



Search Medium

Write



A Quick Review of Coreference Resolution Task



Gal Hever · Follow

5 min read · May 29, 2020



76



1



Introduction

In this blog-post we will go over the Coreference Resolution task. This is one of the tasks in NLP that belongs to the discourse analysis part in which the

research detects how the sentences are combined to one long meaningful text. So, let's start!

What is a Coreference Resolution Task?

A coreference resolution task is a clustering task in NLP that identifies all the nouns in the sentence\document\corpus that refer to the same entity\event. An entity is defined by either a person name, location or organization.

In many natural language applications, such as question answering, automatic document summarization or machine translation, the first step is to preprocess the text for identifying references to entities before start working on the main task.

To understand the coreference task better, let's see an example:

*“Emma said that **she** thinks that **Nelson** really likes to dance, because **he** goes to the dancing studio 4 times a week.”*

Emma and *she* belong to the same entity and *Nelson* and *he* belong to the same entity, while each entity will represent a different cluster.

Nouns or entities that don't have more than one reference in the corpus are called *singletons* (a cluster that contains just one unit).

The coreference task is separated into two sub tasks:

1. Mention Detection

2. Mention Clustering

1. Mention Detection


In this sub-task the main goal is to find all the candidates spans referring to some entities. For example, in the sentence below the mention detection step will color all the candidates spans in blue:

"**Emma** said that **she** thinks that **Nelson** really likes to dance, because **he** goes to the dancing studio 4 times a week."

There are three kinds of mentions that will be detected in this step:

1. Pronouns:

A pronoun is a word that substitutes for noun phrase and usually involves anaphora, where the meaning of the pronoun is dependent on an antecedent. For instance, '*She*' is the pronoun in the sentence "Noa gave an amazing lecture, when she was in the conference at Madrid last year."



| | ENGLISH PRONOUNS | | | | |
|----------------------|------------------|-----------------|-----------------------|---------------------|--------------------|
| | Subject Pronouns | Object Pronouns | Possessive Adjectives | Possessive Pronouns | Reflexive Pronouns |
| 1st person | I | me | my | mine | myself |
| 2nd person | you | you | your | yours | yourself |
| 3rd person (male) | he | him | his | his | himself |
| 3rd person (female) | she | her | her | hers | herself |
| 3rd thing | it | it | its | (not used) | itself |
| 1st person (Plural) | we | us | our | ours | ourselves |
| 2nd person (Plural) | you | you | your | yours | yourselves |
| 3rd person and thing | they | them | their | theirs | themselves |

2. Named-entity recognition (NER):

NER model locates entities in unstructured text and classifies them into pre-defined categories (such as person names, organizations, locations, products, etc).

For example:

"There was nothing about this storm that was as expected," said **Jeff Masters**, a meteorologist and founder of **Weather Underground**. "**Irma** could have been so much worse. If it had traveled 20 miles north of the coast of **Cuba**, you'd have been looking at a (Category) 5 instead of a (Category) 3."

