Coreference Resolution

Natalie Parde
UIC CS 421

What is coreference resolution?

The process of automatically identifying expressions that refer to the same entity

Coreference resolution is essential to creating high-performing NLP systems.

Which NLP course do you want to take next year?

What are my options?

Well, there's CS 421: Natural Language Processing, CS 521: Statistical Natural Language Processing, CS 532: Advanced Topics in NLP, and CS 533: Deep Learning for NLP.

Hmm, I'll do Statistical NLP.

Coreference resolution is essential to creating high-performing NLP systems.

Which NLP course do you want to take next year?

What are my options?

Well, there’s CS 421: Natural Language Processing, **CS 521: Statistical Natural Language Processing**, CS 532: Advanced Topics in NLP, and CS 533: Deep Learning for NLP.

Hmm, I’ll do **Statistical NLP**.

Both humans and NLP systems interpret language with respect to a discourse model.

- **Discourse model:** Mental model that is built incrementally, containing representations of entities, their properties, and the relations between them
- **Referent:** The discourse entity itself
  - (CS 521: Statistical Natural Language Processing)
- **Referring expression:** The linguistic expression referring to a referent
  - "CS 521"
  - "CS 521: Statistical Natural Language Processing"
  - "521"
  - "Statistical NLP"
- Two or more referring expressions that refer to the same discourse entity are said to **corefer**

Anaphora

- **Anaphora:** Referring to an entity that has already been introduced in the discourse
  - First mention is the **antecedent**
  - Subsequent mentions are **anaphors**
  - Entities with only a single mention are **singletons**

The University of Illinois at Chicago is an excellent place to study natural language processing. UIC has many faculty currently working in the area, including but not limited to Natalie Parde, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. The school is located in bustling downtown Chicago, and as a bonus it will be opening a snazzy new (non-brutalist) CS building in 2023.

Anaphora

- **Anaphora:** Referring to an entity that has already been introduced in the discourse
  - First mention is the **antecedent**
  - Subsequent mentions are **anaphors**
  - Entities with only a single mention are **singletons**

The **University of Illinois at Chicago** is an excellent place to study natural language processing. UIC has many faculty currently working in the area, including but not limited to Natalie Parde, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. The school is located in bustling downtown Chicago, and as a bonus it will be opening a snazzy new (non-brutalist) CS building in 2023.

Anaphora

- **Anaphora:** Referring to an entity that has already been introduced in the discourse
  - First mention is the **antecedent**
  - Subsequent mentions are **anaphors**
  - Entities with only a single mention are **singletons**

The **University of Illinois at Chicago** is an excellent place to study natural language processing. **UIC** has many faculty currently working in the area, including but not limited to Natalie Parde, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. **The school** is located in bustling downtown Chicago, and as a bonus **it** will be opening a snazzy new (non-brutalist) CS building in 2023.

Anaphora

- **Anaphora:** Referring to an entity that has already been introduced in the discourse
  - First mention is the **antecedent**
  - Subsequent mentions are **anaphors**
  - Entities with only a single mention are **singletons**

The **University of Illinois at Chicago** is an excellent place to study natural language processing. **UIC** has many faculty currently working in the area, including but not limited to **Natalie Parde**, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. **The school** is located in bustling downtown Chicago, and as a bonus **it** will be opening a snazzy new (non-brutalist) CS building in 2023.

Coreference Chains

A set of coreferring expressions is often called a **coreference chain**

The **University of Illinois at Chicago** is an excellent place to study natural language processing. **UIC** has many faculty currently working in the area, including but not limited to **Natalie Parde**, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. **The school** is located in bustling downtown Chicago, and as a bonus **it** will be opening a snazzy new (non-brutalist) CS building in 2023.

{"University of Illinois at Chicago", "UIC", "The school", "it"}

{"Natalie Parde"}

Two Key Tasks

- **Coreference resolution** thus generally comprises two key tasks:
  - Identify **referring expressions** (mentions of entities)
  - Cluster them into **coreference chains**
- We can also perform **entity linking** to map coreference chains to real-world entities
  - {"University of Illinois at Chicago", "UIC", "The school", "it"} → https://en.wikipedia.org/wiki/University_of_Illinois_at_Chicago

Linguistic Background

- Referring expressions can occur in several forms:
  - **Indefinite noun phrases**
  - **Definite noun phrases**
  - **Pronouns**
  - **Proper nouns (names)**
- These can be used to **evoke** and **access** entities in the discourse model in a variety of ways

Indefinite Noun Phrases

- Usually marked with the determiner *a* or *an*
- Can also be marked with other indefinite terms
  - E.g., *some*
- Generally introduce **new entities** to the discourse

The blue line was experiencing delays so I took **an** Uber.

Definite Noun Phrases

- Usually marked with *the*
- Generally refer to entities that have already been introduced to the discourse
- May refer to entities that haven't been introduced to the discourse, but are identifiable to the receiver due to:
  - World knowledge
  - Implications from the discourse structure

The blue line was experiencing delays so I took **an** Uber. Unfortunately, so did everyone else …**the** Uber got stuck in a traffic jam.

Have you checked out **the** Andy Warhol exhibit?

Make sure to order **the** tiramisu!

Pronouns

- Generally refer to entities that have already been introduced to the discourse and are easily identifiable

The blue line was experiencing delays so I took **an** Uber. Unfortunately, so did everyone else …**the** Uber got stuck in a traffic jam. **It** ended up reaching UIC later than the original train I'd been hoping to catch.

