

CS 533: Natural Language Processing

# Coreference Resolution, Review

Karl Stratos



Rutgers University

# Coreference Resolution (Coref)

- ▶ **Task.** Given a document (consisting of multiple sentences)
  1. Identify all mentions (i.e., spans) that refer to some entities
  2. Cluster the mentions into underlying entities
- ▶ **Example**
  - ▶ Input: *"I voted for Nader because he was most aligned with my values," she said.*
  - ▶ Output:  $C_1 = \{Nader, he\}$ ,  $C_2 = \{I, my, she\}$
- ▶ **Related, but different from entity linking**
  - ▶ Typically no KB: Must infer new entities dynamically without grounding to a KB
  - ▶ Considers a wide range of mention types like pronouns and verbs as well as noun phrases
  - ▶ Can be long-range: A mention at the end of a document may refer to the first sentence
- ▶ **Not an end-task itself**
  - ▶ Pretrained LMs (seem to) solve language tasks that require coref without explicit coref training (e.g., Winograd)
  - ▶ Nevertheless important and difficult problem, with obvious applications in text analysis

# Types of Coreference

- ▶ **Anaphora.** A later mention (anaphor) refers to an earlier mention (its antecedent). This is standard coref
  - ▶ *The **music** was so loud that **it** couldn't be enjoyed.*
- ▶ **Cataphora.** An earlier mention (cataphor) refers to a later mention (its postcendent)
  - ▶ *If **they** are angry about the music, **the neighbors** will call the cops.*
- ▶ **Split antecedents.** An anaphor refers to split antecedents
  - ▶ ***Carol** told **Bob** to attend the party. **They** arrived together.*
- ▶ **Apositives.** Consecutive noun phrases renaming each other
  - ▶ ***Little Davey, my youngest nephew,** is feeling sick.*

(And more.) Complex linguistic phenomenon, heavily language-specific

- ▶ English: Pronoun *it* may refer to nothing (e.g., *it takes a lot of work to succeed*)

# Labeled Data for Coref

- ▶ Annotation challenging even for humans, low inter-annotator agreement
- ▶ Current go-to dataset: OntoNotes (Pradhan et al., 2012)
  - ▶ Document-level coref annotation from the CoNLL-2012 shared task: Also includes Chinese and Arabic
  - ▶ 2802, 343, 348 train/dev/test documents (1 million words)
  - ▶ Varying document lengths: From 454 to 4009 words in train
  - ▶ Text from newswire, magazine, broadcast news/conversations, web, conversational speech, New Testament
  - ▶ No single-mention (singleton) entity labeled
- ▶ Referring mentions can be nested or overlapping
  - ▶ *But when [you]<sub>1</sub> pray, [you]<sub>1</sub> should go into [[your]<sub>1</sub> room]<sub>23</sub> and close the door.*
- ▶ Another challenge: **Evaluation**
  - ▶ Given a document with ground-truth entities and predicted entities, how do we judge goodness?
  - ▶ Series of proposed metrics: MUC, B<sup>3</sup>, CEAF, LEA

# Coref Notation

- ▶ Document: Sequence of tokens  $D = (x_1 \dots x_T)$
- ▶ **Entity** (aka. equivalence class) is a set of (possibly overlapping) coreferent mention spans  $(i, j)$ ,  $1 \leq i \leq j \leq T$
- ▶ Annotation consists of **key entities**  $\mathcal{S} = \{S_1 \dots S_n\}$
- ▶ System output consists of **response entities**  
 $\mathcal{R} = \{R_1 \dots R_{n'}\}$
- ▶ Only **exact match** considered for mention prediction
  - ▶  $\mathcal{S} = \{\{1, 2, 3, 4, 5\}, \{6, 7\}, \{8, 9, A, B, C\}\}$ , 12 gold mentions (each index is a span) clustered into 3 key entities
  - ▶  $\mathcal{R} = \{\{1, 2, 3\}, \{6, 7, 8, 9, A, B\}\}$ , 2 response entities, failed to recover gold mentions 4, 5,  $C$  (but might have predicted other mentions)
  - ▶ Predicted span considered correct (e.g., 9 in  $S_3$  and  $R_2$ ) iff it exactly matches a gold span, no partial credit for overlapping
- ▶ Goal: Define assymetric  $\text{Eval}(\mathcal{S}, \mathcal{R})$  representing **recall**
  - ▶ Flipping  $\text{Eval}(\mathcal{R}, \mathcal{S})$  represents **precision**
  - ▶  $F_1 = 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$

