Empirical Methods in NLP

Coreference Resolution

Amir Zeldes
amir.zeldes@georgetown.edu
What is coreference?

- What spans are candidates?
- What counts as coreference?
- How to do this automatically?
Why coreference resolution?

General premise:
- Reduce ambiguity – every pronoun replaceable by lexical NP
- Theoretical models of discourse comprehension – how do humans know what we’re talking about?

In practice:
- enable information extraction (IR, summarization)
- better input data for entity sensitive tasks (e.g. MT)
- NLG / referring expression generation (QA)
What kind of task is this?

- **Step 1:**
  - Identify *referring expressions*

- **Step 2** – two variants:
  - A: Perform **clustering** into entities
  - B: Perform **linking** of anaphor-antecedent pairs

- **Assumptions:**
  - Referentiality is binary (referring/non-referring)
  - Clustering: coreference is transitive? $A \leftarrow B \leftarrow C \models A \leftarrow C$
  - Linking: coreference always ‘points backward’?
Mention Detection

- Naïve approach:
  - use a parser (which? Errors now inevitable?)
  - take all NPs - recall oriented (what about “on the other hand”?)

- Let’s try an easy one – how many mentions? Where do they start and end?

- New Zealand begins process to consider changing national flag design
Mention Detection

- Naïve approach:
  - use a parser (which? Errors now inevitable?)
  - take all NPs - recall oriented (what about “on the other hand”?)

- Harder example (OntoNotes corpus) – which NPs are referring expressions?
  - If [any part of [the matter]] were in [[my] hand], [no eye] would have read [it] and [no passerby] would have come across [it]
Mention Detection

- Are NPs enough? Where are the borders?
Mention Detection

- Parser input is still the most common approach
- But recently, considering and ranking *any* span of tokens up to length $k$ has been proposed (Lee et al. 2017)

**Advantage:**
- Possible to identify unusual spans from training data
- Potentially better at verb event coreference

**Disadvantage:**
- Possible spurious spans (hard to rule out ‘blunders’)  
- Can’t capitalize on larger training data for parsing
Coreference

- What phenomena should be included?
- Easy! Group cases of **same entity in the world**
- Pronominal NP anaphora: *Kim says she*....
- Lexical coref: *Aamir Khan ... This Indian actor* ...
- Apposition: *Shinzo Abe, The Japanese premier*
- Cataphora: *In her speech the chairwoman said*
- Event anaphora: *Ben visited Rome ... the visit*
- Sense anaphora: *Don’t you like beer? Yes, I’ll have one*
- Bridging: *Looking at Mexico, they said the economy* ...
Annotation schemes and consequences

- Guidelines and goals still debated
- Many discussions begin with the **ACE corpora** (Doddington et al. 2004)
- Current de facto standard: **OntoNotes** (Hovy et al. 2006)
- But many in between (e.g. ARRAU, Poesio & Artstein 2008; GUM, see discussion in Zeldes & Zhang 2016)
Biggest points of contention:

No antecedents for indefinites (BBN 2007, 4)

[Parents]x should be involved with their children's education at home, not in school. [They]x should see to it that [their]x kids don't play truant; [they]x should make certain that the children spend enough time doing homework; [they]x should scrutinize the report card. [Parents]y are too likely to blame schools for the educational limitations of [their]y children. If [parents]z are dissatisfied with a school, [they]z should have the option of switching to another.
No predicatives, no ‘as’ phrases:

\[\text{[George] was [the king] and was treated as [the monarch]}\]

Relations **should** be derivable from syntax but:

- Not all corpora have gold syntax
- ‘as’ can be ambiguous
- Negation, modality...

Sometimes counter-intuitive:

- \textit{It} was a beetle! (no markup whatsoever in OntoNotes)
- Milisanidis **scored** 9.687 ... \textit{It} was the best result for Greek gymnasts since they began taking part in gymnastic internationals. (markup, but only pronouns! Cf. Lee et al. 2013)
- Markup catches less interesting mention: What was best for Greek gymnasts?
OntoNotes - coordination

- No coordination envelope without aggregate mention:

\[\text{[The US] and [Japan]} \ldots \text{[The US] and [Japan]}\]

\[\text{[[The US] and [Japan]]} \ldots \text{[They]}\]

- Difficult for coreferencer to make local decision on coordinate mention
OntoNotes - apposition

- Apposition envelope:
  - A peculiarity of OntoNotes – appositions are a separate entity reference:

```
Emeritus Professor John Burrows, the chairman of the project's panel of twelve, said
New Zealand's flag has never before been open to public choice. Professor Burrows also
```
OntoNotes - i within i

- OntoNotes forbids nested mention coreference
  - *He has in tow [his prescient girlfriend, whose sassy retorts mark [her] ...]* (not annotated!)

