What is Coreference Resolution?

- Identify all mentions that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.
What is Coreference Resolution?

- Identify all mentions that refer to the same real world entity

Barack Obama nominated **Hillary Rodham Clinton** as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former **First Lady**.

Applications

- Full text understanding
  - information extraction, question answering, summarization, ...
- "He was born in 1961" (Who?)

Applications

- Full text understanding
- Machine translation
  - languages have different features for gender, number, dropped pronouns, etc.

Applications

- Full text understanding
- Machine translation
- Dialogue Systems
  - "Book tickets to see **James Bond**"
  - "Spectre is playing near you at 2:00 and 3:00 today. How many tickets would you like?"
  - "Two tickets for the showing at three"
Coreference Resolution in Two Steps

1. Detect the mentions (easy)
   "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said
   • mentions can be nested!

2. Cluster the mentions (hard)
   "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

Outline

• What is Coreference Resolution?
  • Mention Detection
  • Types of Reference
  • Coreference Resolution Models
  • Coreference Resolution Evaluation

Mention Detection

• Mention: span of text referring to some entity
• Three kinds of mentions:
  1. Pronouns
     • I, you, it, she, him, etc.
  2. Named entities
     • People, places, etc.
  3. Noun phrases
     • "a dog," "the big fluffy cat stuck in the tree"

Mention Detection: Not so Simple

• Marking all pronouns, named entities, and NPs as mentions over-generates mentions
• Are these mentions?
  • It is sunny
  • Every student
  • No student
  • The best donut in the world
  • 100 miles

How to deal with these bad mentions?

• Could train a classifier to filter out spurious mentions
• Much more common: keep all mentions as "candidate mentions"
• After your coreference system is done running discard all singleton mentions (i.e., ones that have not been marked as coreference with anything else)
Outline

• What is Coreference Resolution?
• Mention Detection
  ➔ Types of Reference
  • Coreference Resolution Models
  • Coreference Resolution Evaluation

Types of Reference

• Coreference is when two mentions refer to the same entity in the world
  • Barack Obama traveled to ... Obama

• A related linguistic concept is anaphora: when a term (anaphor) refers to another term (antecedent)
  • the interpretation of the anaphor is in some way determined by the interpretation of the antecedent
  • Barack Obama said he would sign the bill.
    antecedent  anaphor

Anaphora vs Coreference

• Coreference with named entities
  text: Barack Obama  Obama
  world

• Anaphora
  text: Barack Obama  he
  world

Anaphora vs. Cataphora

• Usually the antecedent comes before the anaphor (e.g., a pronoun), but not always

Cataphora

“From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum...”

[Oscar Wilde – The Picture of Dorian Gray]

Outline

• What is Coreference Resolution?
• Mention Detection
• Types of Reference
  ➔ Coreference Resolution Models
  • Coreference Resolution Evaluation
Learning-based Models: **Mention Pair**
- Train a binary classifier that assigns every pair of mentions a probability of being coreferent: $p(m_i, m_j)$
- e.g., for “she” look at all candidate antecedents (previously occurring mentions) and decide which are coreferent with it

> I voted for Nader because he was most aligned with my values,” she said.

Mention Pair Training
- $N$ mentions in a document
- $y_{ij} = 1$ if mentions $m_i$ and $m_j$ are coreferent, -1 if otherwise
- Just train with regular cross-entropy loss (looks a bit different because it is binary classification)

$$J = -\sum_{i=2}^{N} \sum_{j=1}^{i-1} y_{ij} \log p(m_j, m_i)$$

Mention Pair Test Time
- Coreference resolution is a clustering task, but we are only scoring pairs of mentions... what to do?
- Pick some threshold (e.g., 0.5) and add coreference links between mention pairs where $p(m_i, m_j)$ is above the threshold
- Take the transitive closure to get the clustering

> I voted for Nader because he was most aligned with my values,” she said.

Even though the model did not predict this coreference link, i and my are coreferent due to transitivity

Mention Pair Test Time
- Coreference resolution is a clustering task, but we are only scoring pairs of mentions... what to do?
- Pick some threshold (e.g., 0.5) and add coreference links between mention pairs where $p(m_i, m_j)$ is above the threshold
- Take the transitive closure to get the clustering

> I voted for Nader because he was most aligned with my values,” she said.

