurant by your house *Speaker-A* [1 I] 'm like dying to go there *Speaker-B* You mean the one right next to the apartment
Prediction:

A You → [1
B the apartment → your house
C the one right next to the apartment → that fresh French restaurant by your house
D SHIFT: next sentence

Input: *Speaker-A* [1 I] still have n't gone to [3 that fresh French restaurant by [2 your house]] *Speaker-A* [1 I] 'm like dying to go there *Speaker-B* [1 You] mean [3 the one right next to [2 the apartment]] *Speaker-B* yeah yeah yeah
Prediction: SHIFT: next sentence

Append: add a mention to a pre-existing cluster

Link: create a new cluster by linking two spans

Shift: done processing sentence

References (Others in J&M)

- Bejan and Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. *ACL*.
- Bohnet et al. 2023. Coreference Resolution through a seq2seq Transition-Based System. *TACL*.
- Durrett and Klein. 2013. Easy Victories and Uphill Battles in Coreference Resolution. *EMNLP*.
- Lee et al. End-to-end Neural Coreference Resolution. 2017. *EMNLP*.
- Ng and Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. *ACL*.
- Porada, Zou and Cheung A Controlled Reevaluation of Coreference Resolution Models. *LREC-COLING 2024*.
- Soon, Ng and Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*.