Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Coreference Annotation and Resolution

David R. Mortensen and Lori Levin

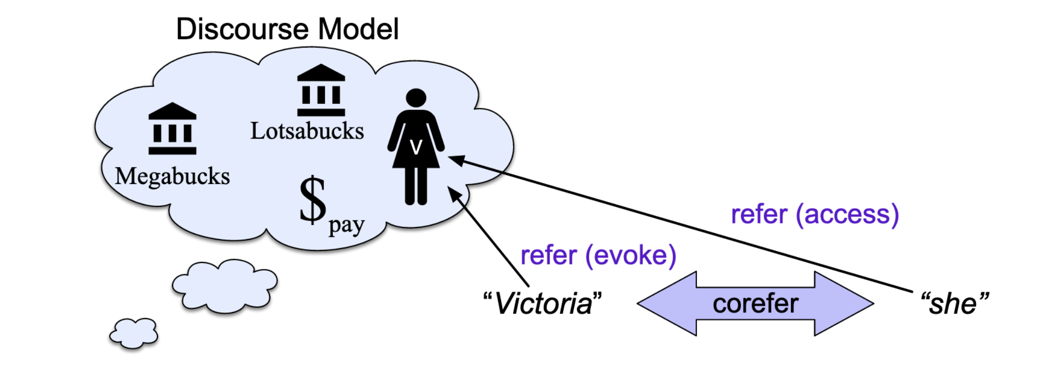
April 11, 2023

Language Technologies Institute

Example

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Mentions and Discourse Entities



Mentions evoke and access discourse entities.

The Coreference Resolution Task

Input: A text that has been divided into mentions. Not all noun phrases are mentions.

[Victoria Chen], [CFO of [Megabucks Banking]], saw [[her] pay] jump to [$2.3 million], as [the 38-year-old] became [[the company's] president]. It is widely known that [she] came to [Megabucks] from [[rival] Lotsabucks].

Output: A set of *coreference chains*, which are sets of mentions. Each color indicates a coreference chain here.

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks

Anaphor and Antecedent

Anaphor: A referring expression that co-refers with a previous referring expression. For example, *she* and *her* are anaphors that refer to *Victoria Chen*.

Antecedent: What the anaphor refers to. For example, *Victoria Chen* is the antecedent of the anaphors *she* and *her*.

Coreference Phenomena: Linguistic Foundations

Linguistic Foundations

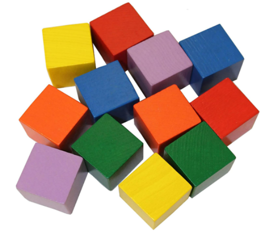
- Reference in a physical world
- Reference in a text

Reference in a blocks world

Hand me **a red block**.
Indefinite description

Hand me **the red block at the top**.
Definite description

Indefinite descriptions and definite descriptions **refer** to entities or sets of entities in the world, which consists of the blocks pictured on the right.



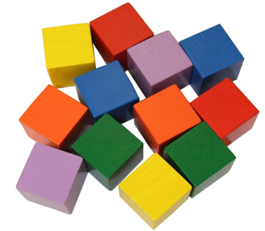
Indefinite Descriptions in Blocks Worlds

Hand me **a red block**: the words **red block** do not refer to one *unique* block, but the set of items satisfying the description **red block** is *identifiable*.

I have **a chartreuse block** *in my pocket*: The chartreuse block is not **identifiable** to the hearer. The speaker is introducing it as new to the world.

Definite Descriptions in Blocks Worlds

Hand me **the red block at the top center**: The red block at the top center is both unique and identifiable to the hearer.



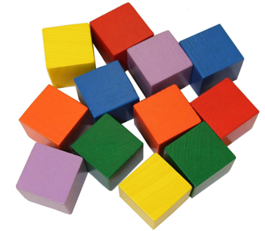
Reference can succeed or fail

Succeed: Hand me **a red block.**

The set of blocks meeting the description “red block” is identifiable.