Proper Nouns (Names)

- Can be used either to introduce new entities to the discourse, or to refer to those that already exist

Chicago, Illinois is one of the largest cities in the United States. **Chicago** is known for its architecture, its thriving arts and music scene, its hot dogs and deep dish pizza, and---of course---its winter weather.

Information Status

- Referring expressions can also be categorized by their **information status**
  - The way they introduce **new information** or access **old information**
- Three main groups:
  - New noun phrases
  - Old noun phrases
  - Inferables

New Noun Phrases

- **Brand new NPs:** Introduce entities that are both **new to the discourse** and **new to the listener**
  - E.g., *an Uber*
- **Unused NPs:** Introduce entities that are **new to the discourse** but **not to the listener**
  - E.g., *Chicago*

Old Noun Phrases

- Introduce entities that already exist in the discourse model (and are thus **not new to the discourse nor to the listener**)
  - E.g., *she*

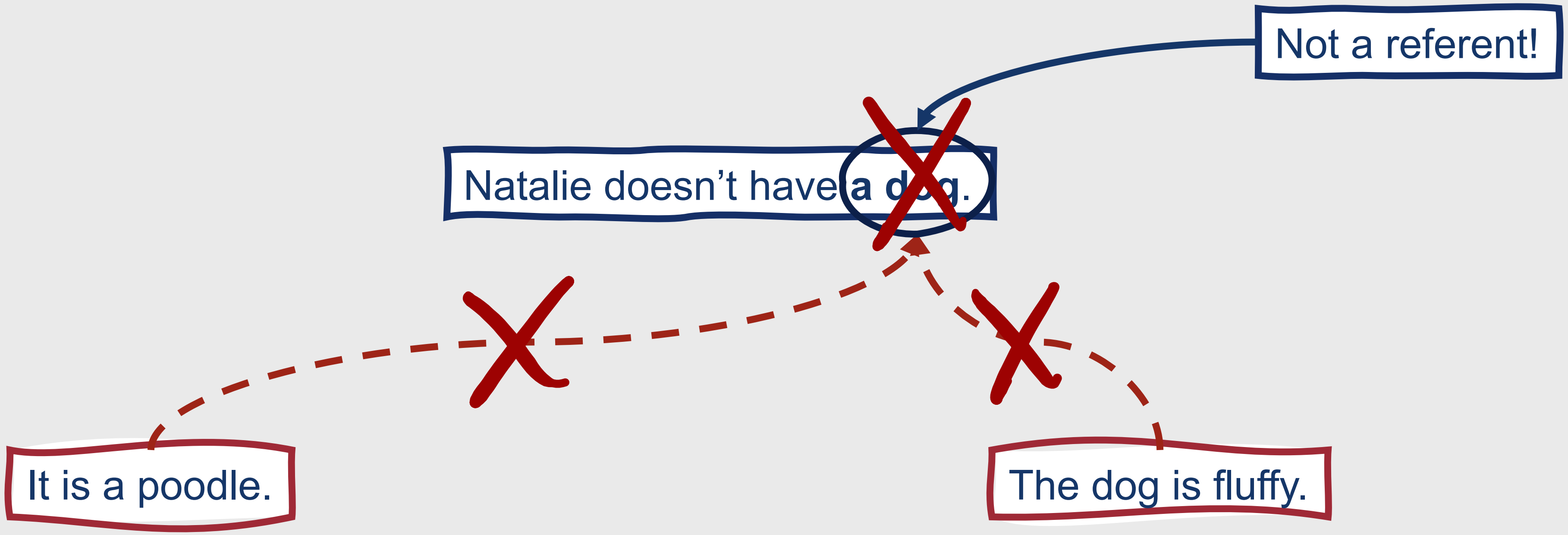
Inferables

- Introduce entities that are **new to the discourse** and **new to the listener** *but* the hearer can infer their existence by reasoning about other entities already introduced
  - E.g., I got in my Uber and told *the driver* to take us to UIC as fast as she could.

Generally, the form of a referring expression gives strong clues about its information status.

- **Very salient** (easily accessible) entities can be referred to using **less linguistic material**
  - E.g., pronouns
- **Less-salient** entities (e.g., those that are discourse-new and hearer-new) require **more linguistic material**
  - E.g., full names

Note: Not all noun phrases are referring expressions!



Structures Easily Confused with Referring Expressions

Appositives	Noun phrases that describe other noun phrases	Natalie Parde, *Assistant Professor of Computer Science*, teaches CS 521.
Predicative and Prenominal Noun Phrases	Noun phrases that describe characteristics of other noun phrases	Natalie Parde is an *Assistant Professor*.
Expletives	Non-referential pronouns	Natalie thought *it* was cool that so many students at UIC were interested in NLP.
Generics	Pronouns that refer to classes of nouns in general, rather than specific instances of those nouns	In Chicago, *you* get to experience all four seasons - summer, early winter, winter, and late winter.

So far, we've focused on linguistic properties of referring expressions….

- What about linguistic properties of coreference relations (relations between an anaphor and its antecedent)?
  - Number agreement
  - Person agreement
  - Gender/noun class agreement
  - Binding theory constraints
  - Recency
  - Grammatical role
  - Verb semantics
  - Selectional restrictions

Number Agreement

- In general, antecedents and their anaphors should agree in number
  - Singular with singular
  - Plural with plural
- A few exceptions:
  - Some semantically plural entities (e.g., companies) can be referred to using either singular or plural pronouns
  - “They” can be used as a singular pronoun

Person Agreement

In general, antecedents and their anaphors should agree in person

First person with first person

- I, my, me

Third person with third person

- They, their, them

An exception:

Text containing quotations

- "**I** spent twelve hours making those slides," **she** pointed out.

