Coreference

CS 685, Spring 2021

Advanced Topics in Natural Language Processing

http://brenocon.com/cs685

https://people.cs.umass.edu/~brenocon/cs685_s21/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- No class on Tues this week!
- Progress reports: due April 30.
- Final in-class presentations: May 3.
- HW3 cancelled / turned into some extra credit questions

- Also! HCI & NLP workshop tomorrow
  - Can write up a talk for HW3 extra credit
  - We'll post joining info to our slack (zoom/gather)

Noun phrase reference

Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities
e.g. real-world entities

Noun phrase reference

Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities
e.g. real-world entities

Noun phrase reference

http://harrypotter.wikia.com/wiki/Harry_Potter

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

Referring expressions reference discourse entities
e.g. real-world entities
(... or non-real-world)

Applications: text inference, search, etc.
- Who tried to kill Harry Potter?

Noun phrase reference

http://harrypotter.wikia.com/wiki/Harry_Potter

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

an **Entity** or **Referent** is a ~real-world object (discourse entity)
("HARRY_POTTER_CONCEPT")

Referring expressions a.k.a. **Mentions**
14 NPs are underlined above (are they all referential?)

Coreference: when referring mentions have the same referent.

Coreference resolution: find which mentions refer to the same entity.
I.e. cluster the mentions into **entity clusters**.

Applications: text inference, search, etc.
- Who tried to kill Harry Potter?

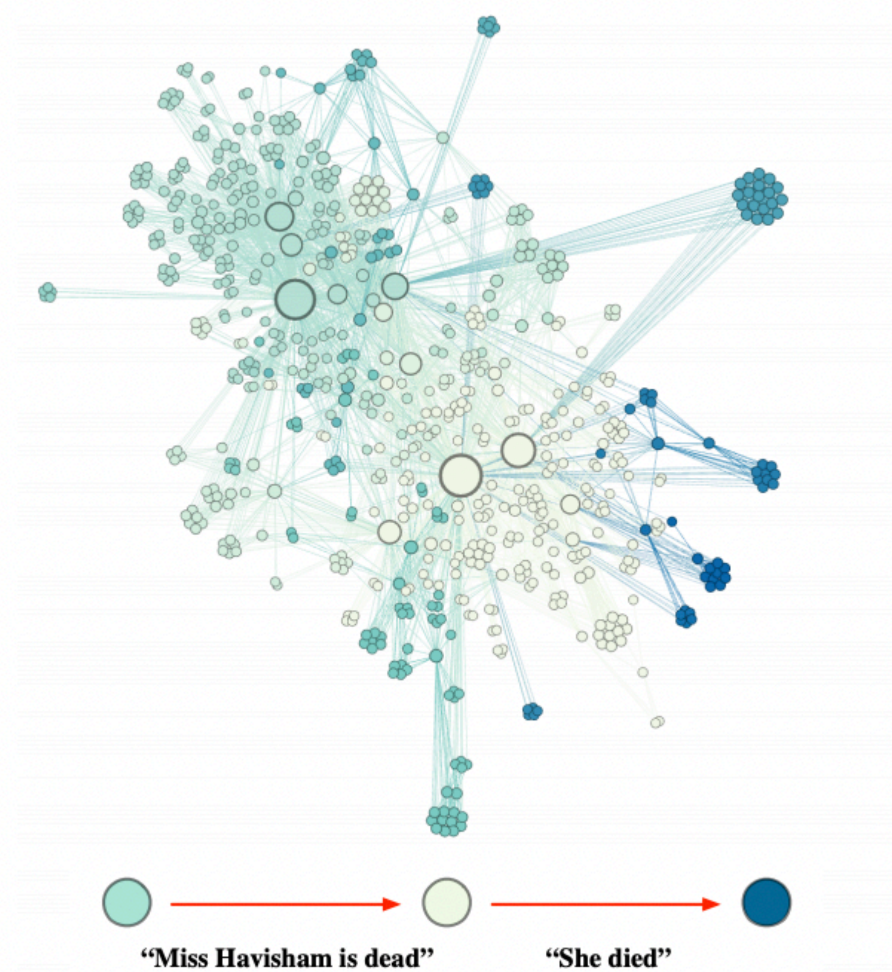
Measuring Information Propagation in Literary Social Networks

Matthew Sims
School of Information
UC Berkeley
msims@berkeley.edu

David Bamman
School of Information
UC Berkeley
dbamman@berkeley.edu

Abstract

We present the task of modeling information propagation in literature, in which we seek to identify pieces of information passing from character A to character B to character C, only given a description of their activity in text. We describe a new pipeline for measuring information propagation in this domain and publish a new dataset for speaker attribution, enabling the evaluation of an important component of this pipeline on a wider range of literary texts than previously studied. Using this pipeline, we analyze the dynamics of information propagation in over 5,000 works of English fiction, finding that information flows through characters that fill structural holes connecting different communities, and that characters who are women are depicted as filling this role much more frequently than characters who are men.