Succeed: Hand me **the red block at the top center**.

The red block at the top center is unique and identifiable.



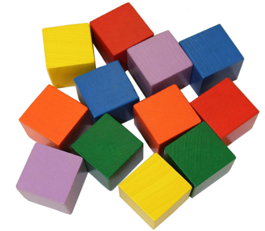
Reference can succeed or fail

Fail: Hand me **the red block.**

There is not a unique red block. There is more than one, but the use of a definite description indicates that the speaker expects the hearer to be able to identify a unique red block.

Fail: Hand me **the striped block.**

There is no striped block. The use of a definite description indicates that the speaker expects the hearer to be able to identify a striped block.



Accessibility Hierarchy, Ariel 1988

The type of a definite description is indicative of the extent to which its referent is *in focus* in the conversation or text.

Proper name plus modifier:	Dr. Garnet Redblock, the top center block
Full/Partial name:	Redblock
Long definite description:	The red block at the top center
Short definite description:	The red block
Distal demonstrative plus modifier:	That red block
Proximal demonstrative plus modifier:	This red block
Distal demonstrative with NP:	That block
Proximal demonstrative with NP:	This block
Distal demonstrative:	That
Proximal Demonstrative:	This
Pronoun:	it
Verb inflection:	Va**mos**
Zero:	Oishii (delicious)

Referring Expressions in Text

Referring expressions in text

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Referring Expressions: known in NLP as mentions

Referring expressions refer to discourse entities. Not every noun phrase is a referring expression.

1. Indefinite noun phrases
2. Definite noun phrases
3. Pronouns
4. Demonstrative pronouns
5. Zero pronouns
6. Names

Indefinite Noun Phrases

Speakers and writers generally use indefinite noun phrases to refer to things that their listener or audience does not yet know about. The indefinite noun phrase introduces a new discourse entity, which will be the beginning of a coreference chain. Indefinite noun phrases in English often start with *a*, *an*, unstressed *some*, or unstressed *this*.

- Mrs. Martin was so very kind as to send Mrs. Goddard *a beautiful goose*.
- He had gone round one day to bring her *some walnuts*.
- I saw *this beautiful cauliflower today*.

Definite Noun Phrases

Definite noun phrases in English may start with *the*, stressed *this*, *that*, *these*, or *those*. Speakers or authors use definite noun phrases to refer to things that the listener or audience can already identify:

- **Identifiable because it was mentioned previously:** these would not be at the beginning of a coreference chain.
  It concerns a white stallion which I have sold to an officer. But the pedigree of *the white stallion* was not fully established.
- **Identifiable from shared context or general knowledge:** these could be anywhere in a coreference chain
  I read about it in *the New York Times*.
  Do you know where *the car keys* are?

Pronouns

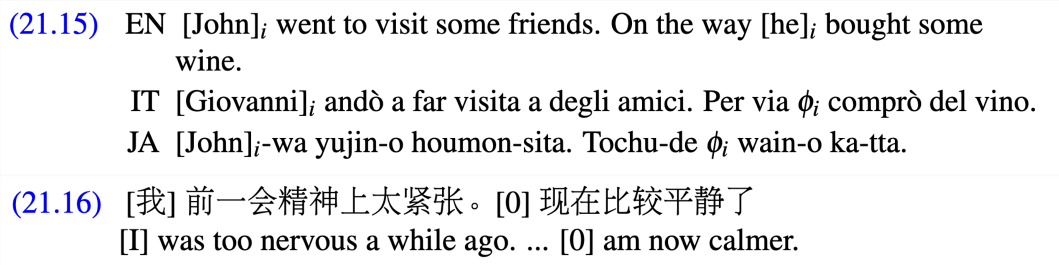
Pronouns refer to discourse entities that are known to the hearer or audience, and are also *in focus*. When something is in focus, you don't need to name it again, and you can shorten it to a pronoun. Pronouns would not generally be at the beginning of a coreference chain.

