Introduction to coreference: training materials for coreference annotators

Courtesy Kevin Bretonnel Cohen
Presented by Arrick

Training Schedule

Session 1: Introduction

Session 2: More guidelines and practice annotation; homework assigned

Session 3: Introduction to Knowtator; begin annotation of (practice) biomedical text; homework assigned

Session 4: Go over annotation homework; begin training texts

Coreference defined

- *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while traveling on a plane.

Coreference defined

- *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while traveling on a plane.




Coreference defined

- *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while traveling on a plane.




Sophia Loren, she, The actress, her, she

Coreference defined

- *Sophia Loren* says *she* will always be grateful to Bono. *The actress* revealed that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while traveling on a plane.

Bono, the U2 singer

How do humans do this?

- Linguistic factors:
  - Kevin saw Larry. He liked him.
- Knowledge about the world:
  - Sophia Loren will always be grateful to Bono. The actress…
  - Sophia Loren will always be grateful to Bono. The singer…
  - Sophia Loren will always be grateful to Bono. The storm…
- A combination of world knowledge and linguistic factors:
  - Sophia Loren says she will always be grateful to Bono…
  - Sophia Loren says he will always be grateful to Bono…

Computers are bad at this

- Linguistic features don't always help.
  - Each child ate a biscuit. They were delicious.
  - Each child ate a biscuit. They were delighted.
- Programming enough knowledge about the world into a computer has proven to be very difficult.

Where you fit in to all of this

- Computers can get better at this if they have "training data" to work from.

- You are going to make training data.

Coreference defined

What you'll start with

- Sophia Loren says she will always be grateful to Bono. The actress revealed that the U2 singer helped her calm down when she became scared by a thunderstorm while traveling on a plane.

What you'll produce

- *Sophia Loren* says *she* will always be grateful to Bono. *The actress revealed* that the U2 singer helped *her* calm down when *she* became scared by a thunderstorm while traveling on a plane.

What you need to know to be able to do this

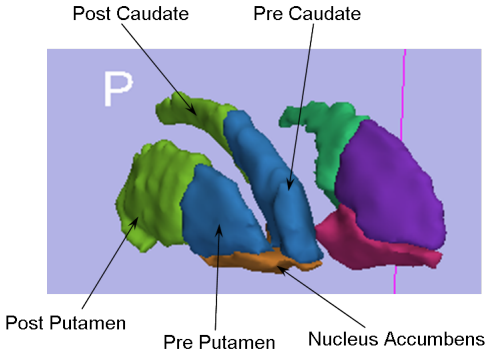
- How to use the tool that you' ll be utilizing to mark the "right answers"
- A long set of rules that dictate what to mark as "the same" and what to ignore

How do you know what "the right answer" is?

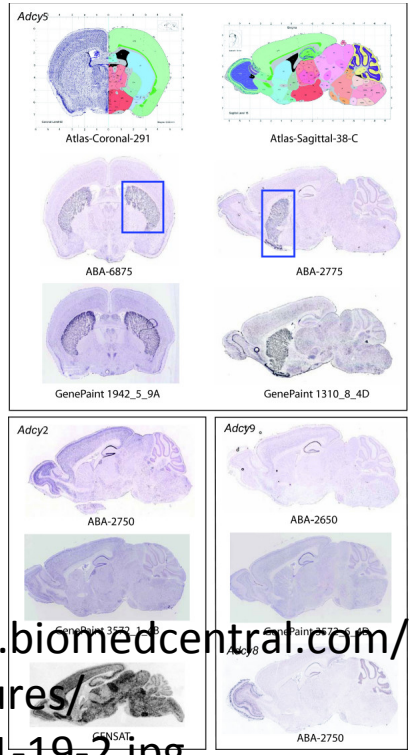
- Trust your intuition. Humans are good at this.