- **But** external reference to embedded mentions is possible:
  - *[The American administration who planned carefully for this event through experts in media and public relations, and [its] tools] ... have caught [them] by surprise* (all three linked!)
Compound modifiers

- Only proper noun modifiers are included:
  - [Hong Kong] government ... [Hong Kong] (annotated)

- No annotation for:
  - small investors seem to be adapting to greater [stock market] volatility ... Glenn Britta ... says he is “factoring” [the market’s] volatility “into investment decisions.”

“Same entity in the world” is not so simple...
Antecedent detection

- After mention detection, check for every referring expression:
  - Given or new?
  - If new: singleton? (Recasens et al. 2013)
  - If given: what is the antecedent?

- Clustering approach: (Lee et al. 2013, Clark & Manning 2016)
  - Add best guess to cluster, recalculate next best guess

- Mention pair/ranking approach: (Durrett & Klein 2013, Lee et al. 2017)

- And in between (Wiseman et al. 2015, Zeldes & Zhang 2016)
Mention pair approach

- Apply binary classification ± anaphoricity ranking (Durrett & Klein 2013, Lee et al. 2017)
- Fast, simple (e.g. loglinear models)
- But: global chain constraints missed
Clustering approach

- Score all possible matches and make best decision first (e.g. Clark & Manning 2015)
- Share features for all clustered mentions
- What will happen here?
  - *Mr. Clinton ... Clinton ... Ms. Clinton ... she ... Clinton*

- And here?
  - *Georgetown is a University in DC. George Washington University, the closest university to it in the city, is also the largest in the District. Both universities offer undergraduate and graduate degrees.*
Candidate selection

  - Search Match And Select using Heuristic

- Basic idea, for each anaphor:
  - Search through all previous mentions
  - Perform feature matching (esp. morphological agreement: gender, number)
  - Discard incompatible mentions
  - Select best candidate (good baseline: most recent)
Sieve approach

- Used e.g. in CoreNLP d-coref (Lee et al. 2013)
Problems

- Precision oriented:
  - Notional agreement: [the New Zealand government] announced the start of a process to determine whether [their] citizens (Zeldes, to appear)
  - Verbal coreference/event anaphora
  - Uphill semantics battle (Durrett & Klein 2013)
    - Synonymy: [this novel idea] == [the new approach]
    - Antonymy: [the good news] != [the bad news]
    - Semantic compatibility: [the gold medalist] .. [this athlete]
    - World knowledge: [The Woman In The Window] ... [the recent New York Times bestseller]
HUGE knowledge bases exist (curated and scraped): DBPedia, FreeBase, Yago, ConceptNet, PPDB...

What do we need to know for coref?
Pure Machine Learning approaches

End-to-end corpus training (Lee et al. 2017)

- Top of the line because:
  - Consider ‘all possible features’?
  - Simple (but slow to train)
  - Best results on (homogeneous) test set

- But:
  - Large corpora unavailable for most languages
  - No way of integrating novel facts
  - Risk of overfitting style, period, other irrelevant properties
• NB: ALL spans up to length K, within sentence are considered
• LSTM learns sentence-wise
Syntactic heads are NOT explicitly learned
Attention mechanism learns something very similar

Hard to rule out...
Evaluation

- Reference scorer implemented by Pradhan et al. (2014)
- Three main metrics:
  - MUC (Vilain et al. 1995)
  - B3 (Bagga & Baldwin 1998)
  - CEAFε (Luo 2005)
Figure 1: Example key and response entities along with the partitions for computing the MUC score.
Scoring

- **MUC** – precision and recall for **links** in gold entities – link based
- **B3** – mention based – each mention in a gold entity gets credit based on ratio of correct mentions in its predicted entity
- **CEAFe** – entity based – calculate best scoring alignment of gold and predicted entities, then get proportion of correct and incorrect links in each entity
  - Other metrics: CEAFm, BLANC (Recasens & Hovy 2011, Luo et al. 2014)
Finding the culprit - p-link

- Use partitioned version of link-based score (Zeldes & Simonson 2016; extension of Martschat et al. 2015)
  - Each segment type accumulates credit (or blame)
  - Precision and recall in terms of correct link end points per partition

\[
p\text{-}link_{R,\pi} = \frac{\sum_{i=1}^{N_k} (|K_i^\pi| - p(K_i^\pi))}{\sum_{i=1}^{N_k} (|K_i^\pi| - 1)}
\]

\[
p\text{-}link_{P,\pi} = \frac{\sum_{i=1}^{N_f} (|R_i^\pi| - p'(R_i^\pi))}{\sum_{i=1}^{N_f} (|R_i^\pi| - 1)}
\]

Implementation available: [https://github.com/amir-zeldes/reference-coreference-scorers](https://github.com/amir-zeldes/reference-coreference-scorers)
Recent criticism

- Moosavi & Strube (2016) point out inconsistent behavior of metrics
- It is possible to construct cases where one metric improves while another degrades
- “Mean of three bad metrics does not make a good one”