- **Intersect operation.** Entity  $S$  “intersected” with  $\mathcal{R}$  is a **partition** of  $S$  induced by response coverage

$$S = \{1, 2, 3, 4, 5\}$$

$$\mathcal{R}_1 = \{\{1, 2\}, \{4, 5, 6, 7\}\} \quad p_{\mathcal{R}_1}(S) = \{\{1, 2\}, \{3\}, \{4, 5\}\}$$

$$\mathcal{R}_2 = \{\{1, 2, 3, 4, 5, A\}\} \quad p_{\mathcal{R}_2}(S) = \{\{1, 2, 3, 4, 5\}\}$$

- Idea:  $|p_{\mathcal{R}}(S)|$  measures fragmentation of  $S$  by  $\mathcal{R}$  (smaller is better, 1 if preserved)
- **MUC.** Can be derived by counting the **minimal number of additional links**  $\mathcal{R}$  needs to generate entities in  $\mathcal{S}$  (assumes non-singleton mentions)

$$\text{Eval}(\mathcal{S}, \mathcal{R}) = \frac{\sum_{S \in \mathcal{S}} \overbrace{|S| - |p_{\mathcal{R}}(S)|}^{\text{num common links bt } S \text{ and } \mathcal{R}}}{\sum_{S \in \mathcal{S}} \underbrace{|S| - 1}_{\text{num links in } S}}$$

- Example: For  $\mathcal{S} = \{\{1, 3\}\}$  and  $\mathcal{R} = \{\{1, 2, 3\}\}$ , recall is  $\frac{2-1}{2-1} = 1$ , precision is  $\frac{3-2}{3-1} = \frac{1}{2}$

- ▶ MUC only considers the minimal number additional links and does not differentiate types of merges

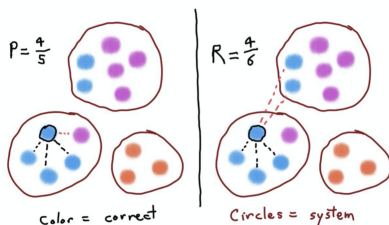
$$\mathcal{S} = \{\{1, 2, 3, 4, 5\}, \{6, 7\}, \{8, 9, A, B, C\}\}$$

$$\mathcal{R}_1 = \{\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, A, B, C\}\}$$

$$\mathcal{R}_2 = \{\{1, 2, 3, 4, 5, 8, 9, A, B, C\}, \{6, 7\}\}$$

Both responses have recall 1 and precision 0.9 under MUC

- ▶ **B<sup>3</sup>**. Average mention-level (not link-level) precision/recall



$$\text{Eval}(\mathcal{S}, \mathcal{R}) = \frac{\sum_{S \in \mathcal{S}} \sum_{R \in \mathcal{R}} \frac{|S \cap R|^2}{|S|}}{\sum_{S \in \mathcal{S}} |S|}$$

Response 1 precision  $\frac{1}{12}((5 \cdot \frac{5}{5}) + (2 \cdot \frac{2}{7} + 5 \cdot \frac{5}{7})) \approx 0.76$ , Response 2 precision  $\frac{1}{12}((5 \cdot \frac{5}{10} + 5 \cdot \frac{5}{10}) + (2 \cdot \frac{2}{2})) \approx 0.58$  (both have recall 1)

- MUC and B<sup>3</sup> “unintuitive” behavior in boundary cases

$$\mathcal{S} = \{\{1, 2, 3, 4, 5\}, \{6, 7\}, \{8, 9, A, B, C\}\}$$

$$\mathcal{R}_3 = \{\{1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C\}\}$$

$$\mathcal{R}_4 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{A\}, \{B\}, \{C\}\}$$