Anaphoric pronoun: the pronoun is after the noun it refers to:
Emma smiled and chatted as cheerfully as *she* could.

Cataphoric pronouns: the pronoun is before the noun it refers to:
Even before *she* saw *it*, Dorothy had been thinking about the Emerald City every day.

Pronoun that is bound by a quantifier:
Every student finished *their* homework.

Zero Pronouns



(21.15) EN [John]$_i$ went to visit some friends. On the way [he]$_i$ bought some wine.

IT [Giovanni]$_i$ andò a far visita a degli amici. Per via ϕ_i comprò del vino.

JA [John]$_i$-wa yujin-o houmon-sita. Tochu-de ϕ_i wain-o ka-tta.

(21.16) [我] 前一会精神上太紧张。[0] 现在比较平静了

[I] was too nervous a while ago. ... [0] am now calmer.

Names

Names can refer to old or new discourse entities.

- Miss Woodhouse certainly had not done him justice.
- International Business Machines sought patent compensation from Amazon; IBM had previously sued other companies.

Complications

Complications

- Inferrables
- Non-referring expressions

Inferrables

I went to a superb restaurant yesterday. *The chef* had just opened it.
Mix flour, butter, and water. Knead *the dough* until shiny.

The chef and the dough are new mentions, but they are inferrable from things that have been previously mentioned. The existence of a chef is inferrable from the existence of a restaurant. The dough came into existence as a result of mixing flour, butter, and water. Because these mentions are inferrable, they can make their first appearance in a definite referring expression, rather than an indefinite referring expression.

The coreference model that we will go over later in this lecture does not connect inferrables to the mentions that evoke them.

Noun phrases that are not referring expressions

Referring:

Janet has a car.

The car is red.

It is a Toyota.

Not Referring:

Janet doesn’t have a car.

# The car is red.

# It is a Toyota.

Non-referring noun phrases are not included in coreference chains. They will be classified as non-mentions.

# means *infelicitous* (not happy). There is a reference failure. When you say *Janet doesn’t have a car*, the mention of *a car* does not establish a discourse entity. Since there is no discourse entity, there is nothing for definite referring expressions (*the car*) or pronouns (*it*) to refer to.

> There was no Armenian genocide. And besides, they deserved it.
>
> —Turkish Nationalists

Other types of non-referring expressions

- Predicate nominals
- Appositives
- Expletive pronouns

Sidetrack on languages without definite and indefinite determiners

What about languages that don’t have *the* and *a*?

Languages in India and China generally don’t have words that are equivalent to *the* and *a*. How do they indicate that a noun phrase is identifiable or not identifiable to the hearer or audience?

Identifiable referring expressions may come earlier in the sentence, or in a special type of sentence.

- In languages of India, the word order is freer than in English, so the order could be SOV or OSV. The first noun phrase is likely to be more identifiable to the hearer.
- In Mandarin, if the words are in the default order (SVO), it is more likely that the subject is identifiable, and the object may or may not be identifiable. But if the word order is S-*ba*-O-V, then the object is usually identifiable, and if the word order is *you-yi-ge*-S-V-O, then the subject is not identifiable.

Languages without *the* and *a*, continued

Languages in India and China have ways to emphasize definiteness or indefiniteness

- using the number one (*ek* in India or *yi* with a classifier in China) to emphasize indefiniteness
- using a word that means *this* or *that* (like *zhe* and *na*) to emphasize definiteness

In languages that have definite and indefinite determiners, the indefinite determiner often started out as the number one, and the definite determiner often started out as a demonstrative determiner.