1 Introduction

- Application: analyze information exchange between characters in a book
  - Book-scale coreference is a prerequisite
  - *Sims and Bamman 2020*

Related tasks

- Within-document coreference
- Entity Linking — named entity recognition with coreference against an entity database (predict entity ID for text spans)
- Record linkage — entity coreference between structured databases

Noun phrase coreference

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer, some nested inside others

John Smith, CFO of Prime Corp. since 1986,

saw his pay jump 20% to $1.3 million

as the 57-year-old also became

the financial services co.'s president.

Exercise

Do within-document coreference in the following document by assigning the mentions entity numbers:

[The government]___ said [today]___ [it]___ 's going to cut back on [[[the enormous number]___ of [people]___]___ who descended on [Yemen]___ to investigate [[the attack]___ on [the " USS Cole]___]___]___. " [[[So many people]___ from [several agencies]___]___]___ wanting to participate that [the Yemenis]___ are feeling somewhat overwhelmed in [[their]___ own country]___. [Investigators]___ have come up with [[another theory]___ on how [the terrorists]___ operated]___. [[ABC 's]___ John Miller]___ on [[the house]___ with [a view]___]___. High on [[a hillside]___, in [[a run - down section]___ of [Aden]___]___]___, [[the house]___ with [the blue door]___]___ has [[a perfect view]___ of [the harbor]___]___. [American and Yemeni investigators]___ believe [that view]___ is what convinced [[a man]___ who used [[the name]___ [Abdullah]___]___]___ to rent [the house]___ [several weeks]___ before [[the bombing]___ of [the " USS Cole]___]___. " Early

Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*

More common in newswire, generally harder in practice

- Free variables
  - Smith saw *his pay* increase
- Bound variables
  - The dancer hurt *herself.*

More interesting grammatical constraints, more linguistic theory, easier in practice

"anaphora resolution"

Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.
- Syntactic cues
  - John bought himself a book.
  - John bought him a book.
- Lexical semantic cues
  - John saw Mary. She was eating salad.
  - John saw Mary. He was eating salad.
- Deeper semantics (world knowledge)
  - The city council denied the demonstrators a permit because they feared violence.
  - The city council denied the demonstrators a permit because they advocated violence.

- State-of-the-art coref uses with the first three (unless NNs are learning the 4th? Probably not…)

Coreference approaches

- Dialogue vs. documents
- Architectures
  - Mention-Mention linking
  - Entity-Mention linking
- Models
  - Rule-based approaches (e.g. *sieves*)
  - Supervised ML, end-to-end NNs
- Datasets: Ontonotes, CoNLL shared tasks (newspapers)
- Available systems (documents)
  - CoreNLP (many variants)
  - BookNLP (supervised, works on book-length texts)
  - Berkeley Coref ... etc. etc.

