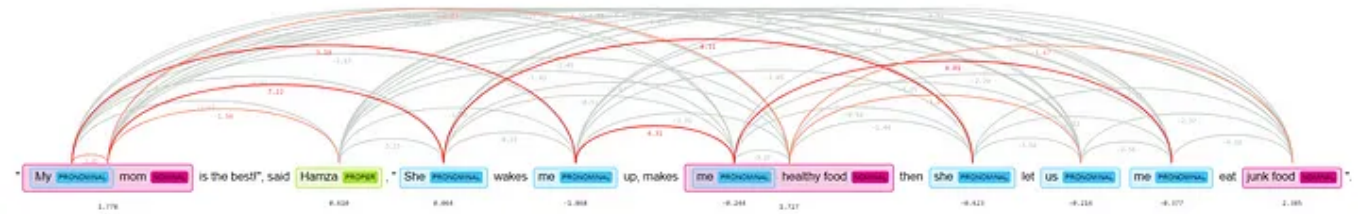
How to Boost Your Topic-Modeling Performance with Coreference Resolution

Improving the accuracy score from 83% to 93% to identify land conflict topics in news articles.

Zaheeda Chauke · Following
Published in Omdena · 4 min read · Nov 26, 2019



Hugging face coreference system in operation with my own example. Try it for yourself!

As a Junior Data Scientist, my Machine Learning journey thus far has led me to NLP challenges which involved good old' fashioned text-classification. So, I was enthused when Omdena presented an opportunity for me to broaden my skill-set and delve into an aspect of NLP I was not familiar with.

After applying to Omdena, I was accepted as a Machine Learning Engineer to collaborate in their AI for Good challenge with the World Resources

Institute.

The challenge

ECOLOGIC

Tell 'em that it's inhuman nature

India has the dubious distinction of more environmental conflicts than any other country

SOUMYA SARKAR
is Managing Editor of
IndiaClimateDialogue.net. @scurve

Development model Protesters in Thoothukudi were met with brutal force by the state. N. RAJESH

The Tamil Nadu government has finally closed down a copper smelter owned by Vedanta Resources in Thoothukudi, a week after police gunned down 13 protesters who had taken to the streets against a proposed expansion of the factory. They said it was poisoning the air they breathe and the water they drink. Had the authorities shown alacrity in listening to them, these lives could have been saved and thousands of others would have been spared serious health problems caused by the fouled air and water in the plant's vicinity.

The shooting on May 22 once again shone a spotlight on India's terrible record of dealing with environmental conflicts, with the administration often directing its firepower on behalf of offending corporations rather than the people. The Thoothukudi shooting was reminiscent of another such killing on January 2, 2006, at Kalinganagar in coastal Odisha, when police opened fire on tribal people protesting the setting up of a steel plant by the Tata Group, and killed 13.

The deaths at Thoothukudi may have downed shutters of the Vedanta factory for now, but the tribal deaths at Kalinganagar failed to stop Tata Steel, which has since commissioned its plant. More than 12 years after the Kalinganagar incident, an inquiry commission found no fault with the authorities. What stands out from the two similar incidents is that, whatever the outcome of civil protests, lives of citizens are cheap and neither the state nor corporations are willing to own up to the cost of harming and degrading the environment.

Cosying interests

India has the dubious distinction of more environmental conflicts than any other country, according to Environmental Justice Atlas, an international database. At last count, the atlas lists 271 cases in India, far ahead of Colombia's 128, the second-worst nation. Even a cursory look at the list of conflicts indicates a cosy relationship between business and political interests, supported by a compliant administration.

Repeat offenders

What else could explain, for instance, the continuing operation of scores of Coca-Cola bottling plants despite widespread civic opposition in places ranging from Uttarakhand and Uttar Pradesh in the north to Kerala in the south?

The global soda conglomerate has been repeatedly accused of abusing local water resources, depleting groundwater levels, and contaminating the soil. Strong protests have shuttered a couple, but scores of others continue, with no indication that the company has taken strong steps to address the environmental damage it causes.

Unfortunately, there are hundreds of such instances of abuse, and only a handful receive any attention, and that too only when the local community protests. For some, the narrative of conflicts is seen as development versus environment. It is patently a false binary because there is enough evidence that economic prosperity does not necessarily mean environmental degradation.

If the happiness and well-being of citizens are compromised what use does society have for such growth?