Gender/Noun Class Agreement

- In general, antecedents and their anaphors should agree in grammatical gender
  - He with his
  - She with hers
  - They with theirs
- This is an even bigger deal in (the many!) languages for which all nouns have grammatical gender
  - La casa 🏠
  - El banco 🏦

Binding Theory Constraints and Recency

- **Binding Theory Constraints:** Antecedents and their anaphors should adhere to the syntactic constraints placed upon them
  - Reflexive pronouns (e.g., herself) corefer with the subject of the most immediate clause that contains them
    - **Natalie** told **herself** that she wouldn't be nearly as busy next week.
- **Recency:** Antecedents introduced recently tend to be more salient than those introduced earlier
  - Pronouns are likelier to be anaphors for the most recent plausible antecedent
    - Natalie went to a **faculty meeting**. Shahla went to a **student government meeting**. **It** was mainly about new policy changes that had recently been approved.

Grammatical Role

- Antecedents in some grammatical roles are more salient than others
  - Subject position > object position
    - **Natalie** went to the Eiffel Tower with **Shahla**. **She** took a selfie.

Verb Semantics

- Salience may be influenced by the types of verbs to which antecedents and anaphors are arguments
  - **Natalie** congratulated **Shahla**. **Her** paper had just been accepted.
  - **Natalie** bragged to **Shahla**. **Her** paper had just been accepted.

Selectional Restrictions

- Finally, salience may also be influenced by other semantic knowledge about the verbs to which antecedents and anaphors are arguments
  - Natalie pulled her **suitcase** out of the **Uber**. **It** sped off into the sunset.

Coreference Tasks

- Now that we have some more linguistic background, we can formalize the task of coreference resolution:
  - **Given a text *T*, find all entities and the coreference links between them**
- This requires a few subtasks:
  - **Detect mentions**
    - Pronominal anaphoras
      - Filter out non-referential pronouns
    - Definite noun phrases
    - Indefinite noun phrases
    - Names
  - **Link those mentions into clusters**

What counts as a mention? What types of links are annotated?

- Depends on the task specifications and dataset
- Some coreference datasets do not include singletons as mentions
  - Makes the task easier
    - Singletons are often difficult to distinguish from non-referential noun phrases, and constitute a majority of mentions
- Some coreference datasets provide human-labeled mentions
  - Task is simply to cluster those mentions into groups

Sample Coreference Task

The University of Illinois at Chicago is an excellent place to study natural language processing. UIC has many faculty currently working in NLP, including but not limited to Natalie Parde, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. The school is located in bustling downtown Chicago, and as a bonus it will be opening a snazzy new (non-brutalist) CS building in 2023.

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus it will be opening a snazzy new (non-brutalist) **CS building** in 2023.

Detect mentions

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus it will be opening a snazzy new (non-brutalist) **CS building** in 2023.

Detect mentions

Cluster mentions

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus it will be opening a snazzy new (non-brutalist) **CS building** in 2023.

Detect mentions

Cluster mentions

Coreference Chains:

- {University of Illinois at Chicago, UIC, The school}
- {natural language processing, NLP}
- {faculty}
- {Natalie Parde}
- {Barbara Di Eugenio}
- {Cornelia Caragea}
- {Bing Liu}
- {Philip Yu}
- {Chicago}
- {CS building}

Popular Coreference Datasets

OntoNotes

- Chinese, English, and Arabic texts in a variety of domains (e.g., news, magazine articles, speech data, etc.)
- No singletons
- https://catalog.ldc.upenn.edu/LDC2013T19

ISNotes

- Adds information status to OntoNotes
- https://github.com/nlpAThits/ISNotes1.0

ARRAU

- English texts in a variety of domains
- Includes singletons
- https://catalog.ldc.upenn.edu/LDC2013T22

Moving on to the finer details….

- Mention detection: The process of finding spans of text that constitute a referring expression (mention)
  - Typically very liberal in predicting mentions, and rely on downstream filtering to prune bad predictions

The **University of Illinois at Chicago** is an excellent ~~**place**~~ to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a ~~**bonus**~~ it will be opening a snazzy new (non-brutalist) **CS building** in 2023.

Mention Detection

- How is filtering performed?
  - Sometimes, rules
  - More often, **classifiers**
    - Referentiality classifier
    - Anaphoricity classifier
    - Discourse-new classifier
- Classifiers for mention filtering often make use of a variety of features characterizing the words, their relationship, and their position in the surrounding text

1. Take all noun phrases, possessive pronouns, and named entities
2. Remove numeric quantities, mentions embedded in larger mentions, and stop words
3. Remove non-referential "it" based on regular expression patterns

“Hard” filtering based on rules or classifiers isn’t necessarily the best option.

- Filter too many → recall suffers
- Filter too few → precision suffers
- Modern solution?
  - Perform mention detection, anaphoricity filtering, and entity clustering jointly in an end-to-end model
- Still an open and active area of investigation

Architectures for Coreference Algorithms

Modern systems:

- Supervised neural machine learning

Several different ways to tackle the problem:

- **Entity-based classification**
  - Make decisions based on a given entity in the discourse model as a whole
- **Mention-based classification**
  - Make decisions locally for each mention
- **Ranking models**
  - Compare potential antecedents with one another (can be combined with either entity-based or mention-based approaches)

The Mention-Pair Architecture

Simple premise:

Given:

- Pair of mentions (candidate anaphor and candidate antecedent)

Decide:

- Whether or not they corefer

How does this work?