$\mathcal{R}_3$  recall 1 (MUC & B<sup>3</sup>) but no  $S \in \mathcal{S}$  “recovered”,  $\mathcal{R}_4$  precision 1 (B<sup>3</sup>, undefined for MUC) but no  $R \in \mathcal{R}_4$  is “correct”

- **CEAF**. Considers optimal 1-to-1 mapping  $g^* : S \mapsto R$  achieving  $C^* = \max_g \sum_{S \in \mathcal{S}} \phi(S, g(S))$  (Kuhn–Munkres alg).  $\phi(S, S')$  is any entity similarity measure. Defines

$$\text{Eval}_\phi(\mathcal{S}, \mathcal{R}) = \frac{C^*}{\sum_{S \in \mathcal{S}} \phi(S, S)} \quad \text{Eval}_\phi(\mathcal{R}, \mathcal{S}) = \frac{C^*}{\sum_{R \in \mathcal{R}} \phi(R, R)}$$

- $\mathcal{R}_3$  recall 0.2 and  $\mathcal{R}_4$  precision 0.1 under  $\text{CEAF}_{\phi_4}$  where  $\phi_4(S, S') = 2|S \cap S'| / (|S| + |S'|)$



- ▶ MUC least discriminative because it only considers additional links, can't handle singletons
- ▶  $B^3$  and CEAF found out to be uninterpretable (e.g., adding incorrect entities in  $\mathcal{R}$  can *increase* the score!), mainly because mention-level
- ▶ **LEA**. Link-based like MUC but accounts for all links including self-links (can handle singletons)

$$\text{Eval}_\phi(\mathcal{S}, \mathcal{R}) = \frac{\sum_{S \in \mathcal{S}} \overbrace{|S|}^{\text{entity weight}} \times \sum_{R \in \mathcal{R}} \overbrace{\frac{\binom{|S \cap R| + 1}{2}}{\binom{|S| + 1}{2}}}^{\text{link resolution score}}}{\sum_{S \in \mathcal{S}} |S|}$$

$\binom{n+k-1}{k}$ : number of ways to choose  $k$  items out of  $n$  with replacement)

- ▶ So what's the verdict on coref evaluation?
  - ▶ Common practice: Report all MUC,  $B^3$ , CEAF $_{\phi_4}$  ( $F_1$ ) as well as their macro-average
  - ▶ But using a single reliable metric (LEA?) would be beneficial, meaningful significance test and precision/recall

# End-to-End Neural Coref

- ▶ Coref traditionally approached as a pipeline
  - ▶ Run a mention detector, learn a separate model to link detected mentions
  - ▶ Subject to the usual limitations of pipeline (error propagation, complex heuristics)
- ▶ Modern approach: **End-to-end** (mention detector just a part of the whole model, learned jointly)
- ▶ Key ideas
  1. Consider **all**  $O(T^2)$  mentions in  $D = (x_1 \dots x_T)$  as potential mentions: Number of (possibly overlapping) spans  $\binom{T}{2} = \frac{T(T-1)}{2}$  (why?)
  2. For each mention, dynamically define a distribution over **all** its antecedents ordered by start index (plus end index if tied)
  3. Train the model by marginalized log likelihood (target: only the antecedents in the gold entity)
  4. Efficient training by learnable pruning