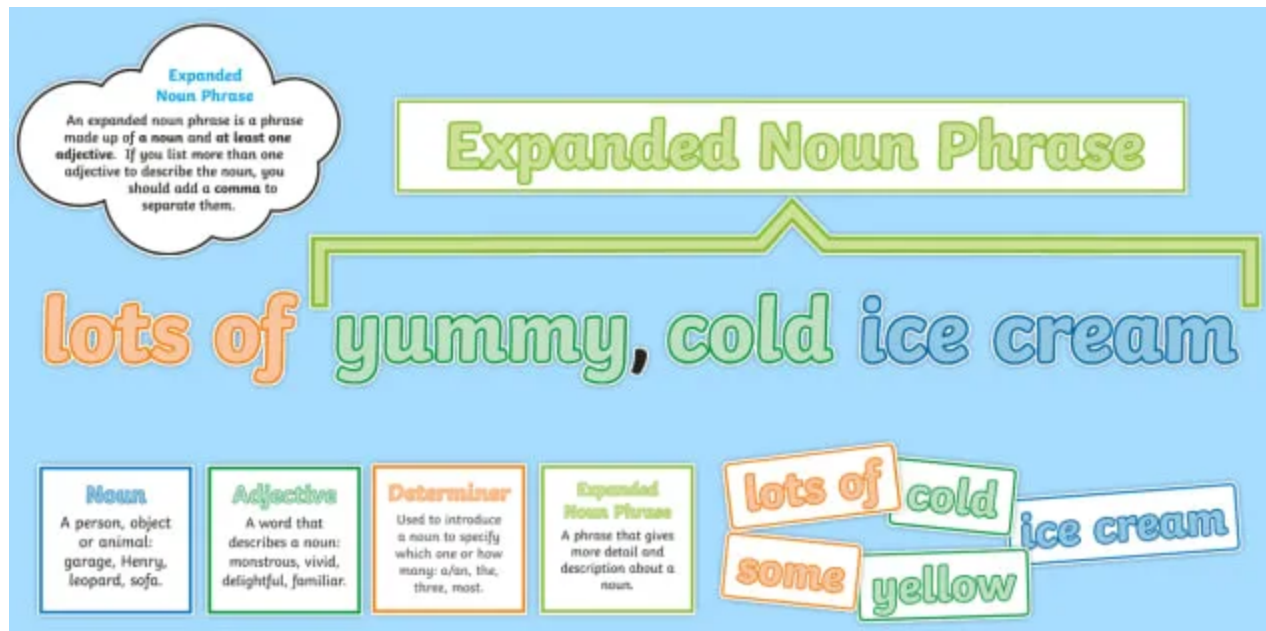
Person

Organization

Location

3. Noun-phrases:

A noun phrase is a bunch words that is headed by a noun and includes modifiers (e.g., 'the,' 'a,' 'of them,' 'with her').



2. Mention Clustering

Once we hold the mentions, the goal of the second sub-task is attempting to identify which ones refer to the same entity. Then, merging the mentions into the cluster corresponding to the entities presented in the text.

"**Emma** said that **she** thinks that **Nelson** really likes to dance, because **he** goes to the dancing studio 4 times a week."

Basic Terms

Let's go over some terms in this domain. We will start with an example that describes them in a simplistic way:

*"Gal came back late from the party, because **she** really enjoyed there."*

Antecedent — An expression in the text that dictates the meaning to all the rest pronouns. For instance, in the example above, the pronoun *she* takes its meaning from *Gal*, so *Gal* is the antecedent of *she*.

Anaphora — Refers to an expression which its interpretation depends upon another expression in context (its antecedent). For example,

*"If you want a **cake**, there is **some** in the kitchen."*

Cataphora — Cataphora is a type of anaphora but the pronoun appears earlier than the noun that it refers to. For example,

*"If you want **some**, there is a **cake** in the kitchen."*

In the sentence above the pronoun *some* appears before the noun *cake*.

Common Features

There are some common features that can help us create the coreference links between mentions just by using the language structure rules, so let's take a look at some examples.

Hand-Crafted Features

Recency:

More recently mentioned entities are likely to be referred. For instance, "Donna went to the dancing class and Ann joined her because she likes to dance." *She* refers to *Ann* because it is recent; although, it can refer to *Donna* but intuitively it refers to the recent one.

Grammatical role:

Subject position entities are more likely to be referred than object position entities. For instance, "Donna went to dancing class and Ann joined her. She likes the dancing lessons." In this case *She* refers to *Donna* because *Ann* is in the object position and *Donna* is the subject position.

Parallelism:

"Donna went to dancing class with Ann, and Maria went with her to watch a movie." *Her* refers to *Ann* while the word *went* create a parallel structure that give us a clue on the antecedent of the pronoun *her*.

Verb semantics:

The semantics in the sentence can also give us a clue of how to link between the mentions in the sentence.

“The animal didn’t cross the street because **it** was too tired.”

“The animal didn’t cross the street because **it** was too wide.”

In the first sentence the word *it* refers to the animal and in the second sentence *it* refers to the street, we know it because of the relation to the word *tired* and *wide*.

Additional Features

There are few more features that can help building a supervised resolution classifier that are called — PNG Constraints:

- Person (1st person, 2nd person, 3rd person)
- Number (single or plural)
- Gender (male or female)

So, how does coreference system works in practice?

The traditional way was to run a kind of pipeline that is composed of separated models, while first we needed to run a part of a speech tagger to find all the pronouns in the text, the second step was running a named-entity recognizer to detect all the entities and then a parser and named mention detector and coreference clustering system which is summarized to

a five-step pipeline. It worked like that until approximately 2016 when the approach changed to be more directed into end-to-end coreference models.

End Notes

This blog-post summarized all the basic terms of coreference task. If you wish to continue reading about the models development of the last few years in this domain you can check out my next blog-post —[Coreference Resolution Models](#), that gives a short review about it.