Compute coreference probabilities for every plausible pair of mentions

Goal: High probability for actual coreferring pairs, and low probability for other pairs

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as **a bonus** **it** will be opening a snazzy new (non-brutalist) **CS building** in 2023.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as **a bonus** **it** will be opening a snazzy new (non-brutalist) **CS building** in 2023.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a **bonus** **it** will be opening a snazzy new (non-brutalist) **CS building** in 2023.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a **bonus** **it** will be opening a snazzy new (non-brutalist) **CS building** in 2023.

How do we learn these probabilities?

- Select training samples
  - One positive instance (m_i, m_j) where m_j is the closest antecedent to m_i
  - A negative instance (m_i, m_k) for each m_k between m_j and m_i
- Extract features
  - Hand-built features, and/or
  - Implicitly learned representations
- Train classification model

How do we make predictions?

- Apply the trained classifier to each test instance in a clustering step
  - **Closest-first clustering**
    - For mention *i*, classifier is run backwards through prior *i*-1 mentions
    - First antecedent with probability > 0.5 is selected and linked to *i*
  - **Best-first clustering**
    - Classifier is run on all possible *i*-1 antecedents
    - Mention with highest probability is selected as the antecedent for *i*

Mention-Pair Architecture

- Advantage:
  - **Simplest** coreference resolution architecture
- Disadvantage:
  - **Doesn't directly compare candidate antecedents** with one another
  - **Considers only mentions**, not overall entities

How can we address these limitations?

- One option: The **Mention-Rank Architecture**
  - Directly compares antecedents with one another
  - Selects the highest-scoring antecedent for each anaphor
- How does this work?
  - For a mention *i*, we have:
    - Random variable y_i ranging over the values $Y(i) = \{1, \ldots, i-1, \varepsilon\}$
      - ε = dummy mention meaning *i* does not have an antecedent
  - At test time, for *i* the model computes a softmax over all possible antecedents
  - When training:
    - Use heuristics to determine the best antecedent for an anaphor (e.g., closest = best)
    - Or, learn more optimal ways to model latent antecedents using machine learning

Another Option: Entity-based Models

- Considers discourse entities, rather than individual mentions
- How does this work?
  - Have the model make decisions over clusters of mentions, where each cluster corresponds to an entity
  - Can be implemented using either feature-based or neural models

We know which architectures we can select …but how do we implement our coreference resolution models?

- Traditional machine learning models using manually-defined features
- Neural models

Feature-based Classification Models

- Common feature types:
  - Features of the candidate anaphor
  - Features of the candidate antecedent
  - Features of the relationship between the pair
- For entity-based models, this can also include:
  - Features of all mentions of the candidate antecedent's entity cluster
  - Features of the relation between the candidate anaphor and the mentions of the candidate antecedent in the entity cluster

What would be examples of these features?

- First word
- Head word
- Gender
- Named entity type
- Length
- Grammatical role
- Document genre
- …and many more!

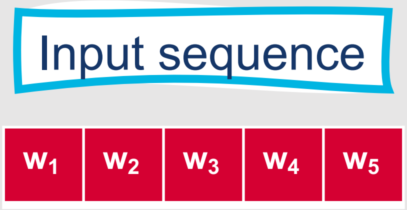
Neural Classification Models

- Generally end-to-end systems
- May not have a separate mention detection step
  - Instead, consider every possible text span of length < *k* as a possible mention
- Same overall goal as usual:
  - Assign to each span *i* an antecedent y_i ranging over the values $Y(i) = \{1, \ldots, i-1, \varepsilon\}$

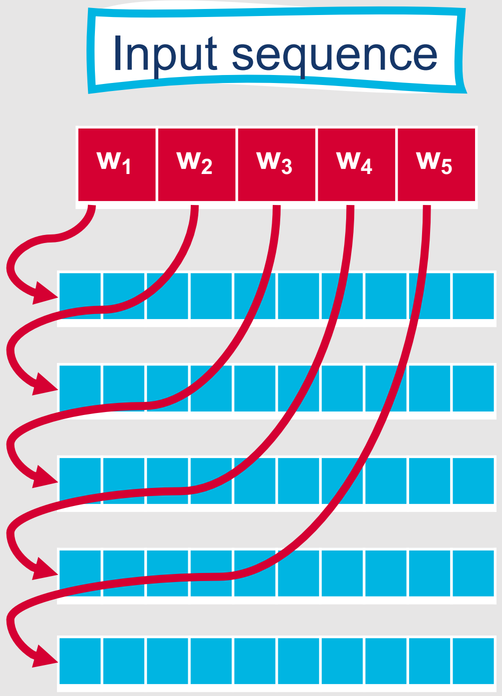
What goes on behind the scenes?

- For each pair of spans *i* and *j*, the system assigns a score $s(i,j)$ for the coreference link between the two
  - $s(i,j) = m(i) + m(j) + c(i,j)$
    - $m(i)$: Whether span *i* is a mention
    - $m(j)$: Whether span *j* is a mention
    - $c(i,j)$: Whether *j* is the antecedent of *i*
- The functions $m(\cdot)$ and $c(\cdot,\cdot)$ are computed using neural models:
  - $m(i) = w_m \cdot FFNN_m(g_i)$
  - $c(i,j) = w_c \cdot FFNN_c([g_i, g_j, g_i \circ g_j, \phi(i,j)])$
    - Where g_i is a vector representation of span *i*, and $\phi(i,j)$ encodes manually-defined characteristics of the relationship between *i* and *j*
    - Note that the exact definition of $c(i,j)$ may differ across models!