End Sidetrack on languages without definite and indefinite determiners

Coreference Datasets

Coreference Datasets

OntoNotes:

- English and Chinese: 1 million words each, annotated by hand, various spoken and written genres
- Arabic: about 300,000 words, newswire
- Does not label singletons (coreference chains of size one), which are 60-70% of all mentions
- Does not mark: generic NPs or appositive NPs

Other Datasets

- ISNotes: newswire; focuses on information status (old, new, inferrable), includes singletons; designed for identifying inferrables
- LitBank: 100 novels; includes singletons, quantified NPs, and negated NPs
- ARRAU: a variety of spoken genres; includes features for generic NPs, quantified, and non-referring NPs, includes singletons and inferrables

Winograd Schema Challenge

Based on Winograd (1992), the Winograd Schema Challenge consists of examples where world knowledge is needed in order to solve coreference:

- The city council denied the demonstrators a permit because *they* feared violence.
- The city council denied the demonstrators a permit because *they* advocated violence.
- Suzi gave Anne a kite for her birthday, but she already had a kite, so she returned *it*.
- Alex passed the controller to Sam because *their* turn was over.
- Alex passed the controller to Sam because *their* turn was next.

Gender Bias in Coreference

The biggest coreference dataset, OntoNotes, consists of articles from the 1990’s or before. Therefore, systems trained on OntoNotes have trouble when feminine pronouns are used to refer to words like *doctor* that were typically male at one time.

The secretary called the physician and told *him* about a new patient.

The secretary called the physician and told *her* about a new patient.

Mention Detection

Mention Detection

Input: a text, perhaps parsed or chunked into noun phrases

Output: the same text, indicating which noun phrases are mentions of discourse entities (as opposed to non-referential)

Methods: Build classifiers for referentiality and anaphoricity using supervised learning from training data. Rule-based heuristics may also be useful. However, modern systems are more likely to do mention detection jointly with coreference resolution.

Hard Problems:
You can make *it* in advance.
You can make *it* in Hollywood.

Architecture for Coreference Algorithms

Types of Coreference Algorithms

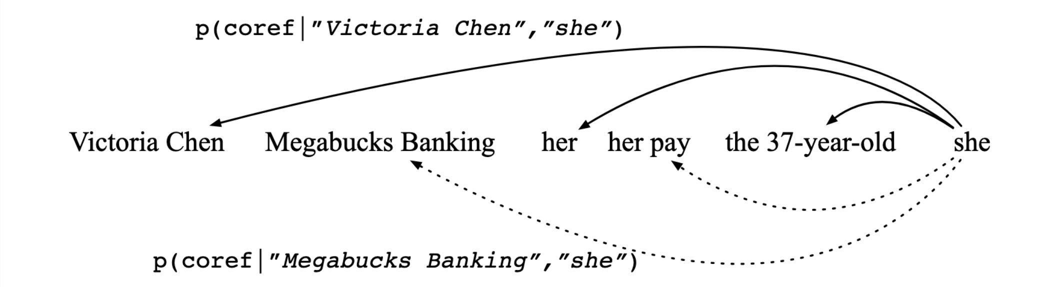
entity-based coreference decisions are made based upon each entity in the discourse model

mention-based decisions consider each mention independently. On another dimension, some algorithms use **ranking models** and others do not.

Mention-Pair Classification

inputs pair of mentions (candidate anaphor and candidate antecedent)

output probability that pair is coreferring


p(coref|"Victoria Chen","she")
Victoria Chen
Megabucks Banking
her
her pay
the 37-year-old
she
p(coref|"Megabucks Banking","she")

Training Mention-Pair Classifiers

Big Problem: Negative examples. Most expressions are not coreferrent with any given expression. How to avoid overwhelming the classifier with irrelevant negative examples?

Solution: A heuristic. Only consider as negative examples the spans that lie **between** the reference anaphor and it's reference antecedent.