In fact, a few opinion makers have the gall to claim that these conflicts stifle the ease of doing business in the country. Such a blinkered view ignores the fact that the growth of business and rapid industrialisation is but a means to provide for a better life for the people of the country. If the happiness and well-being of citizens are compromised and their natural environments are poisoned to fill state coffers and line the pockets of big business, what use does society have for such growth?

Daylight murder

One way authorities have tried to address environmental concerns is to frame more stringent laws to conserve natural resources, limit the environmental impacts of industrialisation, and offer some compensation to people who have borne the brunt of so-called development. But it is a bitter truth that lawmaking in India is often an opportunity to open up loopholes that enable authorities and businesses to flout norms, and sometimes, literally get away with murder.

This is evident when we find that India is ranked 177 among 180 nations on the Environmental Performance Index 2018, dropping precipitously from 141 in 2016, according to the biennial report by Yale and Columbia universities along with the World Economic Forum.

It is at the bottom of the heap on environmental health and third-worst in the world on air quality. Researchers have consistently found that environmental governance in India is poor not because of laws but their pathetic implementation, particularly because the framework does not allow for meaningful public participation, resulting in imprudent resource management.

In a scenario when environmental governance has clearly failed, the judiciary has picked up the slack. The courts have often batted for environmental causes and have been more sympathetic of community concerns than the administration. But they have a tough job ahead. There were 21,145 environmental cases pending for trial in 2016, according to an analysis in the State of India's Environment 2018, which will be released by the Centre for Science and Environment on World Environment Day on June 5.

> Identifying environmental conflict events in India using news media articles.

Part of this project was to scrape news media articles to identify environmental conflict events such as resource conflicts, land appropriation, human-wildlife conflict, and supply chain issues.

With an initial focus on India, we also connected conflict events to their jurisdictional policies to identify how to resolve those conflicts faster or to identify a gap in legislation.

Part of the pipeline in building this Language Model was a semi-supervised Topic Modeling task whose process and the outcome is detailed below.

How-To Identify Land Conflicts in India Through NLP Semi-Supervised Topic Modeling

Semi-supervised learning to identify topics in articles of land conflicts with a model accuracy of 93 percent.

medium.com

In short, in order to make this Topic Modelling model robust, **Coreference Resolution** was suggested as one of the possible additions.

I took the initiative to work on this task and was later elected as the Task Manager.

The team consisted of 27 other collaborators ranging from data wranglers to data engineers, and machine learning engineers. **Together we were ready to contribute!**

Where to begin? Research.

Since I had no experience with Coreference Resolution, I knew my best starting point, as it is with most projects, would be researching this topic.

What exactly is Coreference Resolution?

> ***Coreference resolution*** *is the task of finding all expressions that refer to the same entity in a text (1)*

I like to explain using practical examples, a real example featuring my son :)

> "My mom is the best!", said Hamza, "She wakes me up, makes me healthy food then she lets me eat junk food".

- Entities: "**Hamza**", "**my mom**".
- Expressions which refer to "**Hamza**": "**My**", "**me**", "**me**", "**me**".
- Expressions which refer to "**my mom**": "**She**", "**she**".

Pretty simple!

Use Cases

1. In the context of this project, Coreference Resolution could be best used in order to enrich downstream topic modeling by replacing references with the same entity in order to better model the actual meaning of the text. This increases the Tf-Idf of generalized entities and it removes ambiguous words that are meaningless for classification.
2. Another use-case would be to use the Coreferenced text data as additional features, along with Named Entity Recognition tags, in any classification approach. A one-hot-encoded version of unique entities can

be used as input to factorization machines or other approaches for spare modeling.

Which packages are available to implement it?

I explored almost every available python package out there.

I toyed around with some packages which seemed good in theory but were rather challenging to apply to our specific task. We needed a package that would be user-friendly, as a script would have to be developed for 28 people to take and be able to apply without much struggle.

NeuralCoref, Stanford NLP, Apache Open NLP, and Allennlp. After trying out each package, I personally preferred Allennlp, but as a team, we decided to use NeuralCoref with a short but effective script written by one of the collaborators Srijha Kalyan (add a Github link to the code?).

When applied, the package identifies the entities in the given text then they produce “clusters” or “chains”. These clusters consist of the entity (“**Hamza**”), the references linked to that entity (“**My**”, “**me**”) and also their index (position) in the text.