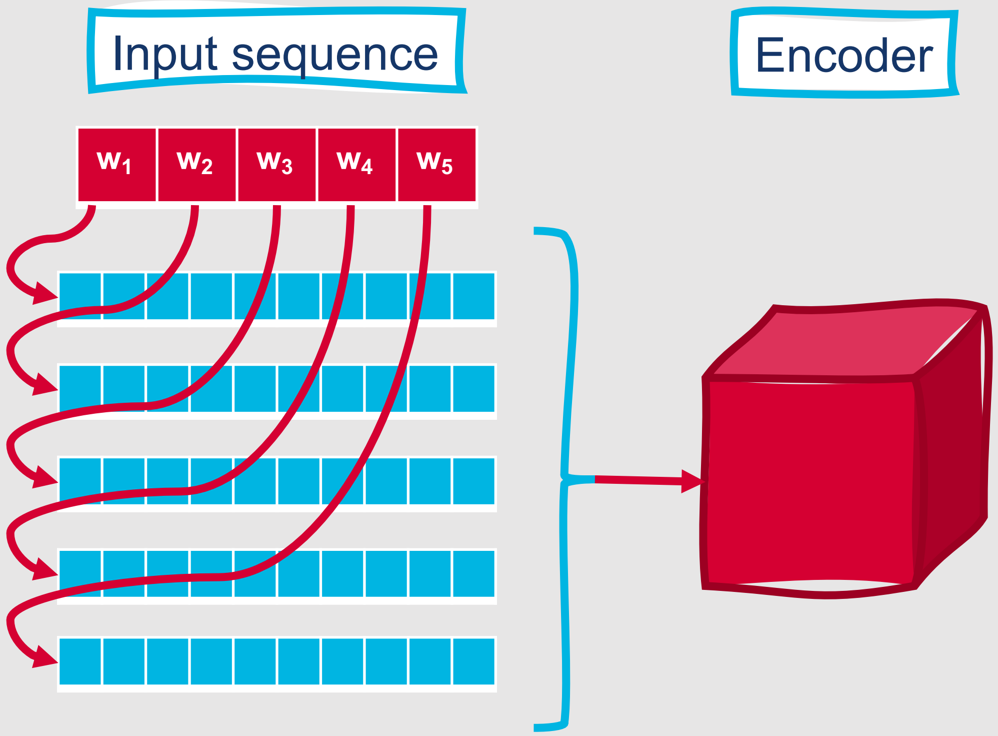
Altogether, a neural coreference resolution model might look like the following….



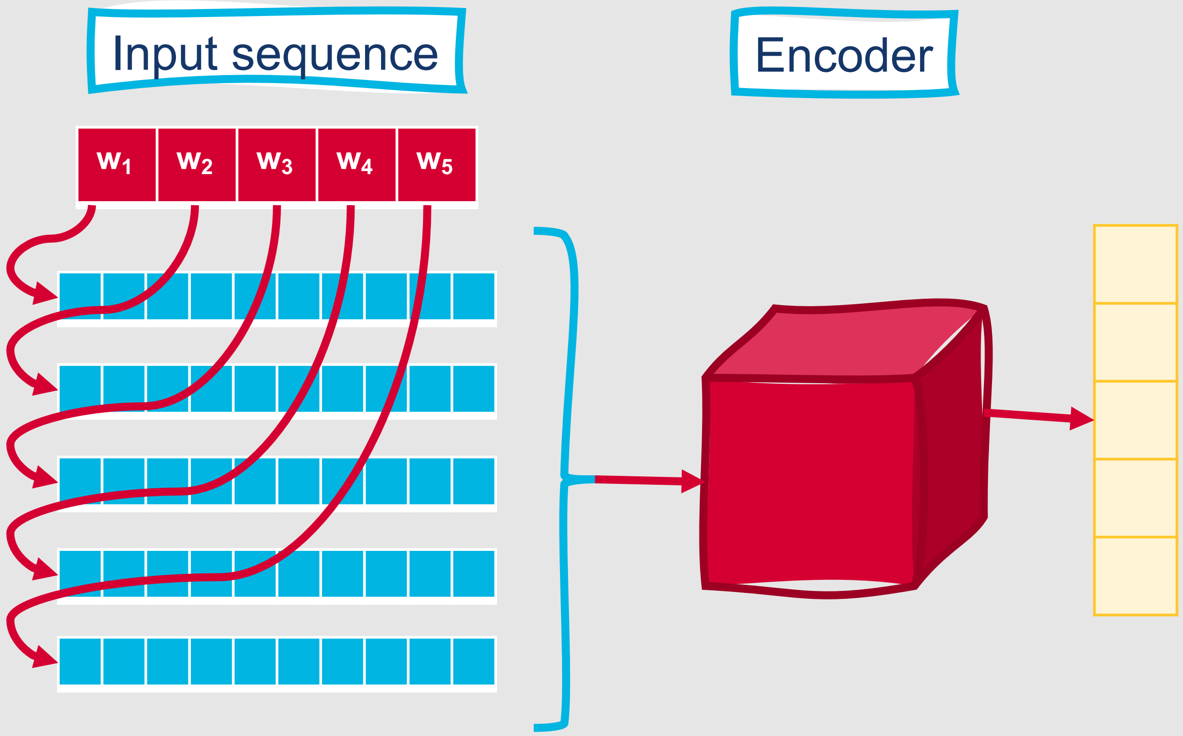
Altogether, a neural coreference resolution model might look like the following....