The Clustering Step

- CLOSEST-FIRST CLUSTERING
    - Classifier is run right to left
    - First potential antecedent with probability > 0.5 is linked to span *i*
- BEST-FIRST CLUSTERING
    - Classifier is run on all potential antecedents
    - The antecedent with the highest probability is linked to span *i*

Strengths and Weaknesses of the mention-pair model

Strengths

- Simple to implement
- Easy to reason about

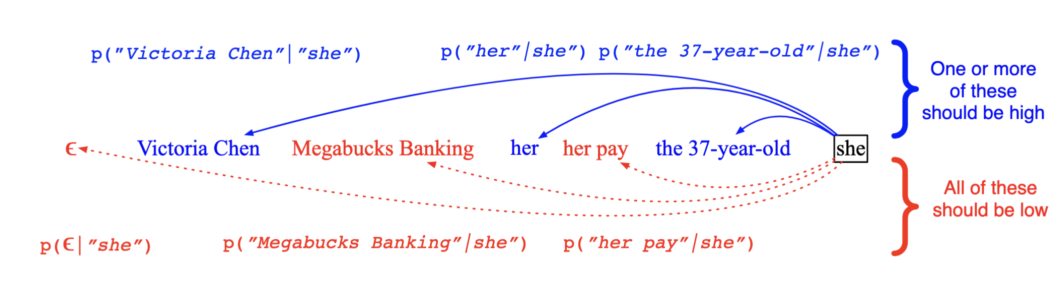
Weaknesses

- Does not compare candidate antecedents to one-another directly
- Looks only at mentions (not entities), ignoring the DISCOURSE MODEL

The mention-rank architecture addresses some of these limitations.

Mention-Rank Architecture

In the mention-rank architecture, candidate antecedents are compared to one another directly and the highest-scoring candidate is assigned to the anaphor:


p("Victoria Chen"|"she")
p("her"|she") p("the 37-year-old"|she")
One or more of these should be high
ϵ
Victoria Chen
Megabucks Banking
her
her pay
the 37-year-old
she
All of these should be low
p(ϵ|"she")
p("Megabucks Banking"|she")
p("her pay"|she")

What if *i* Is not Anaphoric?

- Early formulations of mention-rank simply chose the highest scoring candidate as the antecedent for span *i*
- However, *i* is often not anaphoric
- Solution? Random variable y_i ranging over the values $Y(i) = \{1, ..., i-1, \epsilon\}$
- What is ϵ? It means that *i* has no antecedent because
  - It is discourse-new (starts a new coref chain) OR
  - It is non-anaphoric
- At test time, the model computes one softmax over all the antecedents (**plus ϵ**), giving a probability for each candidate span (**or for none at all**)

It Is Also Possible to Build Entity-Based Models

- Instead of making decisions about individual mentions, make decisions about clusters of mentions (approximately, entities)
- Possible features
    - Size of cluster
    - Shape of cluster (sequence of types of mention)
- Clusters can be learned automatically (e.g., with RNNs over the sequence of cluster mentions)
- However, entity-based models do not perform dramatically better than mention-based models

Classifiers Using Hand-Built Features

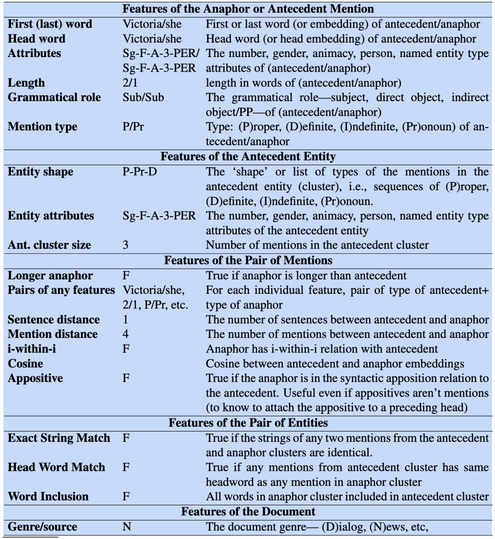
Hand-crafted features continue to play an important role in coreference resolution

Hand-Crafted Features for Coreference Resolution

There are five classes of features which are used in coreference resolution, whether they are rule-based, based on classical classifiers, or neural classifiers*

- Features of anaphor or antecedent mention
- Features of the antecedent entity
- Features of the pair mentions
- Feature of the pair entities
- Features of the document

*Hand-crafted features like Mention distance and Genre often supplement representation learning in coref.



