Applied Natural Language Processing

Info 256
Lecture 23: Coreference (April 18, 2019)

David Bamman, UC Berkeley
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.
We are met on a great battle-field of that war.
We are met on a great battle-field of that war.
Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.
Discourse

• Discourse covers linguistic expression beyond the boundary of the sentence.

  • Dialogues: the structure of turns in conversation

  • Monologues: the structure of entire passages, documents
LUKE

I'll never join you!

VADER

If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE

He told me enough! It was you who killed him.

VADER

No. I am your father.

LUKE

No. No. That's not true! That's impossible!

VADER

Search your feelings. You know it to be true.

LUKE

No! No! No!
LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!
LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!
Coreference resolution

- “Trump met Putin today; he’s the leader of the US.”
Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School from 1992 to 2004. He served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.
Coreference resolution

Did Barack Obama die in an automobile accident in 1982?

attend graduate school at Harvard University on a scholarship. Obama's parents divorced in March 1964. Obama Sr. returned to Kenya in 1964 where he remarried; he visited Barack in Hawaii only once, in 1971. He died in an automobile accident in 1982 when his son was 21 years old.
Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial services company’s president. It has been ten years since she came to Megabucks from rival Lotsabucks.”
“Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial services company’s president. It has been ten years since she came to Megabucks from rival Lotsabucks.”

“Referent”

The entities or individuals in the real world that the text is pointing to.

- Victoria Chen
- Megabucks
- Lotsabucks
Coreference

“Referring expression”

The text that points to entities.

“Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial services company’s president. It has been ten years since she came to Megabucks from rival Lotsabucks.”
Coreference

“coreference”

The set of text strings that all refer to the same ENTITY.

“Victoria Chen, Chief Financial Officer of Megabucks Banking Corp since 2004, saw her pay jump 20%, to $1.3 million, as the 37-year-old also became the Denver-based financial services company’s president. It has been ten years since she came to Megabucks from rival Lotsabucks.”
Event coreference

I stubbed my toe on the chair and it really hurt.
Worth solving?
English constraints

• Number
  • I have a car. They are blue [*they = car]

• Gender
  • My dad is shoveling snow. He’s cold. [*he = snow]

• Person
  • We’re watching a movie. He likes it [*he = you and I]
English exceptions

• Number
  • A: I have a new friend.
    B: What’s their name?
  • We are a grandmother (Margaret Thatcher)

• Gender
  • “The Nellie, a cruising yawl, swung to her anchor without a flutter of the sails, and was at rest.” (Heart of Darkness)
  • It puts the lotion in the basket (Silence of the Lambs)

• Person
  • ???
English preferences

• Recency: more recent NPs are preferred

• Grammatical role: subjects are preferred
  • Billy Bones went to the bar with Jim Hawkins. He called for a glass of rum.

• Repeated mention: more discourse-salient NPs are preferred

• Parallelism
  • Long John Silver went with Jim to the Old Parrot. Billy Bones went with him to the Old Anchor inn.

• Verb semantics

• Selectional restrictions
Verb semantics

• John telephoned Bill. He lost the laptop

• John criticized Bill. He lost the laptop.
Winograd challenge

• The trophy would not fit in the brown suitcase because it was too big. What was too big?

• The town councilors refused to give the demonstrators a permit because they feared violence. Who feared violence?

• The town councilors refused to give the demonstrators a permit because they advocated violence. Who advocated violence?

http://www.commonsensereasoning.org
Selectional restrictions

• John parked his car in the garage after driving it around for hours.
Stanford “Sieve”

Sequence of pattern matching rules starting at high precision coreference links, progressing to higher recall.
Mention Detection

• All NPs, possessive pronouns, and named entity mentions are candidate mentions. Recall is more important than precision.

• Filters to remove candidates:
  • Remove mentions embedded within larger mentions with same headword
  • Remove numeric quantities (100 miles, 9%)
  • Remove existential there, it
  • Remove adjectival forms of nations
  • Remove 8 stop words (there, ltd., hmm)

Lee et al., 2011
John is a musician. He played a new song. A girl was listening to the song. “It is my favorite,” John said to her.
Classification

A mapping $h$ from input data $x$ (drawn from instance space $\mathbf{x}$) to a label (or labels) $y$ from some enumerable output space $\mathbf{y}$

$\mathbf{x} = \text{set of all documents}$

$\mathbf{y} = \{\text{english, mandarin, greek, ...}\}$

$x = \text{a single document}$

$y = \text{ancient greek}$
Classification

Positive examples = pronouns paired with closest antecedent (or coreference chain)

Negative examples = entities not in coreference chain.
Classification

For every possible antecedent y for pronoun x, we frame a binary classification: is y coreferent with x? Every noun phrase is a candidate antecedent.