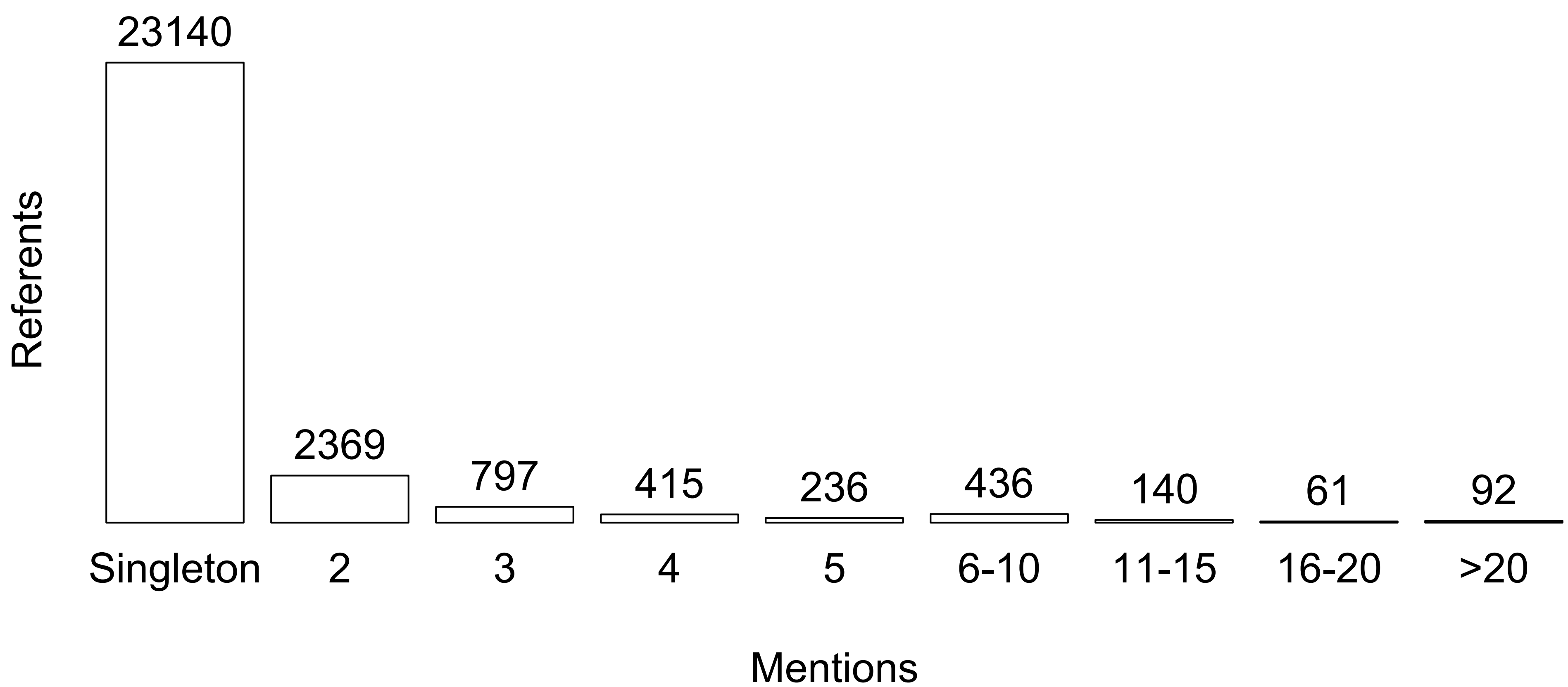


Figure 1: Distribution of referent lifespans in the 2012 OntoNotes development set.

[de Marneffe et al. 2015]

Supervised ML: Mention pair model

Hary Potter was a wizard. Lord Voldemort attempted to murder him.

- View gold standard as defining links between mention pairs
- Think of as binary classification problem: take random pairs as negative examples
- Issues: many mention pairs. Also: have to resolve local decisions into entities

Antecedent selection model

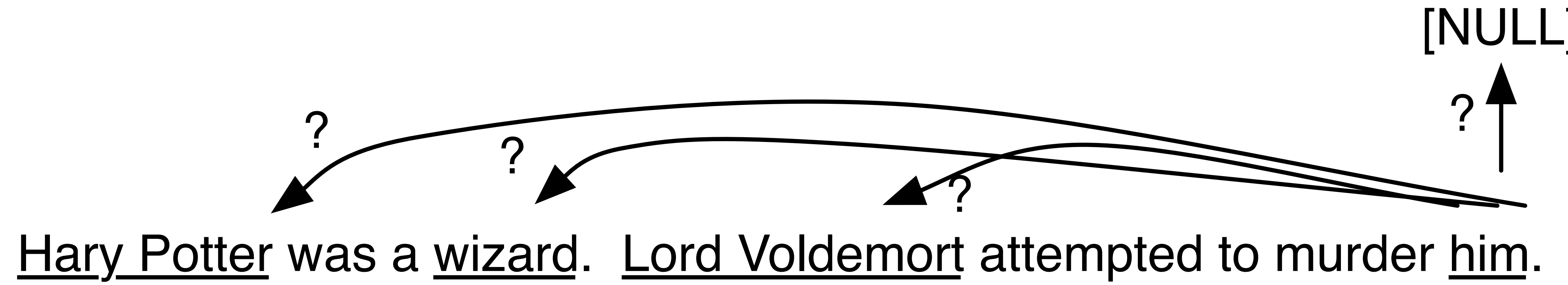
[NULL]

? ? ? ?

Hary Potter was a wizard. Lord Voldemort attempted to murder him.

- View as antecedent selection problem: which previous mention do I corefer with?
  - Makes most sense for pronouns, though can use model for all expressions
- Process mentions left to right. For the *n*'th mention, it's a *n*-way multi-class classification problem: antecedent is one of the *n-1* mentions to the left, or NULL.
  - Features are asymmetric!
  - Use a limited window for antecedent candidates, e.g. last 5 sentences (for news...)
- Score each candidate by a linear function of features.
  Predict antecedent to be the highest-ranking candidate.