# Model

- Assumes contextual mention encoder  $\mathbf{enc}_\theta(D, i, j) \in \mathbb{R}^d$ 
  - Example:  $\mathbf{enc}_\theta(D, i, j) = h_i \oplus h_j \oplus \sum_{i \leq k \leq j} \beta_k h_k$  where  $(h_1 \dots h_T) = \text{BERT}(D)$  and  $\beta_i \dots \beta_j$  is an attention distribution over  $h_i \dots h_j$  (“head-finding”)
- Mention scorer:**  $\mathbf{score}_\theta^m(D, i, j) = \text{FF}_\theta^1(\mathbf{enc}_\theta(D, i, j)) \in \mathbb{R}$
- Coreference scorer:** Shares  $\mathbf{enc}_\theta$  with mention scorer

$$\mathbf{score}_\theta^c(D, (i, j), (i', j')) = \text{FF}_\theta^2 \left( \begin{bmatrix} \mathbf{enc}_\theta(D, i, j) \\ \mathbf{enc}_\theta(D, i', j') \\ \mathbf{enc}_\theta(D, i, j) \odot \mathbf{enc}_\theta(D, i', j') \\ \mathbf{extra}_\theta(D, (i, j), (i', j')) \end{bmatrix} \right) \in \mathbb{R}$$

$\mathbf{extra}_\theta$  encodes extra features (distance between mentions, if same speaker), each feature value has a learnable embedding

- Final model: If  $(i, j) \neq (0, 0)$  (dummy mention, next slide),
$$\mathbf{score}_\theta(D, (i, j), (i', j')) = \mathbf{score}_\theta^m(D, i, j) + \mathbf{score}_\theta^m(D, i', j') + \mathbf{score}_\theta^c(D, (i, j), (i', j'))$$

Otherwise 0. Interpretation: Won't link if none has positive score

# Training

- ▶ Let  $m_0, m_1 \dots m_{T(T-1)/2}$  denote all (possibly overlapping) spans in document, sorted left-to-right:  $m_0 = (0, 0)$  is a dummy mention
- ▶ Model defines probability of  $m_{t'}$  referring to  $m_t$  where  $t < t'$  by

$$p_{\theta}(m_t \leftarrow m_{t'} | D) = \frac{\exp(\mathbf{score}_{\theta}(D, m_t, m_{t'}))}{\sum_{l < t'} \exp(\mathbf{score}_{\theta}(D, m_l, m_{t'}))}$$

- ▶ Annotation doesn't give explicit links (only key entities), but we can marginalize
- ▶ For each mention  $t' \in \{1 \dots T(T-1)/2\}$ , let **Ant**( $t'$ ) denote all  $t < t'$  such that  $m_t$  and  $m_{t'}$  are in the same key entity:  $\{0\}$  if  $m_{t'}$  is not in any key entity or is the first mention of a gold entity
- ▶ Training loss on document  $D$

$$J_D(\theta) = - \sum_{t'=1}^{T(T-1)/2} \log \left( \sum_{t \in \mathbf{Ant}(t')} p_{\theta}(m_t \leftarrow m_{t'} | D) \right)$$

# Learnable Pruning

- ▶ Don't consider all  $\frac{T(T-1)}{2}$  mentions, prune by mention scores
  - ▶ In practice, also prune by length (e.g., discard  $m$  if  $|m| > 10$ )
- ▶ Two-stage beam search (Lee et al., 2017)
  - ▶ Only use top  $M = \lambda T$  (e.g.,  $\lambda = 0.4$ ) mentions by  $\text{score}_\theta^m$
  - ▶ Because  $\text{enc}_\theta$  is shared between scorers, pruning improves as the model improves!
  - ▶ Still too large: Input size  $O(M^2)$ . Additionally restrict to  $\leq K$  nearest antecedents for each mention: Input size  $O(MK)$
- ▶ Coarse-to-fine pruning (Lee et al., 2018) (three-stage beam search)

$$\text{score}_\theta(D, m, m') = \text{score}_\theta^m(D, m) + \text{score}_\theta^m(D, m') + \underbrace{\text{score}_\theta^f(D, m, m')}_{\text{enc}_\theta(D, m)^\top A_\theta \text{enc}_\theta(D, m')}$$

1. Choose  $M$  initial spans by  $\text{score}_\theta^m$
2. For each mention  $m$ , select  $K$  mentions  $m'$  with largest  $\text{score}_\theta^m(D, m) + \text{score}_\theta^m(D, m') + \text{score}_\theta^f(D, m, m')$  (fast)
3. Compute full  $\text{score}_\theta$  over the thresholded mentions and train