The code was applied to the article data which was annotated by fellow collaborators from the Annotation Task Group. This resulted in a CSV file with the original article titles, the original article text and a new column of Coreferenced article text; not as chains but in the same written format as the original article text.

	class	ids	month	text	text_coref	title
0	positive	6477	1	When Skill-India mission was launched, Hon’ble...	When Skill-India mission was launched, Hon’ble...	“Developing skills of ITI Pass-outs in Mining ...
0	negative	7638	1	mumbai\nUpdated: Feb 01, 2018 10:38 IST\nA ses...	mumbai\nUpdated: Feb 01, 2018 10:38 IST\nA ses...	Mumbai man’s dying declaration sends his kille...
0	negative	1670	1	(Eds: Incorporating CM announcing financial as...	(Eds: Incorporating CM announcing financial as...	11 killed after truck hits three-wheeler, crus...
0	negative	4093	1	Caste conflicts in India; Dalits threaten prot...	Caste conflicts in India; Dalits threaten prot...	Caste conflicts in India; Dalits threaten prot...
0	positive	6706	1	Karnataka is ready for talks with Goa on the M...	Karnataka is ready for talks with Goa on the M...	Karnataka ready for talks with Goa on Mahadayi...

> *The output was then sent to the Topic Modeling Task Team which at that point was sitting on an accuracy of 83%, with the Coreference Resolution data, the accuracy jumped to 93%!*

That's an 11% improvement! All the hard work and hours we put into this learning this task was clearly worth it!

I'm proud to say I have a new skill added to my Data Science ninja Resume!

Want to become an Omdena Collaborator and join one of our tough AI for Good challenges, apply here.

Machine Learning | Naturallanguageprocessing | Artificial Intelligence | AI | Technology

More from the list: "NLP"

Curated by Himanshu Birla

Jon Gi... in Towards Data ...
Characteristics of Word Embeddings

Jon Gi... in Towards Data ...
The Word2vec Hyperparameters

Jon Gi... in
The Word2ve

· 11 min read · Sep 4, 2021

· 6 min read · Sep 3, 2021

· 15 min rea

View list

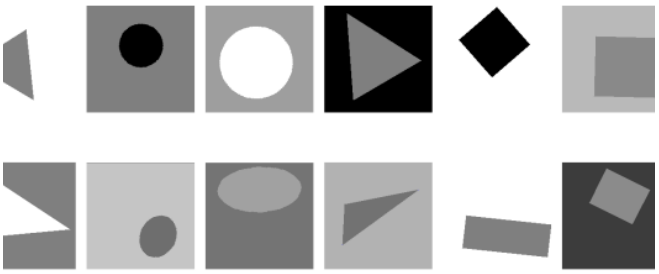
Written by Zaheeda Chauke

Following

115 Followers · Writer for Omdena

Mom | Data Scientist | Machine Learning and Deep Learning Enthusiast | Extrovert

More from Zaheeda Chauke and Omdena



Zaheeda Chauke in Better Programming

Build a Computer Vision Web App — Flask, OpenCV, and MongoDB

Being equipped with some software engineering skills has become a highly value...

12 min read · Sep 2, 2022

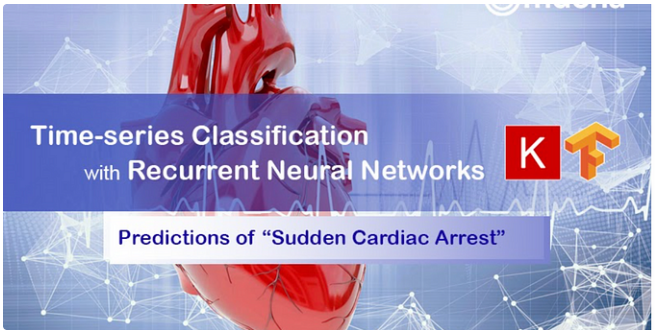
Ahmed ABDELMAGUID in Omdena

Curriculum Learning in Deep

A Technique You May Not Know.

9 min read · Feb 11, 2022

459 3



Sanjana Tule in Omdena

Time-Series Classification Tutorial: Combining Static and Sequential...

Combining Static and Sequential Feature Modeling using Recurrent Neural Networks

7 min read · Dec 22, 2021

113 1

32

Zaheeda Chauke in The Startup

How to set up a Data Science Project

More often than not, those of us who are new to the Data Science field wish to implement...