If it "means" the same thing in the world…If it doesn't…

- *The striatum* is a subcortical part of the telencephalon. <u>The mouse striatum</u> is an excellent model for…



http://www.na-mic.org/Wiki/images/1/1a/Striatum1.png



http://www.biomedcentral.com/content/figures/1752-0509-1-19-2.jpg

Learning the OntoNotes guidelines

- OntoNotes: A large project to build data for helping computers learn to understand language
- You will always have these at hand—we will now spend some time helping you understand them.

General principle #1

It's easier to remember what not to link

up than what to link up.

Copular structures

- A copular structure consists of a referent (usually the subject), an attribute of that referent (usually the predicate), and a copula (most often, though not always, a 'linking' verb). The copula serves to equate (or link) the referent with the attribute….
  - [John] is [a linguist] (no co-ref.)
- Some common copular verbs are: *be, appear, feel, look, seem, remain, stay, become, end up, get*….
- Not all copular structures include a verb. In this example, "or" functions as a copula; therefore, [no] IDENT…relation is marked.
  - Among other things, Mr. Bologna said that the sale will facilitate Gen-Probe's marketing of a diagnostic test for [acquired immune deficiency syndrome], <u>or</u> [AIDS].

OntoNotes guidelines, Section 2.5, Pages 8-9

Pronouns that don't "mean" anything get marked separately

- It is evident that...
- It is raining.
- It is six o'clock.
- It is the case that...
- It is the proteins that do the work.

Appositives get marked separately

- Appositive construction: two (or more) things linked together without a verb
  - John Smith, noted linguist…
  - A famous linguist, he…
  - Heat shock protein 60 (HSP-60)

So, what does get marked?

- "Names, nominal mentions, pronominal mentions, and verbal mentions of the same entity, concept, or event are coreferenced…." (OntoNotes guidelines Section 1.2.1, p. 3)

How to handle specific cases

- Appositives
- Nonreferential pronouns
- Generics
- Copular constructions
- Verbs
- Premodifiers
- Nested mentions

More detail on appositives

- Official definition: "Appositive constructions consist of two (or more) immediately-adjacent noun phrases, separated only by a comma, colon, dash, or parenthesis." (OntoNotes guidelines Sect. 1.2.2, p. 3)

Appositive examples

- John Smith, noted linguist, …
- A famous linguist, he…
- …the president of the linguistics club, J. Smith…
- …heat shock protein 60 (HSP-60)
- The likelihood ratio statistic—a value that can be read as a chi-square—peaks in this 20 cM interval…
- …Grk2, a member of the family of ionotropic receptor genes that is thought to play a role in modulating Huntington disease…

Appositives have a head and one or more attributes

- The "most specific" part of the appositive construction is the "head" of the construction.
- The other parts are its "attributes."
  - John Smith, noted linguist,…

head attribute

How to decide what the most specific element of the appositive is

- The "specificity scale:"

proper noun > pronoun > definite NP > indefinite specific NP > non-specific NP

John > he > the linguist > a linguist I know > noted linguist

What a proper noun is

- A proper noun is a name.
- In our texts, this includes names of genes, proteins, and protein families, even if they are not capitalized.

  - …Vax1 interacts with several molecules including sonic hedgehog, Pax2, Pax6, and Rx that are known to be important during development of the basal forebrain… (Rosen and Williams 2001)

What a noun phrase is

- A noun phrase is a noun plus any words that modify it.
- Modifiers can be on the left…
  - Dog
  - Big dog
  - Big yellow dog
  - This big yellow dog
- …or on the right…
  - Dog on the couch
  - Dog that I saw yesterday
- …or both.
  - This big yellow dog on the couch that I saw yesterday

What a definite NP is

- Definite noun phrases start with words that let you know that the noun phrase refers to a specific individual or specific set of individuals in the world, and that we know who/what it is.
- The, this, these, that, those, my, your, its…

What an indefinite NP is

- An indefinite NP starts with a word that lets us know that we don't know which specific individual is being referred to.
- A, an, some…
- Watch out for indefinites as they require special treatment!