- I
- you
- you
- the power
- the power of the dark side
- the dark side
- Obi-Wan
- you
- your
- your father
- He
- me
- you

LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!
Classifier

Let’s brainstorm a supervised classifier.
Features

• John saw a beautiful 1961 Ford Falcon at the used car dealership

• He showed it to Bob.

• He bought it.
Features

- Unary features *(valid of a single token)*
  - token, lemma, part of speech
  - salience

- Binary features *(valid of a pair of tokens)*
  - number agreement (plural pronoun/plural NP)
  - compatible number (plural pronoun/??? NP)
  - gender agreement
  - compatible gender
  - sentence distance
  - Hobbs distance
  - syntax: grammatical role
Nominal coreference

• Pronominal coreference is a subset of the full coreference resolution problem because pronouns are nearly always coreferent.

• How would we extend the classification approach to general nominal referents?
Evaluation

- Evaluating general reference resolution (i.e., all noun phrase entities) is more complicated than straightforward accuracy/precision/recall

\[ B^3_{\text{precision}} = \frac{1}{n} \sum^n_i \frac{|Gold_i \cap System_i|}{|System_i|} \]

\[ B^3_{\text{recall}} = \frac{1}{n} \sum^n_i \frac{|Gold_i \cap System_i|}{|Gold_i|} \]
LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!
LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!
LUKE
I'll never join you!

VAIDER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VAIDER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VAIDER
Search your feelings. You know it to be true.

LUKE
No! No! No!
LUKE

I'll never join you!

VADER

If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE

He told me enough! It was you who killed him.

VADER

No. I am your father.

LUKE

No. No. That's not true! That's impossible!

VADER

Search your feelings. You know it to be true.

LUKE

No! No! No!
I'll never join you!

If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

He told me enough! It was you who killed him.

No. I am your father.

No. No. That's not true! That's impossible!

Search your feelings. You know it to be true.

No! No! No!

Example system output: 4 entities

3 = {I, me, I}
8 = {you, you, you, your, you, your, your, you}
3 = {Obi-Wan, your father, your father}
2 = {He, him}
Evaluation

• More complicated than straightforward accuracy/precision/recall

\[ B^3_{\text{precision}} = \frac{1}{n} \sum_{i}^{n} \frac{|Gold_i \cap System_i|}{|System_i|} \]

\[ B^3_{\text{recall}} = \frac{1}{n} \sum_{i}^{n} \frac{|Gold_i \cap System_i|}{|Gold_i|} \]

\( n \) ranges over all entities in gold and system output
LUKE
I’ll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That’s not true! That’s impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!

| Gold$_i$ ∩ System$_i$ | = 2 | | Gold$_i$ | = 8 | | System$_i$ | = 3 |
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!

| Gold_i ∩ System_i | = 2 | Gold_i | = 6 | System_i | = 8
I'll never join you!

If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

He told me enough! It was you who killed him.

No. I am your father.

No. No. That's not true! That's impossible!

Search your feelings. You know it to be true.

No! No! No!

| Gold$\cap$ System$| = 6 |  | Gold$| = 8 |  | System$| = 8 |
LUKE
I'll never join you!

VADER
If you only knew the power of the dark side. Obi-Wan never told you what happened to your father.

LUKE
He told me enough! It was you who killed him.

VADER
No. I am your father.

LUKE
No. No. That's not true! That's impossible!

VADER
Search your feelings. You know it to be true.

LUKE
No! No! No!

| Gold_i ∩ System_i | = 1 | | Gold_i | = 2 | | System_i | = 3 |
I'll never join you!

You only knew the power of the dark side. Obi-Wan never told you what happened to your father.

He told me enough! It was you who killed him.

No. I am your father.

No. No. That's not true! That's impossible!

Search your feelings. You know it to be true.