Antecedent selection model



- Training: simple way is to process the gold standard coref chains (entity clusters) into positive and negative links. Train binary classifier.
- Prediction: select the highest-scoring candidate as the antecedent. (Though multiple may be ok.)
- Using for applications: take these links and form entity clusters from connected components [whiteboard]

Features for pronoun resolution

- English pronouns have some grammatical markings that restrict the semantic categories they can match. Use as features against antecedent candidate properties.
  - Number agreement
    - he/she/it vs. they/them
  - Animacy/human-ness? agreement
    - it vs. he/she/him/her/his
  - Gender agreement
    - he/him/his vs. she/her vs. it
- Grammatical person - interacts with dialogue/ discourse structure
  - I/me vs you/y’all vs he/she/it/they

Other syntactic constraints

- High-precision patterns
  - Predicate-Nominatives: “X was a Y …”
  - Appositives: “X, a Y, …”
  - Role Appositives: “president Lincoln”

Features for Pronominal Anaphora Resolution

- Preferences:
  - Recency: More recently mentioned entities are more likely to be referred to
    - John went to a movie. Jack went as well. He was not busy.
  - Grammatical Role: Entities in the subject position is more likely to be referred to than entities in the object position
    - John went to a movie with Jack. He was not busy.
  - Parallelism:
    - John went with Jack to a movie. Joe went with him to a bar.

Recency

- Not too recent, but can override
  - (1) John likes him
  - (2) John likes his mother
  - (3) John likes himself
  - (4) John likes that jerk
- Typical relative distances *[via Brian Dillon, UMass Ling.]*
  - reflexive < possessive < pronoun < anaphoric NP
- Salience: Subject of *previous* sentence is typical antecedent for a pronoun
  - Hobbs distance on constituent trees

Features for Pronominal Anaphora Resolution

- Preferences:
  - Verb Semantics: Certain verbs seem to bias whether the subsequent pronouns should be referring to their subjects or objects
    - John telephoned Bill. He lost the laptop.
    - John criticized Bill. He lost the laptop.
  - Selectional Restrictions: Restrictions because of semantics
    - John parked his car in the garage after driving it around for hours.
- Encode all these and maybe more as features

Features for non-pronoun resolution

- Generally harder!
  - String match
  - Head string match
    - I saw a green house. The house was old.
  - Substrings, edit distance
  - For names: Jaro-Winkler edit distance...

- *Cross-document coreference* and *entity linking*
  - Name matching: string comparisons
  - Contextual information

End-to-end neural coref

- Traditional architectures: mention detection, then mention linking
- End-to-end: directly compare all/most spans
  - For each span *i* (all T(T-1)/2 or T(maxwidth) of them),
    - Predict antecedent $y_i \in$ {NULL, 1, 2, … i-1}
  - $\mathbf{s_m}$ mention score: is the span a mention?
    - This is weirdly effective in a way specific to their training set, IMO
  - $\mathbf{s_a}$ antecedent score: are two spans linked?
- Naively O(T^4) runtime; aggressively prune based on s_m (mention detection as pruning)

$$
\begin{aligned}
& P(y_1, \ldots, y_N \mid D) \\
&= \prod_{i=1}^{N} P(y_i \mid D) \\
&= \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))}
\end{aligned}
$$

$$
s(i, j) = \begin{cases} 0 & j = \epsilon \\ s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i, j) & j \neq \epsilon \end{cases}
$$

[Lee et al. (2017)]

