Search Medium

Write

Member-only story

Coreference Resolution in Python

Integrate Neural Network-Based Coreference Resolution into your NLP Pipeline using NeuralCoref

Chris Thornton · Following

Published in Towards Data Science · 3 min read · Dec 22, 2019

115 3

In human language, **endophoric awareness** plays a key part in comprehension (decoding) skills, writing (encoding) skills, and general linguistic awareness. Endophora consists of anaphoric, cataphoric, and self-references within a text.

Anaphoric references occur when a word refers back to other ideas in the text for its meaning.

```
David went to the concert. He said it was an amazing experience.

He refers to David.
It refers to the concert.
```

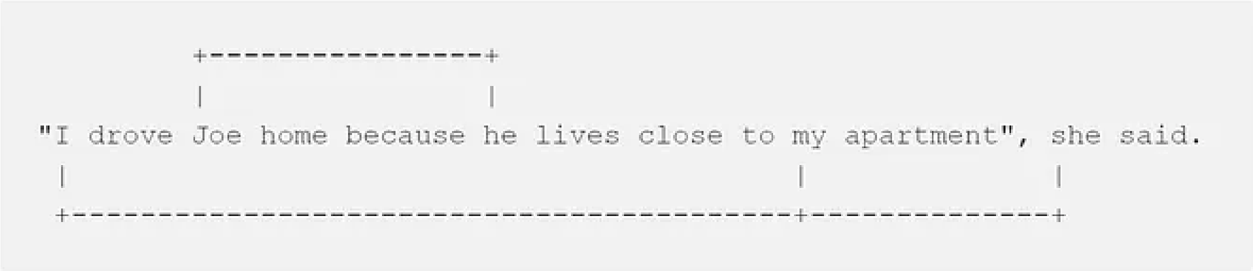
Cataphoric references occur when a word refers to ideas later in the text.

```
Every time I visit her, my grandma bakes me cookies.

Her refers to my grandma.
```

Coreference resolution is the **NLP** (Natural Language Processing) equivalent of endophoric awareness used in information retrieval systems, conversational agents, and virtual assistants like Amazon’s Alexa. It is the task of clustering mentions in text that refer to the same underlying entities.

For example:



```
         +----------------+
         |                |
"I drove Joe home because he lives close to my apartment", she said.
 |                                          |              |
 +------------------------------------------+--------------+
```

“I”, “my”, and *“she”* belong to the same cluster and *“Joe”* and *“he”* belong to the same cluster.

Algorithms which resolve coreferences commonly look for the nearest preceding mention that is compatible with the referring expression. Instead of using rule-based dependency parse trees, neural networks can also be trained which take into account word embeddings and distance between mentions as features.

NeuralCoref is an open source python packgage integrated in SpaCy’s NLP pipeline. You can install NeuralCoref with pip:

```
pip install neuralcoref
```

or from sources with dependencies in a virtual environment:

SpaCy and NeuralCoref can be used to create production-ready NLP applications with little fine-tuning. For example, let's parse through the historical **United States v. Nixon** case to retrieve facts referencing the former U.S. President Richard Nixon:

Output:

Fact count: 108

1. *Following indictment alleging violation of federal statutes by certain staff members of the White House and political supporters of the President, the Special Prosecutor filed a motion under Fed.*

2. *Proc. 17(c) for a subpoena for the production before trial of certain tapes and documents relating to precisely identified conversations and meetings between the President and others.*
3. *the President, claiming executive privilege, filed a motion to quash the subpoena.*

The script scrapes the webpage with Urllib and parses HTML using Beautiful Soup. We load the text into a SpaCy model of our choice; you can download pre-trained SpaCy models from the terminal as shown below:

```
python -m spacy download en_core_web_lg
```

The SpaCy pipeline assigns word vectors, context-specific token vectors, part-of-speech tags, dependency parsing, and named entities. by extending the SpaCy's pipeline of annotations you can resolve coreferences.

You can retrieve a list of all the clusters of corefering mentions using the `doc._.coref_clusters` attribute and replace corefering mentions with the main mentions in each cluster by using the `doc._.coref_resolved` attribute.

SpaCy has a built-in unsupervised sentence tokenizer to split the text into a list of sentences. Use lowercased lemmatized sentences for approximate string searching to the topic of your interest (e.g. President).

Machine Learning | Artificial Intelligence | Python | Programming | Data Science

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 Word2ve

✦ · 15 min rea

View list

Written by Chris Thornton

Following

345 Followers · Writer for Towards Data Science

Sharing ideas and research about ML, NLP, Data Science | Toronto, Canada | https://www.linkedin.com/in/christopher-thornton1

More from Chris Thornton and Towards Data Science

Chris Thornton in Towards Data Science

Fuzzy Name Matching with Machine Learning

Stacking Phonetic Algorithms, String Metrics and Character Embedding for Semantic...

✦ · 4 min read · Jul 30, 2020

186 5

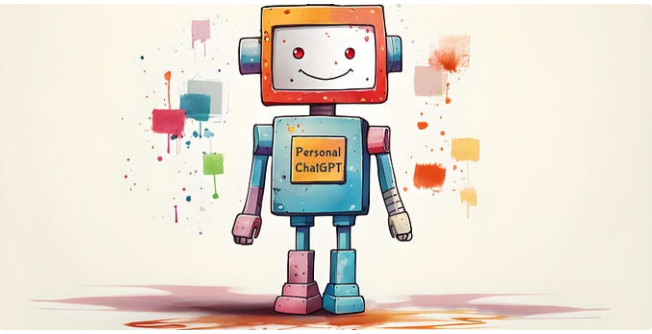
Antonis Makropoulos in Towards Data Science

How to Build a Multi-GPU System for Deep Learning in 2023

This story provides a guide on how to build a multi-GPU system for deep learning and...