No! No! No!
Evaluation

• More complicated than straightforward accuracy/precision/recall

\[ B^3_{\text{precision}} = \frac{1}{n} \sum_{i}^{n} \frac{|Gold_i \cap System_i|}{|System_i|} \]

\[ B^3_{\text{recall}} = \frac{1}{n} \sum_{i}^{n} \frac{|Gold_i \cap System_i|}{|Gold_i|} \]

n ranges over all entities in gold and system output
“Between him and Darcy there was a very steady friendship, in spite of great opposition of character. Bingley was endeared to Darcy by the easiness, openness, and ductility of his temper, though no disposition could offer a greater contrast to his own, and though with his own he never appeared dissatisfied."
The Clinton campaign is circulating a fake photo of Barack Obama in Muslim clothes to damage his reputation. In the photo, Obama wears a long sari-like garment.

The Clinton campaign is circulating a fake photo of Barack Obama in Muslim clothes to damage his reputation, but Obama never wore Muslim clothes.
• You cannot read Cyril Connolly for very long without wanting to acquire—and then developing—a relationship with the personality of the man himself. [. . . ] With Connolly there is a marked difference and the difference is that the artist and the man are so conjoined and intermingled that you cannot savour the one without the other and vice versa.
Non-identity

• Non-Identity. The two NPs point to two different DEs. Even if they share any feature, they are not ‘the same thing.’

• “President Samaranch sent a letter to Sydney in which he asked for information. A similar missive has also been received by all the candidate cities to host the Olympic Games of 1996.”

Recasens et al. 2010
Identity

• Identity. The two NPs point to the same DE (i.e., they have the same set of attributes, as far as one can tell). They are (almost certainly) ‘the same thing.’

• “It began when a Hasidic Jewish family bought one of the town’s two meat-packing plants 13 years ago. First they brought in other Hasidic Jews, then Mexicans, Palestinians, Ukrainians.”

Recasens et al. 2010
Identity

• Identity. The two NPs point to the same DE (i.e., they have the same set of attributes, as far as one can tell). They are (almost certainly) ‘the same thing.’

• “It began when a Hasidic Jewish family bought one of the town’s two meat-packing plants 13 years ago. First they brought in other Hasidic Jews, then Mexicans, Palestinians, Ukrainians.”

Recasens et al. 2010
Near-identity

- A proper noun appears first, and a subsequent noun phrase refers to some aspect of the discourse entity
  - Role
  - Location
  - Organization
  - Information realization
  - Representation
  - Other

Recasens et al. 2010
Role near-identity: A specific role or function performed by a human, animal or object, is distinguished from their other facets.

“Our father was the greatest” commented an anonymous old lady while she was shaking Alessandro’s hand — Gassman’s best known son. “I will miss the actor, but I will be lacking my father especially,” he said.

Recasens et al. 2010
Location near-identity: The name of a location can be used to describe facets such as the physical place, the place associated with a (political) organization, the population living in that location, the ruling government, an affiliated organization, an event celebrated at that location, etc.

“The Jordan authorities arrested, on arriving in Iraq, an Italian pilot who violated the air embargo to this country.”

Recasens et al. 2010
Organization near-identity: The name of a company or other social organization can be used to describe facets such as the legal organization itself, the facility that houses the organization or one of its branches, the company shares, a product manufactured by the company, etc.

“The strategy has been a popular one for McDonalds . . . It’s a very wise move on for them because if they would have only just original McDonalds, I don’t think they would have done so great.”

Recasens et al. 2010
Information realization near-identity: A discourse entity corresponding to an informational object (e.g., story, law, review, etc.) can be split according to the format in which the information is presented or manifested (FRBR abstraction hierarchy)

She hasn’t seen Gone with the Wind, but she’s read it.

Recasens et al. 2010
Representation near-identity: One noun phrase is a representation of the other--as in a picture or a starring of a person, or a toy replica of a real object.

We stand staring at two paintings of Queen Elizabeth. In the one on the left, she is dressed as Empress of India. In the one on the right, she is dressed in an elegant blue gown.

Recasens et al. 2010
Part-whole near-identity: One noun phrase mentions a part to refer to the whole expressed by the other noun phrase.

The City Council approved legislation prohibiting selling *alcoholic drinks* during night hours ...Bars not officially categorized as bars will not be allowed to sell alcohol.

Recasens et al. 2010
Solve it

• Ontonotes
  • http://catalog.ldc.upenn.edu/LDC2013T19

• MUC 7
  • http://catalog.ldc.upenn.edu/LDC2001T02

• ACE 2003
  • http://catalog.ldc.upenn.edu/LDC2001T02
Activity

• 15.coref/CorefSetup.ipynb