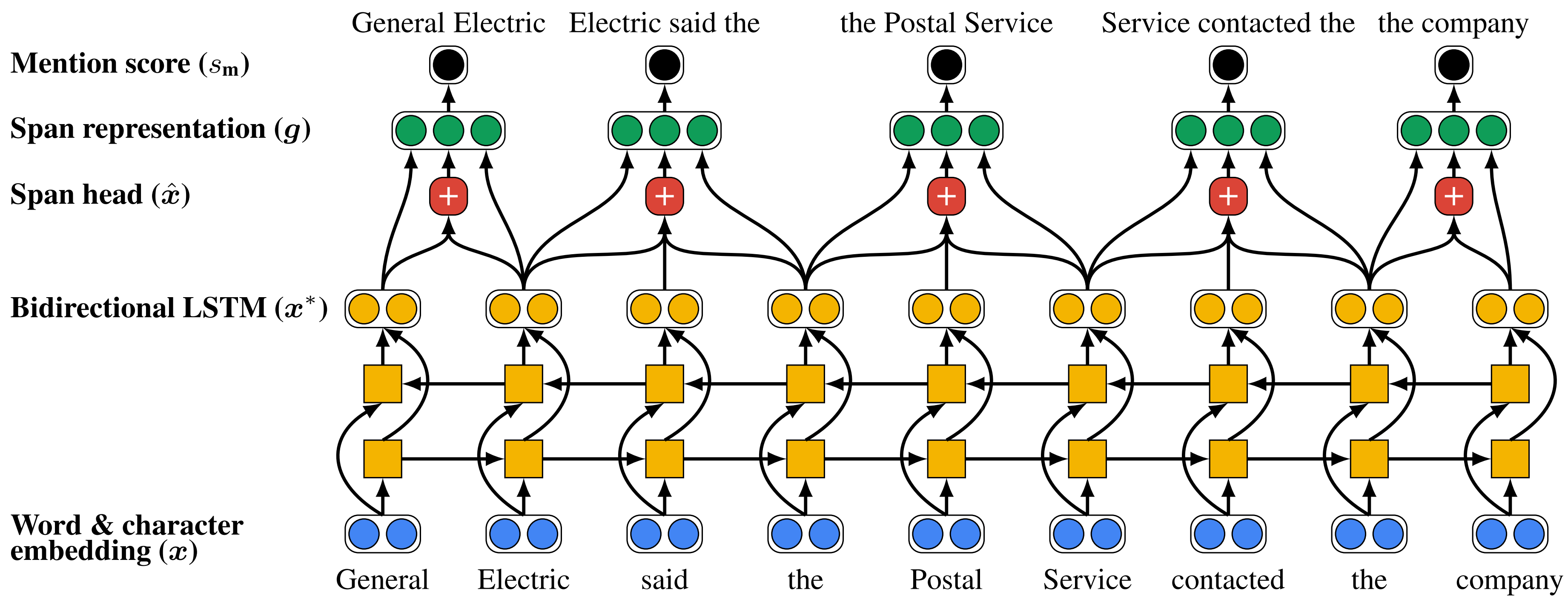


Figure 1: First step of the end-to-end coreference resolution model, which computes embedding representations of spans for scoring potential entity mentions. Low-scoring spans are pruned, so that only a manageable number of spans is considered for coreference decisions. In general, the model considers all possible spans up to a maximum width, but we depict here only a small subset.

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

$$\begin{aligned} & P(y_1, \ldots, y_N \mid D) \\ = & \prod_{i=1}^{N} P(y_i \mid D) \\ = & \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))} \end{aligned}$$

$$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_\mathrm{m}(i) + s_\mathrm{m}(j) + s_\mathrm{a}(i,j) & j \neq \epsilon \end{cases}$$

[Lee et al. (2017)]