10 min read · Sep 17

549 11



Robert A. Gonsalves in Towards Data Science

Your Own Personal ChatGPT

How you can fine-tune OpenAI's GPT-3.5 Turbo model to perform new tasks using you...

✦ · 15 min read · Sep 8

595 7

Chris Thornton in Towards Data Science

Auto-Generated Knowledge Graphs

Utilize an ensemble of web scraping bots, computational linguistics, natural language...

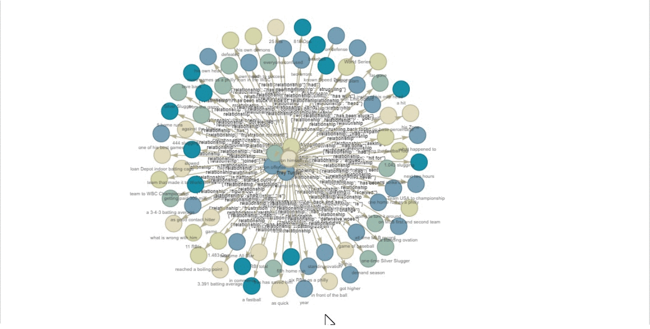
✦ · 4 min read · Feb 9, 2020

1.3K 9

See all from Chris Thornton

See all from Towards Data Science

Recommended from Medium



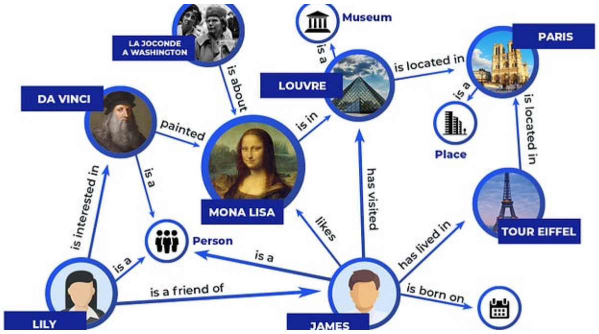
Wenqi Glantz in Better Programming

7 Query Strategies for Navigating Knowledge Graphs With...

Exploring NebulaGraph RAG Pipeline with the Philadelphia Phillies

✦ · 17 min read · 4 days ago

👏 501 💬 4



Alla Chepurova in DeepPavlov

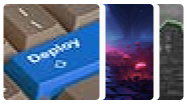
Improving Knowledge Graph Completion with Generative LM...

Combining both internal LM knowledge and external data from KG

13 min read · Sep 5

👏 36 💬

Lists



Predictive Modeling w/ Python

20 stories · 452 saves

Practical Guides to Machine Learning

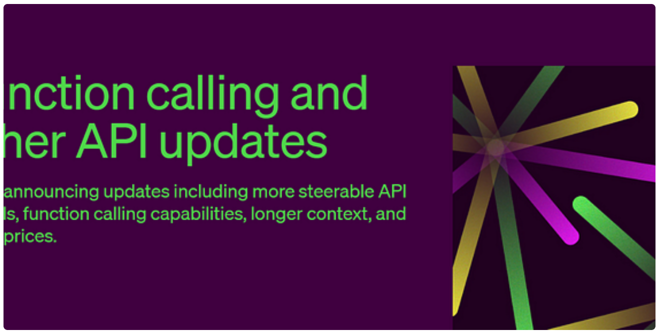
10 stories · 519 saves

Coding & Development

11 stories · 200 saves

ChatGPT

21 stories · 179 saves


nction calling and
her API updates
announcing updates including more steerable API
s, function calling capabilities, longer context, and
prices.

Benjamin De Kraker

Fun with Functions: (Almost) Everything About GPT API...

OpenAI has announced support for "Functions," a new feature within the GPT-3.5...

11 min read · Jun 14

5 1

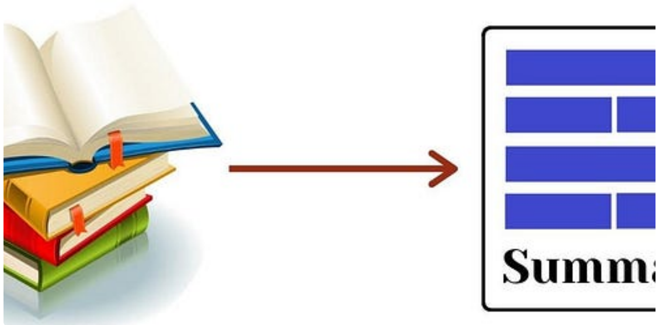
Dixn Jakindah

Top P, Temperature and Other Parameters

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

3 min read · May 18

7



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

Build	Stable (2.0.0)		Preview (Nightly)	
	Linux	Mac		Windows
	Conda	Pip	LibTorch	Source
e	Python		C++ / Java	
e Platform	CUDA 11.7	CUDA 11.8	~~ROCm 5.4.2~~	CPU

Command:

```
pip3 install torch torchvision torchaudio --index-url https://download.
orch.org/whl/cu117
```

Bluetick Consultants

The Future of Automatic Speech Recognition—Whisper

Introduction

4 min read · Jun 2

4

1

See more recommendations