More from the list: "NLP"

Curated by Himanshu Birla



Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

★ . 11 min read . Sep 4, 2021



Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

★ . 6 min read . Sep 3, 2021



Jon Gi... in

The Word2vec

★ . 15 min rea



[View list](#)



Written by Gal Hever

108 Followers

Data Scientist

Follow

More from Gal Hever



Gal Hever

Getting Started with NVIDIA NeMo ASR

NVIDIA NeMo—Quick Start Guide

3 min read · Apr 20, 2021



89



Gal Hever

Sentiment Analysis with Pytorch—Part 4—LSTM\BiLSTM Model

Introduction

8 min read · Apr 11, 2020



74



2





 Gal Hever

Coreference Resolution Models

A Review of the Latest Models

12 min read · Aug 10, 2020

 17 



 Gal Hever

Sentiment Analysis with Pytorch— Part 1—Data Preprocessing

Introduction

8 min read · Apr 8, 2020

 59 

See all from Gal Hever

Recommended from Medium





Dixn Jakindah

Top P, Temperature and Other Parameters

Large Language Models(LLMs) are essential tools in natural language processing (NLP)...

3 min read · May 18



7



Tomas Vykruta

Understanding Causal LLM's, Masked LLM's, and Seq2Seq: A...

In the world of natural language processing (NLP), choosing the right training approach i...

7 min read · Apr 30



20

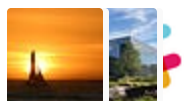


Lists



Staff Picks

465 stories · 317 saves



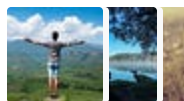
Stories to Help You Level-Up at Work

19 stories · 235 saves



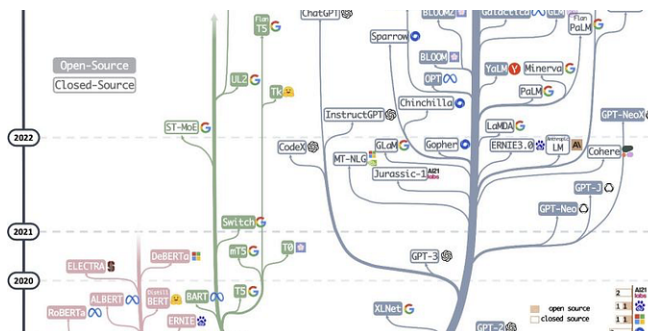
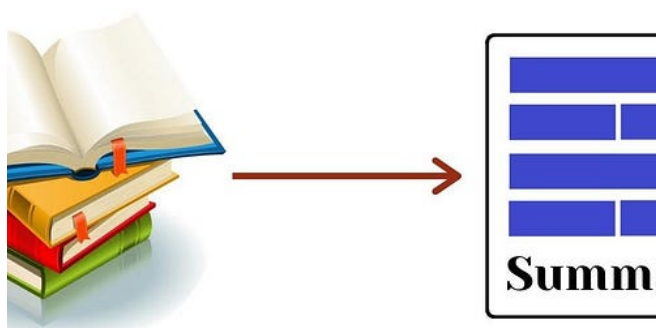
Self-Improvement 101

20 stories · 643 saves



Productivity 101

20 stories · 597 saves



Fabiano Falcão

Metrics for evaluating summarization of texts performe...

Text summarization performed by Transformers is one of the most fascinating...

7 min read · Apr 23

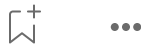


Haifeng Li

A Tutorial on LLM

Generative artificial intelligence (GenAI), especially ChatGPT, captures everyone's...

15 min read · Sep 14

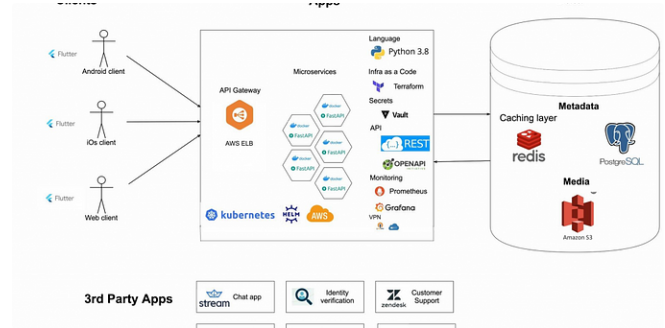


Ritesh

Speaker diarization using Whisper ASR and Pyannote

What is speaker diarization?

8 min read · Jul 22



Dmitry Kruglov in Better Programming

The Architecture of a Modern Startup

Hype wave, pragmatic evidence vs the need to move fast

16 min read · Nov 7, 2022



See more recommendations