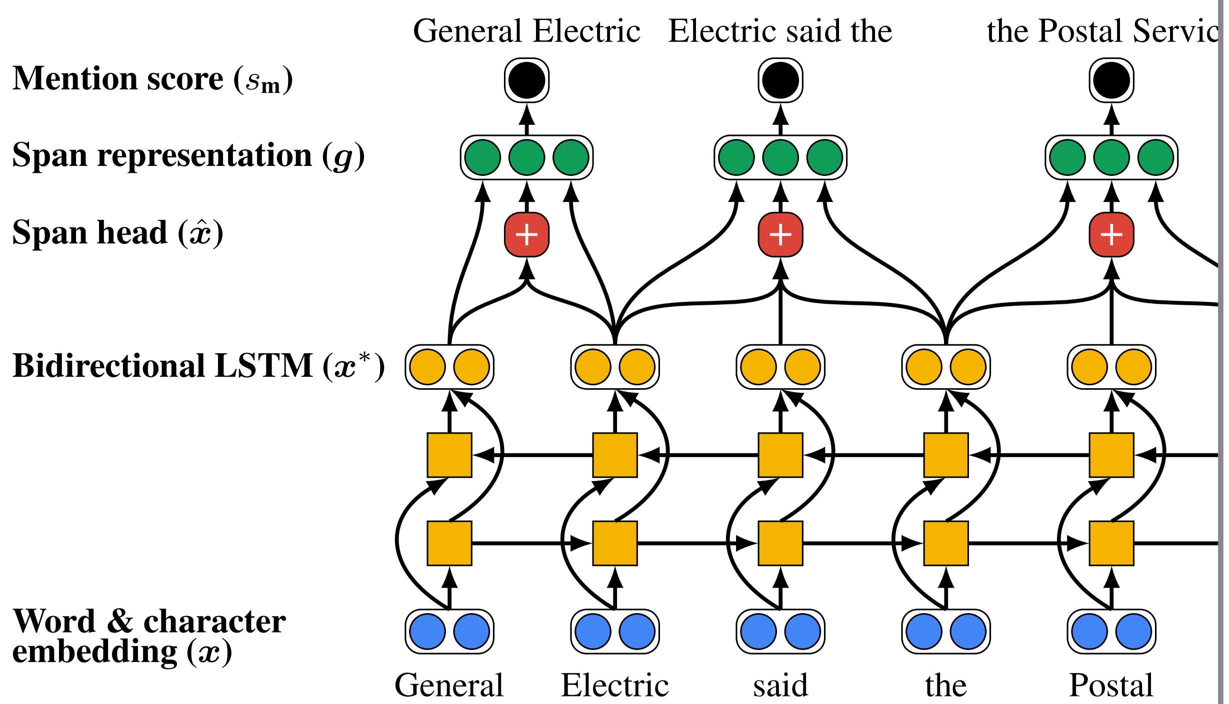


Figure 1: First step of the end-to-end coreference resolution mod sentations of spans for scoring potential entity mentions. Low-sc manageable number of spans is considered for coreference decisio possible spans up to a maximum width, but we depict here only a

Span representation uses attention mechanism, in order to get syntactic head info

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \text{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\text{START}(i)}, \boldsymbol{x}^*_{\text{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

$$\begin{aligned} &P(y_1, \ldots, y_N \mid D) \\ &= \prod_{i=1}^{N} P(y_i \mid D) \\ &= \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))} \end{aligned}$$

$$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \end{cases}$$

[Lee et al. (2017)]

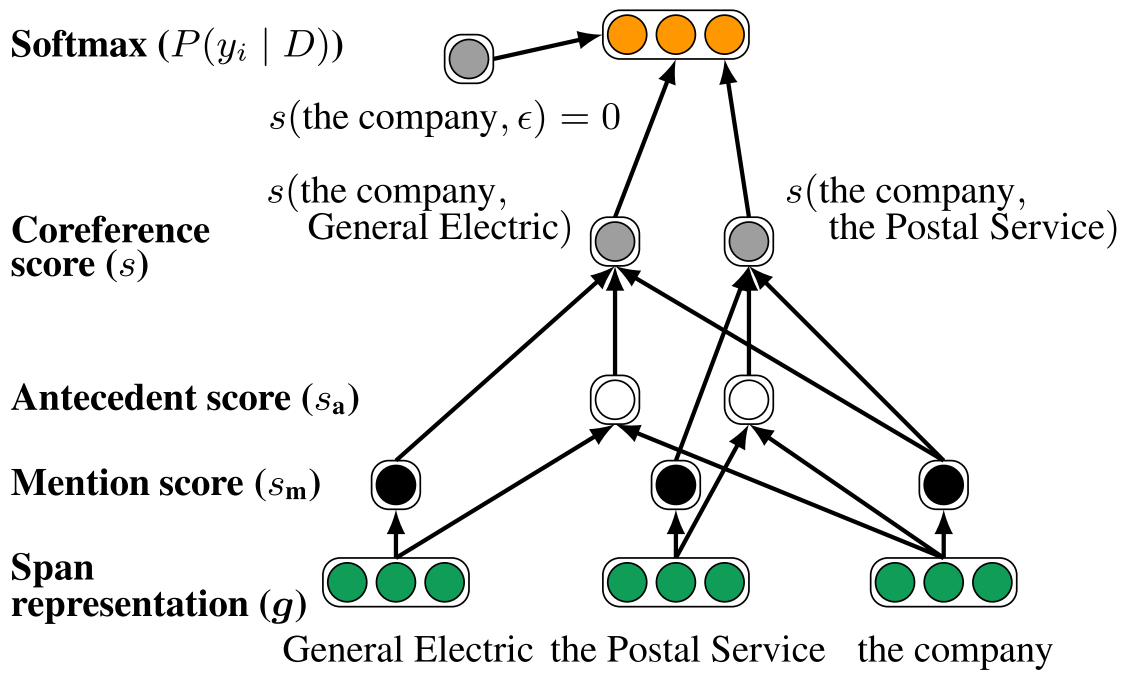


Figure 2: Second step of our model. Antecedent scores are computed from pairs of span representations. The final coreference score of a pair of spans is computed by summing the mention scores of both spans and their pairwise antecedent score.

