SI485i : NLP

Set 14
Reference Resolution
Kraken, also called the Crab-fish, which is not that huge, for heads and tails counted, he is no larger than our Öland is wide [i.e., less than 16 km] ... He stays at the sea floor, constantly surrounded by innumerable small fishes, who serve as his food and are fed by him in return: for his meal, (if I remember correctly what E. Pontoppidan writes,) lasts no longer than three months, and another three are then needed to digest it.
Why Reference Resolution?

Q: What is the second-oldest service academy in the US?

The United States Naval Academy (also known as USNA, Annapolis, or Navy) is a four-year coeducational federal service academy located in Annapolis, Maryland, United States. Established in 1845 under Secretary of the Navy George Bancroft, it is the second-oldest of the United States' five service academies.
Reference Resolution, Defined

- Identify all noun phrases that refer to the same real-world entity.

- *Alternate*: ground each noun phrase (entity mention) to its real-world referent

- **Input**: noun phrases
- **Output**: sets of noun phrases
Terminology

- **Coreference**: two noun phrases both refer to the same real-world entity

- **Anaphora**: a noun phrase (the anaphor) refers to a previously mention noun phrase (the antecedent)
  - Key: interpretation of the anaphor is dependent on interpreting the antecedent
  - *John won the race. He is happy.*
  - Not all anaphora is coreference, but in general, most are.

*We went to see a movie yesterday. The tickets were expensive!*
Anaphora vs Coreference

- Anaphora

- Coreference
Terminology

- **Entity**: the real-world concept
- **Entity mention**: the noun phrase
- **Antecedent**: the previous NP to which our current NP refers

Text contains entity mentions, each of which belongs to an entity. Some entity mentions have antecedents in the text.
Hobbs Algorithm (for pronouns)

1. Start at the pronoun NP
2. Climb parse tree until NP or S. Call this node X, and path p that you climbed up.
3. Traverse all branches below X to the left of path p, breadth first. Propose any seen NP/S as the antecedent.
4. If node X is the highest S node, traverse previous sentences left-to-right breadth-first…most recent sentences first. When an NP is seen, propose as antecedent.
5. If node X not highest S node, go up tree to first NP or S. Call this the new X node.
6. If X is an NP node and if path p did not pass through a Nominal node that X immediately dominates, propose X.
7. Traverse breadth-first all branches below X to the left of path p. Propose any NP as the antecedent.
8. If X is an S node, traverse all branches of X to the right of path p. But don’t go below any NP or S encountered. Propose any seen NP.
Supervised Coreference

- Use labeled data, and learn to perform coreference!
- For each entity mention, find the best previous antecedent.

The man jumped the fence. It scraped his leg as he went over.

\[ P(\text{coreferent} | \text{“the man”}, \text{“it”}, \text{sentence}) \]
Learning in Coreference

- Create classifiers over pairs of NPs

1. Step through each NP in the text, in order
2. Choose a preceding NP as antecedent, or “new”
3. Start with the closest NP to the left
   - Compute $P(\text{coreferent} \mid \text{NP, NP}_{\text{left}}, \text{document})$
   - Determine if probability over a threshold
   - OR compute probability for all NPs to the left, choose best

- Many, many variants of this
  - However, the vast majority make these pairwise decisions
Features for Pronominal Coreference

- **Number**
  - Singular (he/she/it…) vs Plural (we/us/they…)

- **Gender**
  - He and she are obvious. John is male. Pat is ???

- **Person**
  - First/Second/Third … he/she/they refer to a third person name

- **Syntax**
  - English contains certain hard (and soft!) constraints
  - “Sally gave her a book” – (her cannot be Sally)
  - “Sally gave herself a treat” – (herself must be Sally)
Features for Pronominal Coreference

- Sentence distance between mentions
- Hobbs distance!
- Grammatical role of potential antecedent
  - Is the antecedent the subject, object, or PP?
- Linguistic form
  - Is the antecedent a definite noun (the car), an indefinite noun (a car), a pronoun (it), or a proper name (John)?
Features for Reference Resolution

- Features can work on any mentions, not pronouns!

  *I saw a 2013 Audi A3 yesterday. The red car zoomed past me.*

- Word edit distance
  - How similar are the two mentions?
- Word subphrase
  - Does one mention contain a subset of the other’s words?
  - “Tom Hanks” and “Hanks”
- Named entity type match
  - Are both mentions labeled as PERSON?
- Etc.
It’s sometimes very difficult

• Common nouns can differ in number
  • “A patrol moved down our street. The soldiers saw us.”

• Common nouns can refer to proper nouns
  • “Barack Obama … the president”

• Split antecedence!
  • “Bilbo found Gollum. They traveled to Mordor.”
Exercise!

“What a pity that Bilbo did not stab that vile creature, when he had a chance!”, Frodo said. “Pity? It was Pity that stayed his hand on Gollum. Pity, and Mercy: not to strike without need. And he has been well rewarded, Frodo. Be sure that he took so little hurt from the evil, and escaped in the end, because he began his ownership of the Ring so. With Pity.”

1. Identify the entities in this passage, and cluster the entity mentions. Write down sets of entity mentions. Some sets will only have one mention!

2. How do you know what other mention *Gollum* goes with?
How to Evaluate?

- We have labeled data (pairs of NPs linked to each other)

- **B-CUBED algorithm** *(Bagga and Baldwin, 1998)*
  - Precision: % of NPs in your guessed entity that are in the same gold entity
  - Recall: % of NPs in the gold entity that are in your guessed entity
  - You sum over all your guessed entities (sets of NPs) to get the overall precision and recall.
  - Greedy matching…choose best matching gold entities
B-CUBED

Precision(e) = 4/5

Recall(e) = 4/6

From Amigo et al. 2009
Alternative Approaches

• Multi-pass systems.

1. Run through all entity mentions, and only resolve the ones you are most certain about.
   • e.g., “John” and “John” and “John”

2. Run through again, resolving only the ones you are pretty certain about.
   • e.g., “John Smith” and “Smith”

3. Run through again, resolving only those you are reasonably sure of.

4. Etc.
Example Passes

Lee et al. 2012

- **Pass 3**
  - Yes: “the Florida Supreme Court”, “the Florida court”
  - No: “researchers”, “two Chinese researchers”
- **Pass 4** (Compatible Modifiers)
  - Yes: “President Clinton”, {American President, American President Bill Clinton, Clinton}
- **Pass 5** (Word Inclusion)
Multi-pass system

- This is **not** a learning system!
- It is a series of **rules**!

- It is contrary to much of NLP these days in that it outperforms learning approaches
- Why? It is very flexible, and the rules are gradually applied. Also...coreference is a somewhat unique task compared to other things we’ve covered in this course.
Key Points

1. Language is full of entity mentions
2. Most systems resolve *pairs of mentions*, find the single mention earlier in the text that matches
3. Coreference output: *sets of mentions*
4. Hobbs Algorithm: deterministic rules
5. Supervised Learning: labeled data and probability
   - Features include number, gender, word match, named entity types, etc.
6. Evaluation: precision/recall of the output sets