	Features of the Anaphor or Antecedent Mention	
First (last) word	Victoria/she	First or last word (or embedding) of antecedent/anaphor
Head word	Victoria/she	Head word (or head embedding) of antecedent/anaphor
Attributes	Sg-F-A-3-PER/ Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of (antecedent/anaphor)
Length	2/1	length in words of (antecedent/anaphor)
Grammatical role	Sub/Sub	The grammatical role—subject, direct object, indirect object/PP—of (antecedent/anaphor)
Mention type	P/Pr	Type: (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun) of antecedent/anaphor
	Features of the Antecedent Entity	
Entity shape	P-Pr-D	The 'shape' or list of types of the mentions in the antecedent entity (cluster), i.e., sequences of (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun.
Entity attributes	Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of the antecedent entity
Ant. cluster size	3	Number of mentions in the antecedent cluster
	Features of the Pair of Mentions	
Longer anaphor	F	True if anaphor is longer than antecedent
Pairs of any features	Victoria/she, 2/1, P/Pr, etc.	For each individual feature, pair of type of antecedent+ type of anaphor
Sentence distance	1	The number of sentences between antecedent and anaphor
Mention distance	4	The number of mentions between antecedent and anaphor
i-within-i	F	Anaphor has i-within-i relation with antecedent
Cosine		Cosine between antecedent and anaphor embeddings
Appositive	F	True if the anaphor is in the syntactic apposition relation to the antecedent. Useful even if appositives aren't mentions (to know to attach the appositive to a preceding head)
	Features of the Pair of Entities	
Exact String Match	F	True if the strings of any two mentions from the antecedent and anaphor clusters are identical.
Head Word Match	F	True if any mentions from antecedent cluster has same headword as any mention in anaphor cluster
Word Inclusion	F	All words in anaphor cluster included in antecedent cluster
	Features of the Document	
Genre/source	N	The document genre— (D)ialog, (N)ews, etc,

Conjunctions of Features

- It is usually ineffective just to use the features as is
- Instead, it is common to take conjunctions of features (either by hand-engineering them or by discovering optimal conjunctions empirically)
    - Decision tree classifier
    - Random forest classifier

Neural Mention-Ranking Algorithm (after Jurafsky and Martin)

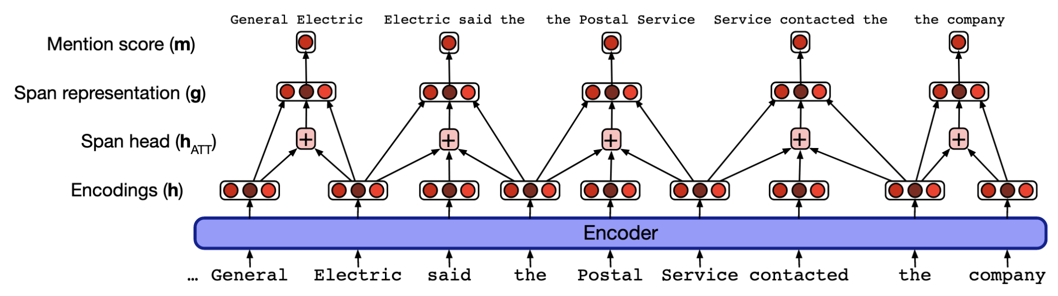
A Neural e2e-coref Algorithm

- Considers all $\frac{T(T-1)}{2}$ spans in a document (unigrams, bigrams, trigrams, etc.,), assigning a mention score to each
- Prunes mentions based on this score
- Assigns coreference links to the mentions that remain

Task: for each span i, assign an antecedent y_i (random variable ranging over the values $Y(i) = \{1, ..., i-1, \epsilon\}$) where ϵ means that i is either discourse new or non-anaphoric.