Span representation uses attention mechanism, in order to get syntactic head info

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \text{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum\limits_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\text{START}(i)}, \boldsymbol{x}^*_{\text{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

$$P(y_1, \ldots, y_N \mid D) = \prod_{i=1}^{N} P(y_i \mid D) = \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))}$$

$$s(i,j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i,j) & j \neq \epsilon \end{cases}$$

[Lee et al. (2017)]

Learning

- Training data only specifies clustering information. Which antecedent is *latent* variable.
- Maximize marginal log-likelihood of observed data
  - (Compare to EM)

$$\log \prod_{i=1}^{N} \sum_{\hat{y} \in \mathcal{Y}(i) \cap \textsc{gold}(i)} P(\hat{y})$$

[Lee et al. (2017)]

Results, devset

	Avg. F1	Δ
Our model (ensemble)	69.0	+1.3
Our model (single)	67.7	
– distance and width features	63.9	-3.8
– GloVe embeddings	65.3	-2.4
– speaker and genre metadata	66.3	-1.4
– head-finding attention	66.4	-1.3
– character CNN	66.8	-0.9
– Turian embeddings	66.9	-0.8

- Add ELMO: 67.2 => 70.4 (dev set)

[Lee et al. (2017)]

Results

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

Table 1: Results on the test set on the English data from the CoNLL-2012 shared task. The final column (Avg. F1) is the main evaluation metric, computed by averaging the F1 of MUC, B^3, and $CEAF_{\phi_4}$. We improve state-of-the-art performance by 1.5 F1 for the single model and by 3.1 F1.

[Lee et al. (2017)]

But!

- Moosavi and Strube (2017): very heavy lexical overlap between CoNLL train/test splits. Are coref systems just memorizing domain-specific entity information? Lexical features overfit.
- How to make coreference work on *really* different domains: dialogue, web forums, books?
- How to expand to many languages? With low training data?

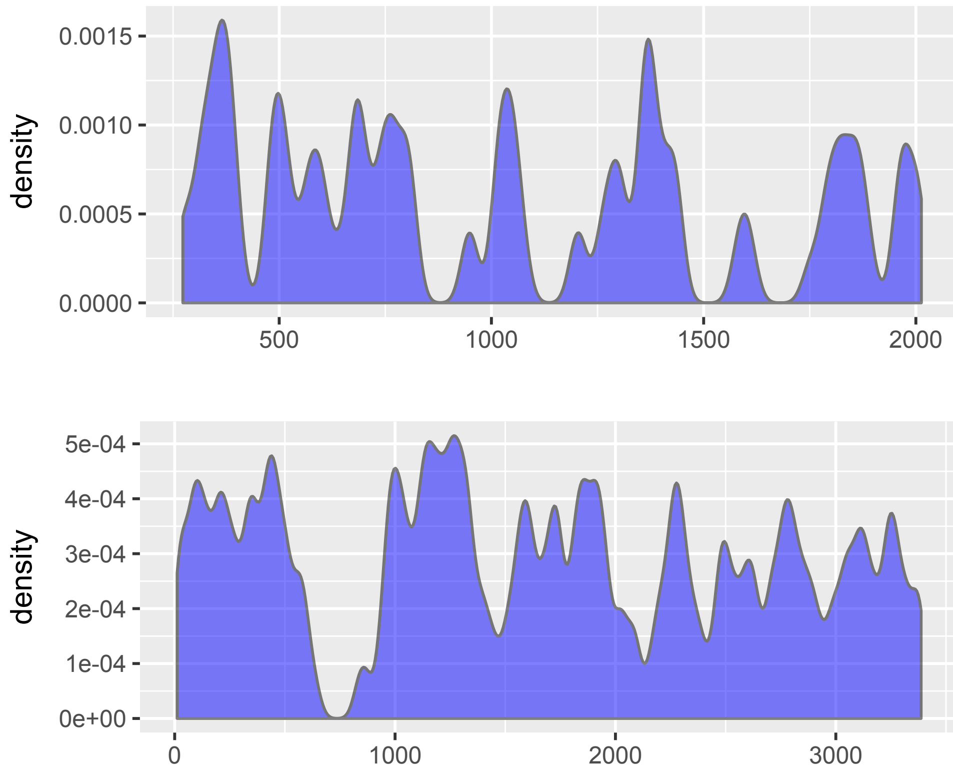


Figure 3: Long-range entities are bursty; the distributio of mentions over narrative time for the entity with the low est entropy (top; Basil Hallward in Wilde's *The Picture c Dorian Gray*) and highest entropy (bottom; the narrator i Swift's *Gulliver's Travels*).