The same thing might get referred to with an indefinite NP first and then with definite NPs later

- A guy walks into a bar. He walks up to the bartender. The guy is wearing a t-shirt that says "I hate foodservice workers."

Now that you know those definitions, here is the specificity scale again

proper noun > pronoun > definite NP > indefinite specific NP > non-specific NP

John > he > the linguist > a linguist I know > noted linguist

Which element is the head?

John Smith, noted linguist, …

A famous linguist, he…

The president of the linguistics club, J. Smith

If two things are equally specific, pick the one to the left as the head

- Heat shock protein 60 (HSP-60)
  - Head: Heat shock protein 60
  - Attribute: HSP-60

Linking appositives to other things

- You link the appositive head to any subsequent things with which it corefers.
- Appositive: Richard Godown$_{\text{HEAD}}$, president of the company$_{\text{ATTRIBUTE}}$, gave a speech today. He said…
- Coreference: Richard Godown, head of the company, gave a speech today. He said…

Generics

- Ignore the section on generics in the guidelines.
- We have a different definition of what counts as a generic mention.
- We'll discuss this later

Premodifiers: defined

- "Premodifiers" are things to the left of a noun.
  - *Big yellow* dog
  - *Army Corps* spokesman
  - *Wheat* field

Link all premodifying nouns, whether proper or not

- But the Army Corps of Engineers expects the river level to continue falling this month. "The flow of the Missouri River is slowed," an Army Corps spokesman said.

- The premodifier *Army Corps* is proper, so it gets grouped with the preceding noun phrase

Non-proper Pre-modifying nouns

- Wheat is an important part of the economy in the Midwest. In Kansas, wheat fields stretch as far as the eye can see.

- The premodifier *wheat* in *wheat fields* is not proper, but it *does* get grouped with the preceding noun phrase

- This differs from the OntoNotes policy on generics (*wheat* is considered generic according to the OntoNotes guidelines.)

Examples of Premodifying Nouns

- “BXD5 mice” would yield two NPs,

  1. [BXD5 mice] to be linked with other mentions of [BXD5 mice]

  2. [BXD5] to be linked with other mentions of [BXD5]

- For “liver cells”, link [liver cells] with [liver cells] and [liver] with [liver]

- heterogeneous [F2] animals (refer to pg 2, ex. (22))

  Do not group adjectives with a premodifier if they modify the larger NP span, as in the NP above

- However, for “striatal neuron packing density,” ‘striatal’ modifies ‘neuron’ so it is included in the premodifying span.

  [striatal neuron] packing density

Only link premodifiers when they are nouns

- Charles Dickens was famous for his memorable characters. The Dickensian character has since become a literary archetype.

- The premodifier *Dickensian* is an adjective, not a proper noun, so it does not get grouped with the preceding noun phrase

Verbs

- Verbs can corefer with a noun.
- This will most likely happen when there is a nominalization (a noun that is formed from a verb, e.g. "phosphorylate/phosphorylation."
- Select just the verb itself.

Verbs

- Sales of passenger cars grew 22%. The strong growth followed year-to-year increases.

Nested mentions

- When linking premodifiers from a larger NP, in general, select the longest logical span.

- For example, if you see *epidermal growth factor* and [*epidermal growth factor] receptor,* do link the two mentions of *epidermal growth factor* together.

- Do not break up names more than once.

- For example, **do not** link [*growth factor*] to other mentions of *growth factor* if you have already extracted [*epidermal growth factor*]

What CANNOT be linked to other mentions?

•Nested mentions containing gene and gene family names cannot be broken up

-platelet derived growth factor receptor

a) 'platelet' cannot be extracted, as it breaks up the full NP

b) 'growth factor' also cannot be extracted

•Left-most NP that includes a prepositional phrase

-neuron number in mice (notice that [neuron] number in mice is valid, though)

-number of neurons

•Adjectives

-the mammalian CNS

-male mice ('male' functions as an adjective when it occurs with a head noun)