Computation of Span Representation and Mention Score in e2e-coref

The computation of **m** from **g** from **h**$_{\text{ATT}}$ from **h** from text is shown below (for some bigrams and trigrams):



Computing $\mathrm{h}_{\mathrm{ATT}(i)}$

$$\alpha_t = \mathrm{w}_\alpha \cdot \mathrm{FFNN}_\alpha(\mathrm{h}_t) \tag{1}$$

where w is a learned weight vector and h_t is the hidden state. Attention score is then normalized into a distribution, via softmax:

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp(\alpha_k)} \tag{2}$$

Attention distribution is then used to create an attention-weighted sum of words in span i:

$$\mathrm{h}_{\mathrm{ATT}(i)} = \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \mathrm{w}_t \tag{3}$$

(This should be familiar from our discussion of self attention in Transformers).

How Do We Compute g?

$$\mathrm{g}_i = [\mathrm{h}_{\mathrm{START}(i)}, \mathrm{h}_{\mathrm{END}(i)}, \mathrm{h}_{ATT(i)}] \tag{4}$$

g_i is the concatenations of three embeddings from a BERT encoder:

- The first token in the span
- The last token in the span
- The “most important” token in the span.

We Can Finally Compute $m(i)$ and $c(i,j)$

Remember that $\mathbf{g}_i$ is the concatenation of the first, last, and attention-weighted tokens in span i:

$$\mathbf{g}_i = [\mathbf{h}_{\mathrm{START}(i)}, \mathbf{h}_{\mathrm{END}(i)}, \mathbf{h}_{ATT(i)}] \tag{5}$$

$m(i)$ and $c(i,j)$ are then computed from g_i and g_j using FFNNs:

$$m(i) = w_m \cdot \mathrm{FFNN}_m(\mathbf{g}_i) \tag{6}$$

$$c(i,j) = w_c \cdot \mathrm{FFNN}_c([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \odot \mathbf{g}_j]) \tag{7}$$

Where $\odot$ is being used to indicate element-wise multiplication.

Computing the coreference score

The system assigns a score $s(i, j)$ for the coreference link between span i and span j, then learns a distribution over the antecendents for span i:

$$P(y_i) = \frac{\exp(s(i, y_i))}{\sum_{y' \in Y(i)} \exp(s(i, y'))} \tag{8}$$

There are three components to the score $s(i, j)$;

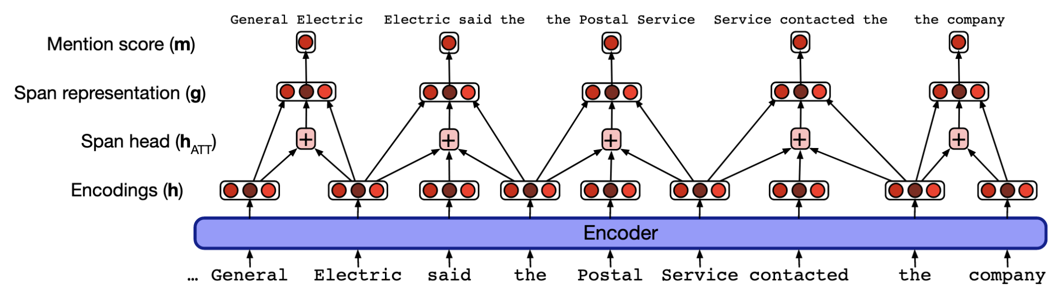
- Whether span i is a mention: $m(i)$
- Whether span j is a mention: $m(j)$
- Whether j is the antecedent of i: $c(i, j)$