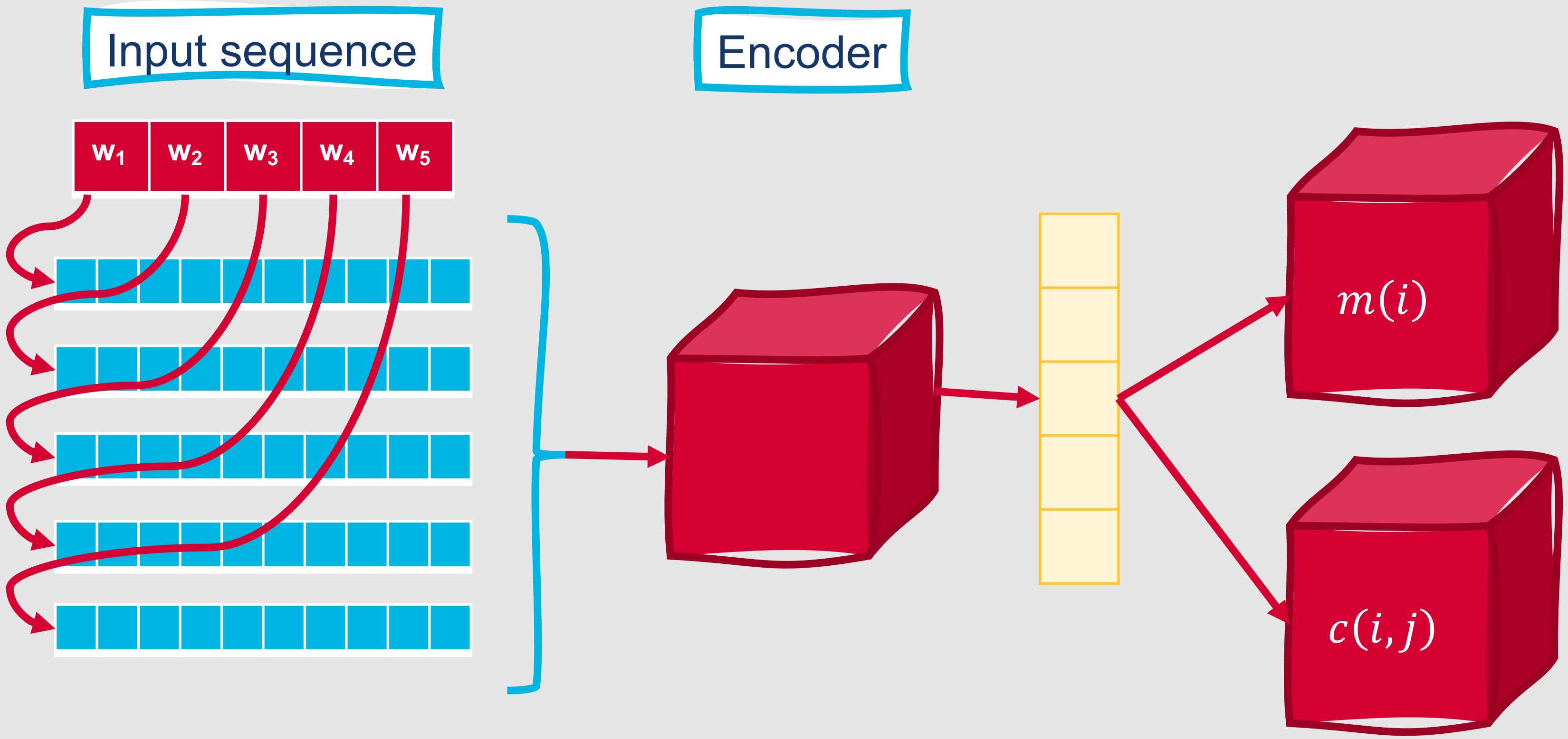
Altogether, a neural coreference resolution model might look like the following….



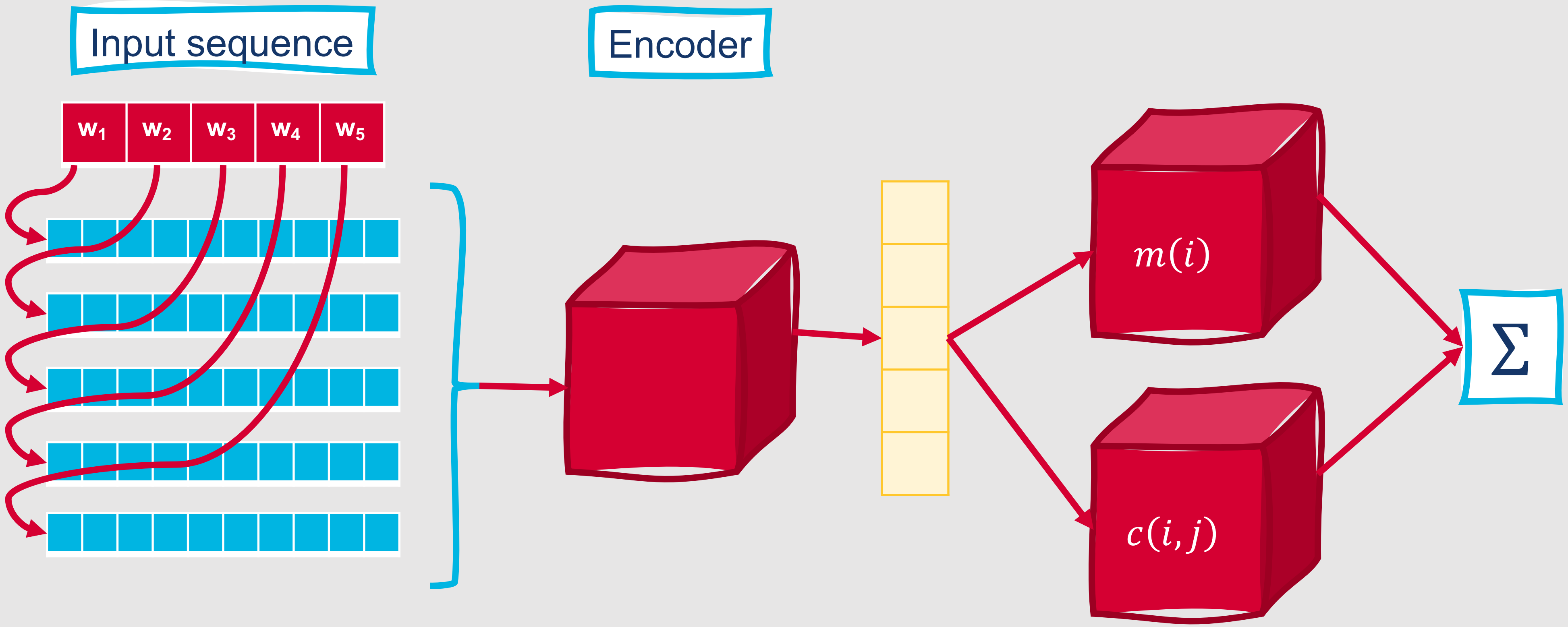
Altogether, a neural coreference resolution model might look like the following….



