Coreference Resolution

Introduction

What is Coreference Resolution?

- Manning and Clark:

> Coreference Resolution is the task of identifying which mentions in a text refer to the same real world entity. A mention is a text span, usually coherent, which serves as a referent to an entity.

Improving Coreference Resolution by Learning Entity-Level Distributed Representations

Kevin Clark
Computer Science Department
Stanford University
kevclark@cs.stanford.edu

Christopher D. Manning
Computer Science Department
Stanford University
manning@cs.stanford.edu

What is Coreference Resolution?

Beispiel

$[\text{Barack Obama}]_1^1$ nominated $[\text{Hillary Clinton}]_2^2$ as $[[\text{his}]_3^1$ secretary of state$]_4^3$ on $[\text{Monday}]_5^4$. $[\text{He}]_6^1$

- Superscript: ID of an entity
- Subscript: ID of a mention

What is Coreference Resolution?

- What is the structure of the problem?
- Goal is it to predict an ID for every mention (the entity ID)
- Example:

Mention-ID	Entity-ID
1	1
2	1
3	1
4	1

➔ This means, the first 4 mentions are all the same entity

What is Coreference Resolution?

- The amount of entities is unknown, but it resides in $[1, n]$, with n being the amount of mentions in the text

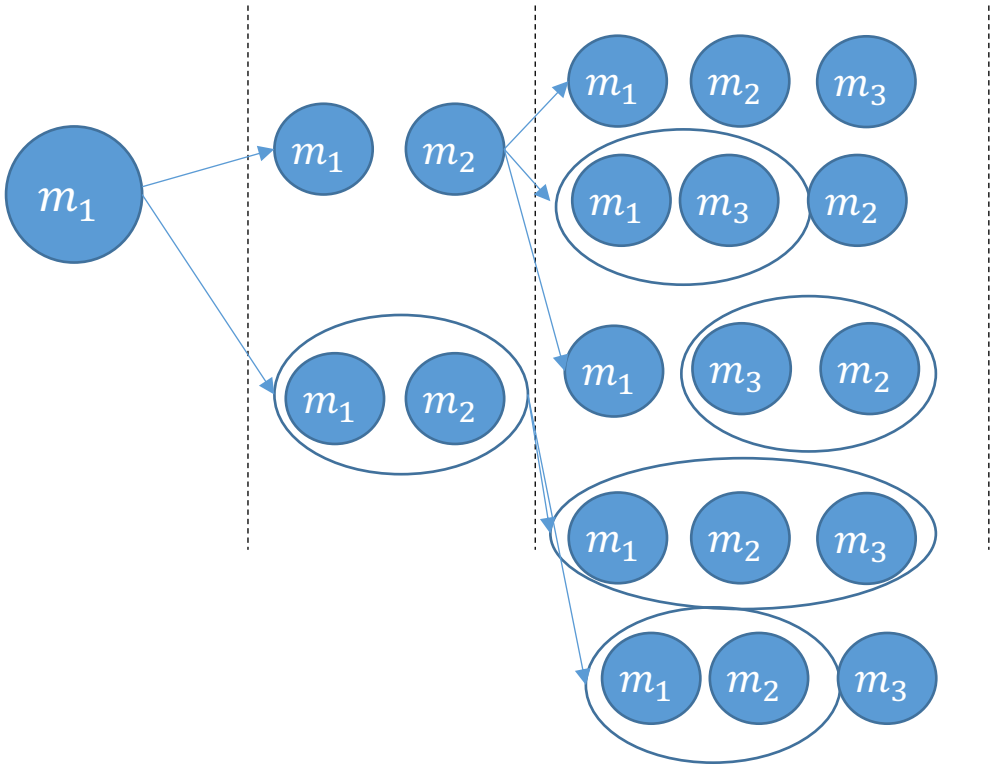
Coreference, more formal

The task of Coreference Resolution is to provide an entity ID for every mention, this results in a **clustering** of mentions

➔ The structure of Coreference Resolution is to predict a clustering among all possible clusterings

Coreference as Clustering

- How many possibilities to cluster n mentions
- Idea: Swipe iteratively through the text and keep track of all possible sub-clusterings



Coreference as Clustering

- We observe:
  - Every clustering with k clusters, creates $k + 1$ clusters in the next step

- You can now try to sum this and you will end up at the **Bell number**

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$$

Coreference as Clustering

- Bell number:

Amount mentions	Bell number
10	115975
20	51724158236496
30	846749014529889671069667
50	1.8572414972124*E*+47
100	$2.3 \cdot 10^{117}$

➔The amount of clusters is growing extremely fast, scoring and ranking them is infeasible!

Coreference Resolution – Literary Phenomena

- If you are dealing with Coreference literature, you should know some literary phenomena

Phenomenon	Example
Apposition	$[\text{Otto}]_1^1, [\text{her oldest son}]_2^1$
Predicative	$[\text{He}]_1^1$ was $[\text{a baker}]_2^?$
Expletive	[It] is raining
Singleton	An entity with just a single reference
Split Antecedents	$[[\text{Hansel}]_1^1$ and $[\text{Gretel}]_2^2]_3^3$. $[\text{They}]_4^3$
Bridging Anaphora	I met [two interesting people] yesterday. [The woman]