$$s(i, j) = m(i) + m(j) + c(i, j) \tag{9}$$

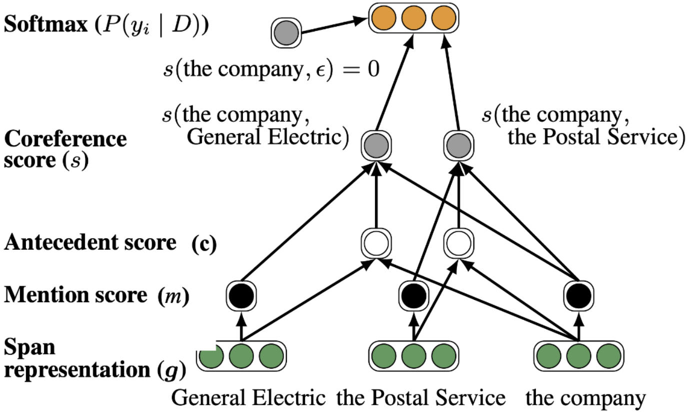
The score of $s(i, \epsilon)$ is fixed to 0.

Computation of Span Representation and Mention Score in e2e-coref

The computation of **m** from **g** from $\mathbf{h}_{\text{ATT}}$ from **h** from text is shown below (for some bigrams and trigrams):



Putting e2e-coref Together


Softmax ($P(y_i \mid D)$)
$s(\text{the company}, \epsilon) = 0$
s(the company, General Electric)
s(the company, the Postal Service)
Coreference score (s)
Antecedent score (c)
Mention score (m)
Span representation (g)
General Electric
the Postal Service
the company

Evaluation of Coreference Resolution

Coreference is evaluated by comparing a set H of hypothetical chains or clusters to a set R of reference chains or clusters and reporting precision and recall

This Sounds Simple, but Is Not

No less that five metrics are commonly used to evaluate entity coreference:

- MUC F-measure (link-based)
- BLANC (link-based)
- B^3 (mention-based)
- CEAF (entity-based)
- LEA (link-based, entity-aware)

We will discuss just the MUC F-measure and B^3.

The MUC F-Measure Is a Link-Based Metric

- Based on the number of coreference **links** (mention-mention pairs) common to H and R
- PRECISION: $\frac{|H \cup R|}{|H|}$ (number of common links divided by the number of links in H)
- RECALL: $\frac{|H \cup R|}{|R|}$ (number of common links divided by the number of links in R)
- Biased towards systems that produce long chains
- Ignores singletons

B^3 Is a Mention-Based Metric

- Mention-based rather than link based

$$\text{Precision} = \sum_{i=1}^{N} w_i \frac{\text{\# of correct mentions in hypothesis chain containing entity}_i}{\text{\# of mentions in hypothesis chain containing entity}_i} \tag{10}$$

$$\text{Recall} = \sum_{i=1}^{N} w_i \frac{\text{\# of correct mentions in hypothesis chain containing entity}_i}{\text{\# of mentions in reference chain containing entity}_i} \tag{11}$$

w_i is a by-entity weight that can adjusted to create different versions of the metric

Gender Bias in Coreference

ML Models Have a Gender Bias Problem

Because ML models, including coref models, are trained on empirical data from real people, they tend to replicate the biases of those people.

That is, a “good” ML model will reproduce the distribution of labels in the data even if this data is biased in socially harmful ways.

Not infrequently, ML models (including coref models) will actually magnify bias (produce output that is more biased than the input.

An Example from WinBias

Zhao et al. (2018) built a dataset to test bias in coref called WinoBias. Here is an example:

1. The secretary called the physician$_i$ and told him$_i$ about a new patient [pro-stereotypical]
2. The secretary called the physician$_i$ and told her$_i$ about a new patient [anti-stereotypical]

Get 1. right more than 2. $\Rightarrow$ BIASED!

Data Augmentation is One Approach to Bias Mitigation

- Data augmentation means adding training data with specific properties in order to improve performance on some metric (in this case, WinoBias and friends)
- Most coref systems are trained on OntoNotes (which has relatively few feminine pronouns)
- Automatically create a new version of OntoNotes with the pronouns switched in gender
- Train on the concatenation of old OntoNotes and new OntoNotes
- Gender bias is largely eliminated

Questions?