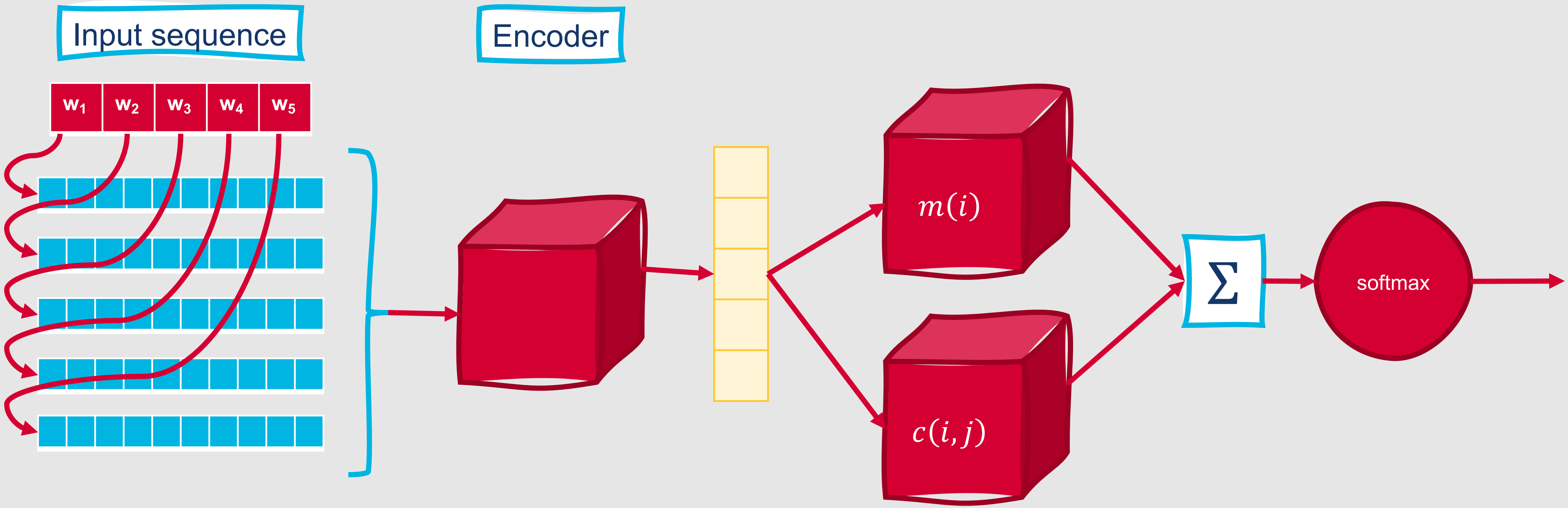
Altogether, a neural coreference resolution model might look like the following….



Altogether, a neural coreference resolution model might look like the following....