Adding this extra link would merge everything into one big coreference cluster!
Group discussion: Disadvantages of Mention Pair Models and Features for Computing Probability

\[
P(\text{"she"}, \text{"I"})?\]

Mention Pair Models: Disadvantage

- Suppose we have a long document with the following mentions:
  - Ralph Nader... he... his... him...<several paragraphs>
  - ... voted for Nader because he...

Mention Pair Models: Disadvantage

- Many mentions only have one clear antecedent
- But we are asking the model to predict all of them
- Solution: Instead train the model to predict only one antecedent for each mention

How do we compute the probabilities?

A. Non-neural statistical classifier
B. Simple neural network
C. More advanced model using LSTMs, attention

A. Non-Neural Coref Model: Features

- Person/Number/Gender agreement
- Jack gave Mary a gift. She was excited.
- Semantic compatibility
- John bought him a new car. (him can not be John)
- More recently mentioned entities preferred for referenced
- John went to a movie. Jack went as well. He was not busy.
- Grammatical Role: Prefer entities in the subject position
- John went to a move with Jack. He was not busy.
- Parallelism:
  - John went with Jack to a movie. Joe went with him to a bar.
B. Neural Coref Model
- Standard feed-forward neural network
- Input layer: word embeddings and a few categorical features

Neural Coref Model: Inputs
- Embeddings
  - Previous two words, first word, last word, head word, ... of each mention
  - The *head* word is the “most important” word in the mention – you can find it using a parser, e.g., *The fluffy cat stuck in the tree*
- Still need some other features:
  - Distance
  - Document genre
  - Speaker information

Learning-based Models: Mention Ranking
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

Learning-based Models: Mention Ranking
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

Learning-based Models: Mention Ranking
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

Learning-based Models: Mention Ranking
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)

Learning-based Models: Mention Ranking
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything (“singleton” or “first” mention)
Coreference Models: Training

- We want the current mention $m_i$ to be linked to any one of the candidate antecedents it's coreferent with.
- Mathematically, we want to maximize this probability:
  \[ \sum_{j=1}^{i-1} I(y_{ij} = 1) p(m_j, m_i) \]
  \[ \text{iterate through candidate antecedents (previously occurring mentions)} \]
  \[ \text{For ones that are coreferent to } m_i \]
  \[ \text{...we want the model to assign a high probability} \]
- The model could produce 0.9 probability for one of the correct antecedents and low probability for everything else, and the sum will still be large.

/coreference models: training

Coreference Models: Training

- We want the current mention $m_i$ to be linked to any one of the candidate antecedents it’s coreferent with.
- Mathematically, we want to maximize this probability:
  \[ \sum_{j=1}^{i-1} I(y_{ij} = 1) p(m_j, m_i) \]
  \[ \text{iterate over all the mentions in the document} \]
  \[ \text{Usual trick of taking negative log to ease from likelihood to loss} \]

Mention Ranking Models: Test Time

- Pretty much the same as mention-pair model except each mention is assigned only one antecedent

Outline

- What is Coreference Resolution?
- Mention Detection
- Types of Reference
  - Coreference Resolution Models (Rule-based, Learning-based)
    - Coreference Resolution Evaluation

Coreference Evaluation

- Many different metrics: MUC, CEA-F, LEA, B-CUBED, BLANC
  - Often report the average over a few different metrics
Coreference Evaluation

- An example: B-cubed
  - For each mention, compute a precision and a recall

\[ P = \frac{4}{5} \]
\[ R = \frac{4}{5} \]

Gold Cluster 1
Gold Cluster 2

System Cluster 1
System Cluster 2

Knowledge-based Pronominal Coreference

- She poured water from the pitcher into the cup until it was full
- She poured water from the pitcher into the cup until it was empty

- The city council refused the women a permit because they feared violence.
- The city council refused the women a permit because they advocated violence.

- Winograd (1972)
- These are called Winograd Schema

- Recently proposed as an alternative to the Turing test
  - See: Hector J. Levesque “On our best behavior” IJCAI 2013
  - http://www.onourbestbehavior.org/