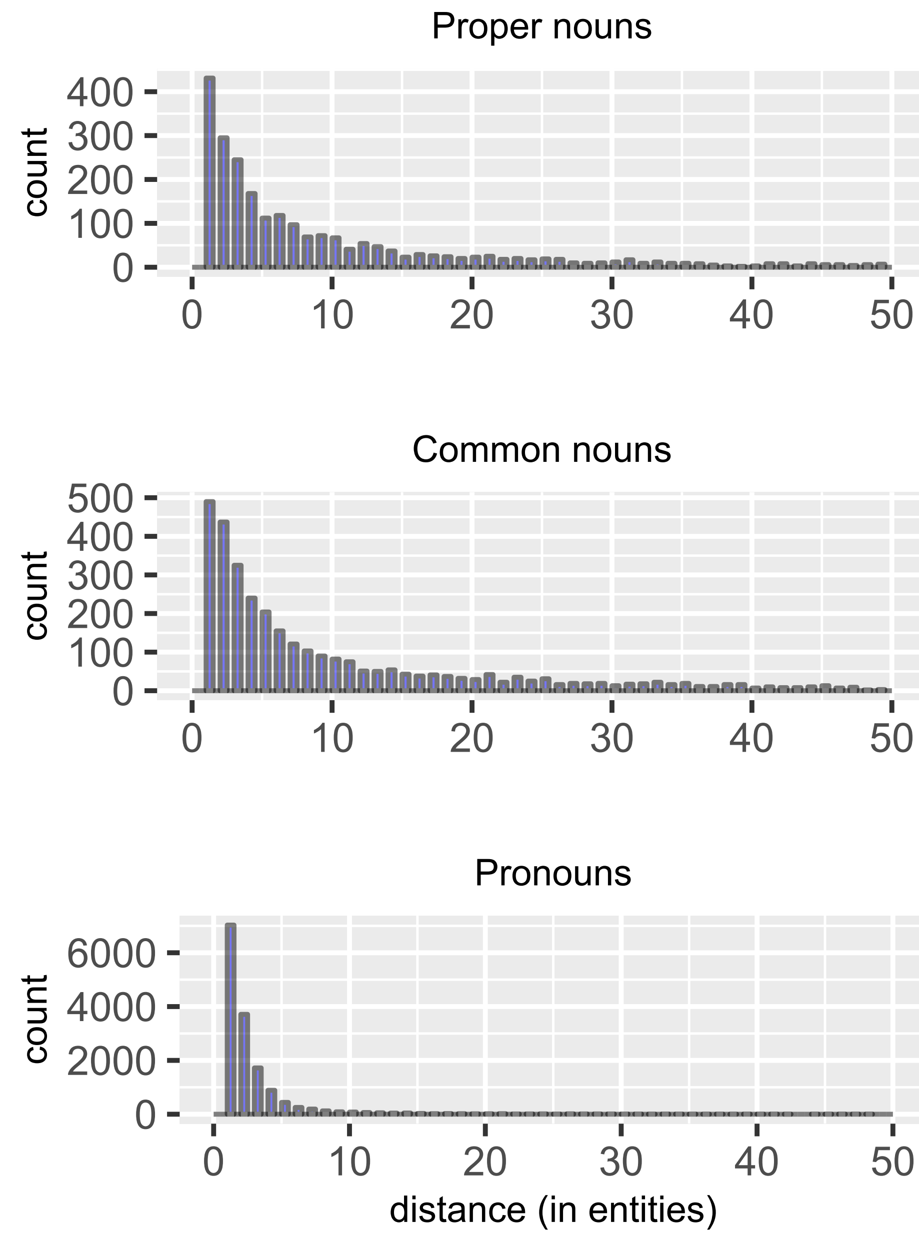


Figure 4: Distance to antecedent in entities.

- *LitBank* book coreference annotations
  - Selections from 100 books, avg 2000 tokens long
  - Command-line annotation software (!)
  - *Bamman et al. 2020*

- Two step pipeline
  - 1. Mention detection (span classification)
  - 2. Mention pair coreference prediction with Lee-style BERT model

Task	Precision	Recall	F
Mention span detection	90.7	87.6	89.1
+ PROP/NOM/PRON	90.2	86.5	88.3
+ Entity class	89.2	85.5	87.3

Table 4: Mention identification performance.

Training source	B^3	MUC	CEAF$_{\phi_4}$	Average
OntoNotes	57.7	81.2	49.7	62.9
PreCo	63.5	84.2	55.1	67.6
LitBank	62.7	84.3	57.3	68.1

Table 5: Coreference resolution performance on predicted mentions.

In-domain annotations matter!
PreCo is 100x larger than LitBank.
Could transfer learning help?