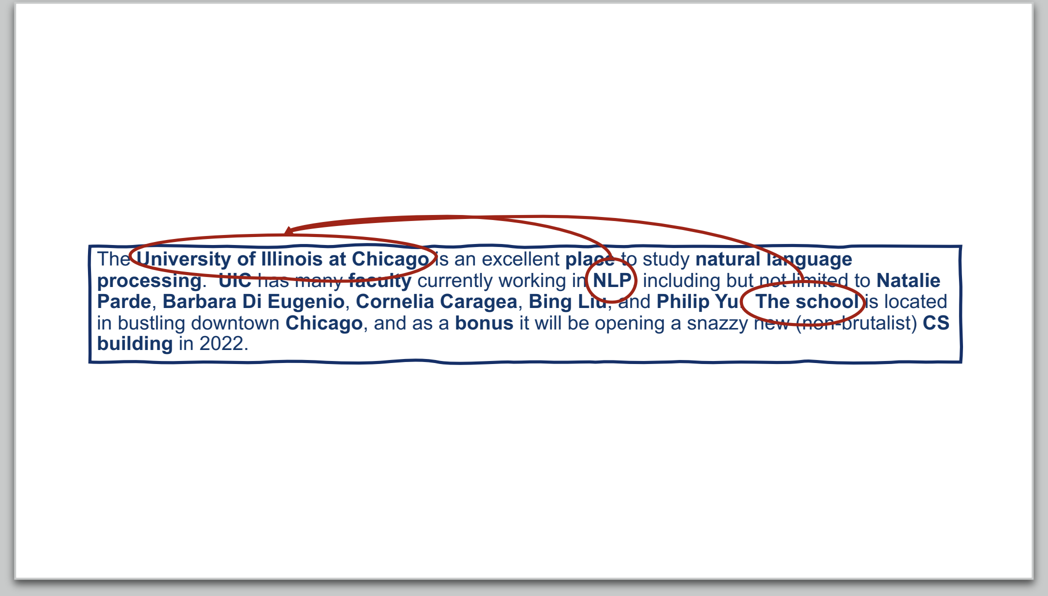


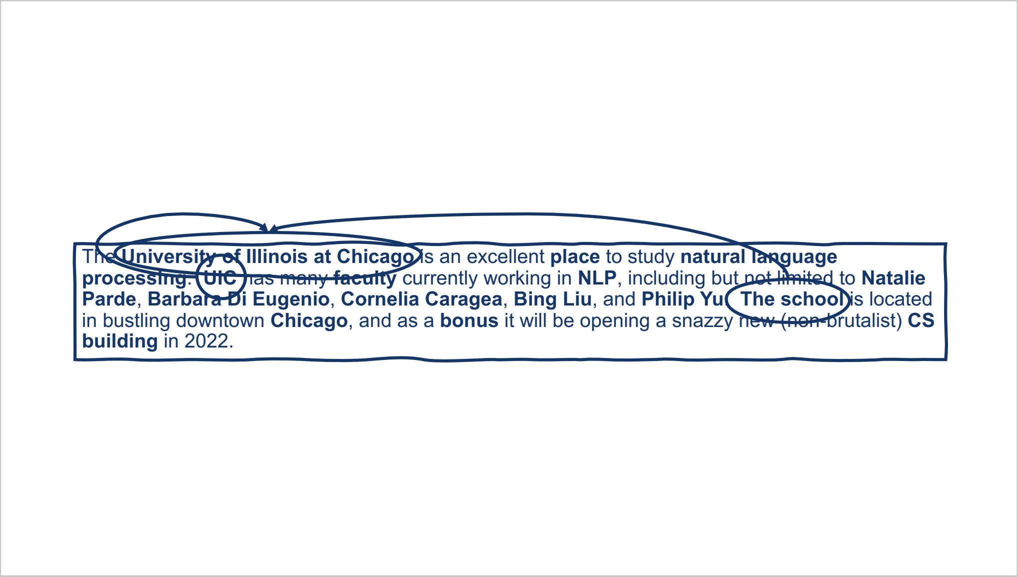
Altogether, a neural coreference resolution model might look like the following….



How do we evaluate coreference resolution models?

- Compare hypothesis coreference chains or clusters with a gold standard
- Compute precision and recall





How do we compute precision and recall?

- Several approaches:
  - **Link-based:** MUC F-measure
  - **Mention-based:** B^3

MUC F-Measure

- Message Understanding Conference (MUC)
- True positives = Common coreference links (anaphor-antecedent pairs) between hypotheses and gold standard
- Precision = # Common links / # Links in hypotheses
- Recall = # Common links / # Links in gold standard
- A couple downsides to this approach:
  - Biased towards systems that produce large coreference chains
  - Ignores singletons (no links to count)

B^3

- Mention-based
- True positives for a given mention, *i* = # Common mentions in hypothesis and gold standard coreference chain including *i*
- Precision for a given mention, *i* = TP / # Mentions in hypothesis coreference chain including *i*
- Recall for a given mention, *i* = TP / # Mentions in gold standard coreference chain including *i*
- Total precision and recall are the weighted sums of precision and recall across all mentions
  - Different weights correspond to different variations of the metric

So …where are we now?

- Still plenty of room for growth in coreference resolution!
- Recently, lots of interest in **Winograd Schema** problems
  - Coreference resolution problems that are:
    - Easy for humans to solve
    - Particularly challenging for computers to solve, due to their reliance on world knowledge and common sense reasoning

Winograd Schema Problems

- Winograd Schema problems are characterized by the following:
  - There are two statements that differ by only one word or phrase
  - There are two entities that remain the same across statements
  - A pronoun preferentially refers to one of the entities, but could grammatically also refer to the other
  - A question asks to which entity the pronoun refers
  - If one word/phrase in the question is changed, the human-preferred answer changes to the other entity

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was slower?

Nikolaos

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was slower?

Nikolaos

Nikolaos lost the race to Giuseppe because he was **faster**.

Who was faster?

Giuseppe

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was slower?

Nikolaos

Nikolaos lost the race to Giuseppe because he was **faster**.

Who was faster?

Giuseppe

Best way to solve Winograd Schema problems computationally?

- Currently, a mix of language modeling and external knowledge bases

Gender Bias in Coreference Resolution

- As with language modeling, coreference resolution systems can exhibit harmful gender biases
- How can we avoid these issues?
  - One solution: Increase sample size for underrepresented genders
    - Artificially: Generate gender-swapped versions of existing training corpora
    - Manually: Collect new, gender-balanced corpora
  - Other solutions?
    - Still very much an active research question!

Summary: Coreference Resolution

- **Coreference resolution** is the process of automatically identifying expressions that refer to the same entity
- This involves two tasks:
  - Identifying **referring expressions**
  - Clustering them into **coreference chains**
- Architectures for coreference resolution systems may be **mention-based** or **entity-based**, and may or may not compare potential **antecedents** with one another
- Models for coreference resolution may learn based on **manually defined features**, **neural features**, or a combination of the two
- Computing precision and recall for coreference resolution systems may be done using either **link-based** or **mention-based** methods
- **Winograd Schema** problems are particularly challenging coreference resolution tasks that rely on world knowledge and commonsense reasoning
- Care should be taken to avoid introducing harmful **gender biases** into coreference resolution systems