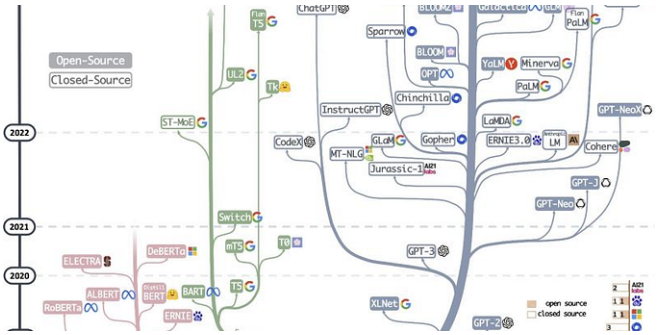
6 min read · Jun 30, 2020

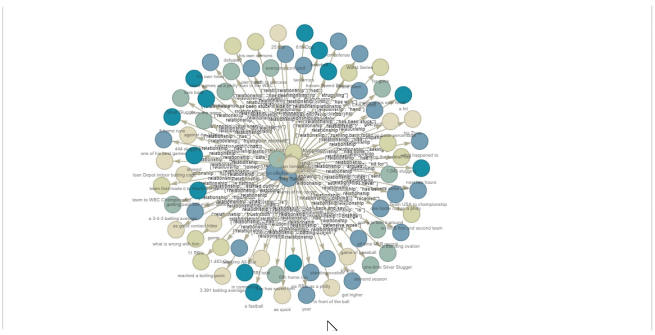
180 5

See all from Zaheeda Chauke

See all from Omdena

Recommended from Medium





Haifeng Li

A Tutorial on LLM

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

15 min read · Sep 14

372

Wenqi Glantz in Better Programming

7 Query Strategies for Navigating Knowledge Graphs With...

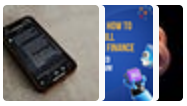
Exploring NebulaGraph RAG Pipeline with the Philadelphia Phillies

· 17 min read · 4 days ago

501 4

Lists

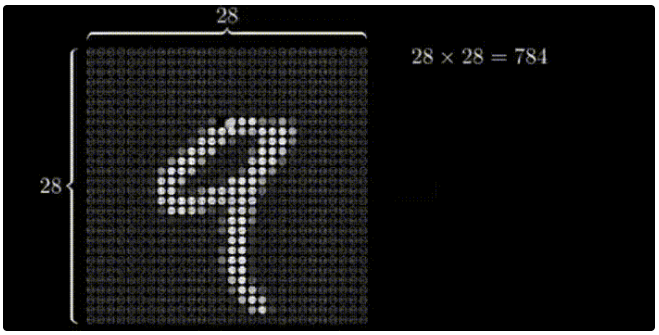
AI Regulation
6 stories · 138 saves



ChatGPT prompts
24 stories · 459 saves

Generative AI Recommended Reading
52 stories · 274 saves

The New Chatbots: ChatGPT, Bard, and Beyond
13 stories · 133 saves



Maximilian Vogel in MLearning.ai

The ChatGPT list of lists: A collection of 3000+ prompts,...

Updated Sep-09, 2023. Added new prompt engineering courses, masterclasses and...

10 min read · Feb 8

8.7K 114

Sadaf Saleem

Neural Networks in 10mins. Simply Explained!

What are Neural Networks?

9 min read · May 15

252 2

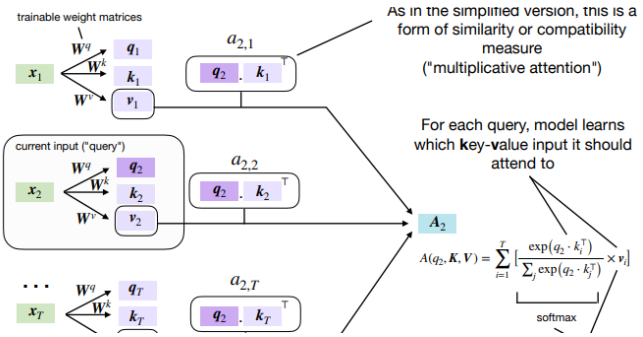
Jesús Cantú

Advancing Speech Recognition with Transfer Learning Techniques

In our previous blog posts, we explored the fundamentals of speech recognition and the...

5 min read · Jun 3

52



Zain ul Abideen

Attention Is All You Need: The Core Idea of the Transformer

An overview of the Transformer model and its key components.

6 min read · Jun 26

144

See more recommendations