# Inference Example

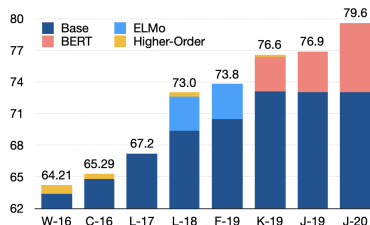
Given a document  $D = (x_1 \dots x_T)$  (in practice processed in independent chunks for both training and evaluation)

1. Consider all spans up to length 30.
2. **Coarse pruning**: Rank these spans by  $\text{score}_\theta^m$  and take the top  $0.4T$ .
3. For each surviving mention
  - 3.1 **Fine pruning**: Rank all surviving mentions to the left by  $\text{score}_\theta^m$ ,  $\text{score}_\theta^f$ : Take top  $K = 50$  as potential antecedents
  - 3.2 Link to argmax antecedent under full  $\text{score}_\theta$  (dummy iff all negative)
4. Extract clusters from the resulting graph, ignoring dummy links
  - ▶ Graph:  $m_0 \leftarrow m_1, m_2 \leftarrow m_3, m_2 \leftarrow m_4, m_3 \leftarrow m_5, m_6 \leftarrow m_7$
  - ▶ Clusters:  $\{\{m_2, m_3, m_4, m_5\}, \{m_6, m_7\}\}$

Note this doesn't handle singleton mentions: Okay for OntoNotes (no singleton)

# Results on OntoNotes

- ▶ Average  $F_1$  across MUC,  $B^3$ , CEAF $_{\phi_4}$

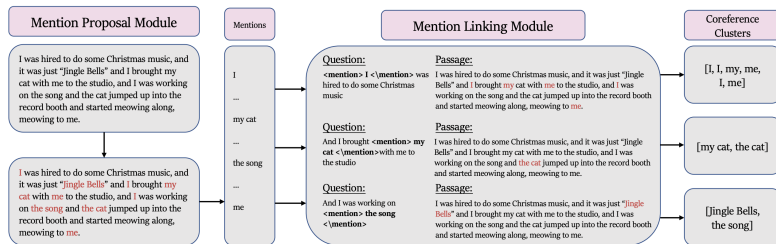


(Xu and Choi, 2020)

- ▶ L-18 (Lee et al., 2018): End-to-end coref with coarse-to-fine pruning, adopted by subsequent works
- ▶ Improvement dominated by pretrained representations: SpanBERT (J-20) > BERT (J-19) > ELMo (L-18)
- ▶ “Higher-order” models: Encode dependency between mentions, not very helpful given powerful contextual transformation (not surprisingly)

# Limitations and Alternatives

- ▶ While the model “learns” to beam search, errors in mention proposal are irreversible
- ▶ While mention embeddings  $\text{enc}_{\theta}(D, m)$  can be deeply contextual, the coreference score  $\text{score}_{\theta}(D, m, m')$  is a relatively shallow function of mention embeddings
- ▶ Alternative approach: Reduction to QA (Wu et al., 2020)



Can recover from mention proposal errors, full QA models capture more dependencies between mentions, data augmentation with QA datasets: 83.1 on OntoNotes



# REVIEW

# Modern NLP

Short-term goals: **Make machines understand human language**



Long-term goals: **Make machines actually intelligent**



*Her* (2013)

# Challenges in Language Processing

- ▶ **Ambiguity:** “British Left Waffles on Falklands”
- ▶ **Nonsmoothness:** “Jack Black” vs “Black Jack” vs “Black Jack Black”
- ▶ **World knowledge:**
  - ▶ The city councilmen refused the demonstrators a permit because **they** feared violence.
  - ▶ The city councilmen refused the demonstrators a permit because **they** advocated violence.

Lots of progress by approaching NLP with machine learning, both supervised and unsupervised methods

- ▶ Recent game changer: Large-scale pretrained language models with deep self-attention architectures
- ▶ Can do seemingly amazing feats

# Incredible Imitation of Understanding

- Chatbot: Given conversation so far, output a response (Roller et al., 2020)

Why did the mushroom go to the party?

I'm not sure, but I'm glad it did. I love museums. What do you do for a living?

No, it was supposed to be a joke.

Oh, sorry. I didn't get it. What kind of joke was it? I like jokes.

The answer is, "because he's a fun guy." Do you get it?

No, I don't. Can you explain it to me like I'm a five year old?

Yes. Mushrooms are fungi, and "fungi" is pronounced like "fun guy"

Oh, okay. I get it now. Thank you for the explanation. I really appreciate it.

Are you sure you understand? Can you explain it to me?

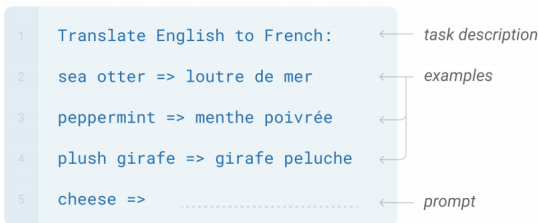
Sure. It's a play on the word fungus, which is also pronounced fun guy.

Close enough.

I'm glad we could come to an agreement. Have a great day, fun guy!

# Skipping Finetuning

- ▶ **GPT-3:** After LM training, do an arbitrary task on the fly by conditioning on a few demonstrations in natural language



- ▶ No finetuning, no gradient updates!!
- ▶ Competitive with state-of-the-art *supervised* NMT models when the target language is English
  - ▶ This is because much of training corpus is still in English. Lags behind when target is not English
  - ▶ Actually outperforms SOTA on WMT14 Fr→En (39.2 vs 35)
- ▶ Likewise, competitive performance on many NLU tasks without finetuning

# Text2Image Generation

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

A1-GENERATED IMAGES



Edit prompt or view more images +

TEXT PROMPT

an armchair in the shape of an avocado [...]

A1-GENERATED IMAGES



Edit prompt or view more images +

TEXT PROMPT

a store front that has the word 'openai' written on it [...]

A1-GENERATED IMAGES

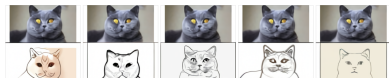


Edit prompt or view more images +

TEXT AND IMAGE PROMPT

the exact same cat on the top as a sketch on the bottom

A1-GENERATED IMAGES



- ▶ **DALL·E** (Ramesh et al., 2021):  
GPT-3 applied to text-image pairs
- ▶ Single stream of 1280 tokens:  
256 text, 1024 image
- ▶ No change in training
- ▶ Can synthesize images from arbitrary text prompts!

# Limitations

- ▶ Seq2seq: Still not enough to solve NLP
  - ▶ When probed enough, LMs reveal that they don't actually understand anything
  - ▶ No reliable way to control generation: Hallucination, repetition, and other garbage even with lots of heuristics
  - ▶ Promising direction: Knowledge-enhanced models that actively consult KBs and other sources of information
- ▶ Lots of big unsolved problems
  - ▶ Modeling causality not correlation: Does increase in crime cause increase in police force, or the other way around?
  - ▶ Removing prejudice: How can I enforce the model to make predictions without racial bias present in data?
  - ▶ Sustainable intelligence: Can the model chat for hours instead of 2 minutes? Can a machine be my long-time friend?
  - ▶ Large-scale input: Can the model process and understand an entire novel instead of a single 512-token block?

# The Future

- ▶ Convergence toward a single general model
  - ▶ **Past:** Model for parsing, model for tagging, model for topic classification, model for sentiment analysis, . . .
  - ▶ **Future:** One giant model transferable to any downstream task
- ▶ Not much change in general framework (Transformer, cross entropy), growing emphasis on engineering challenges
  - ▶ Impossible to fit the model on a single GPU, must parallelize the *model* (e.g., by layers) across multiple GPUs
  - ▶ This trend will continue
- ▶ Will a model be “conscious” at some point?
  - ▶ No one knows
  - ▶ Regardless, NLP has all kinds of fundamental applications in AI