Intro to coreference resolution in NLP

How coreference resolution leads to a substantial information gain, with no extra context needed.

Paweł Mielniczuk · Following

Published in Towards Data Science · 8 min read · Jan 7, 2021

Written by Paweł Mielniczuk and Marta Maślankowska.

Photo by Richa Sharma

Introduction

Natural language processing (NLP) refers to the communication between humans and machines. NLP is one of the most challenging branches of Artificial Intelligence mainly because our human language is full of exceptions and ambiguities which are hard for computers to learn. One way of making it easier for them is to get rid of any imprecise expressions that need a context to be clearly understood. A good example is pronouns (e.g. it, he, her) which can be replaced with specific nouns they are referring to.

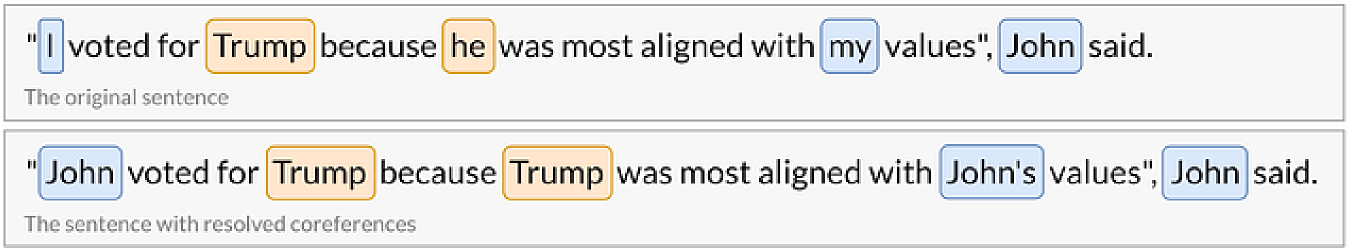
But how about a real-world application?

While working on a Question Answering System for the LMS platform we've encountered several problems. Especially with sentence embeddings —

vector representations of text. It happens that sometimes a sentence consists of many pronouns. Such embeddings often don't reflect the original sentence correctly when sufficient context isn't provided. In order to obtain richer embeddings, we've applied coreference resolution to our pipeline.

What is coreference resolution?

Coreference resolution (CR) is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity. After finding and grouping these mentions we can resolve them by replacing, as stated above, pronouns with noun phrases.



"I voted for Trump because he was most aligned with my values", John said.

The original sentence

"John voted for Trump because Trump was most aligned with John's values", John said.

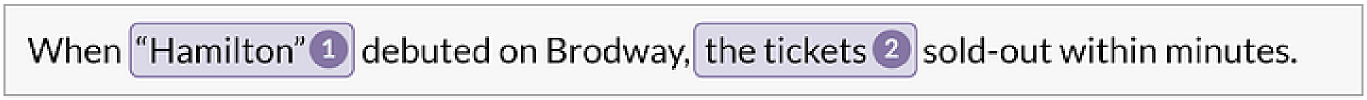
The sentence with resolved coreferences

Coreference resolution is an exceptionally versatile tool and can be applied to a variety of NLP tasks such as text understanding, information extraction, machine translation, sentiment analysis, or document summarization. It is a great way to obtain unambiguous sentences which can be much more easily understood by computers.

Coreference vs. anaphora resolution

It should be noted that we refer to coreference resolution as to a general problem of finding and resolving references in the text. However, technically there are several kinds of references and their definitions are a matter of dispute.

The one case most distinguished from coreference resolution (CR) is anaphora resolution (AR). The relation of anaphora occurs in a text when one term refers to another determining the second's one interpretation [3]. In the example below, we see that (1) and (2) directly refer to different real-world entities however they are used in the same context and our interpretation of (2) relies on (1). These mentions do not co-refer but are in the relation of anaphora.



When “Hamilton” 1 debuted on Brodway, the tickets 2 sold-out within minutes.

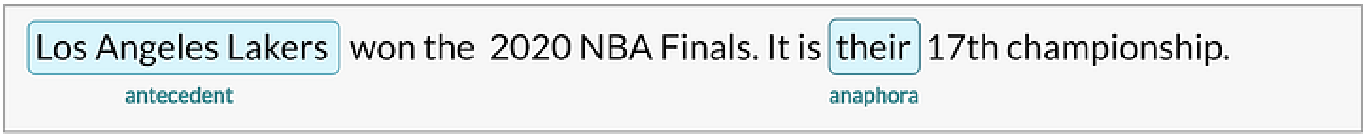
Even though anaphora resolution is distinct from coreference resolution, in the vast majority of cases one equals the other. There are many more examples of such differences and various other kinds of references. However, CR has the broadest scope and covers the vast majority of cases. As we would like to simplify this topic, from now on we are going to assume that all types of relations between terms are coreferential.

Different types of references

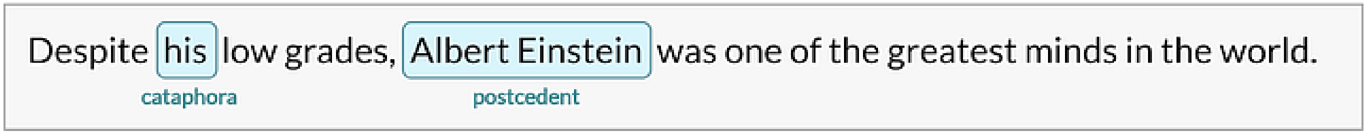
Even if we assume that we can treat all kinds of references as a coreference, there are still many different forms of relations between terms that are worth noting. That's because every kind can be treated differently and most classic natural language processing algorithms are designed to target only specific types of references. [1]

Anaphora and cataphora

These are the bread and butter of our topic. The main difference is that anaphora occurs in the sentence after the word it refers to and cataphora is found before it. The word occurring before an anaphora is called an antecedent and the one following a cataphora is a postcedent.



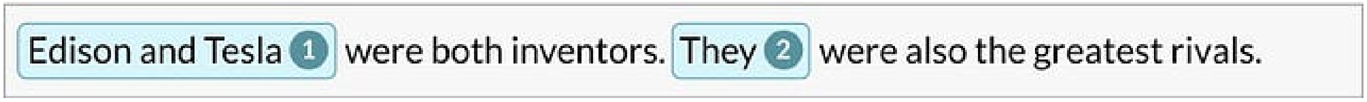
Los Angeles Lakers won the 2020 NBA Finals. It is their 17th championship.



Despite his low grades, Albert Einstein was one of the greatest minds in the world.

Split antecedents

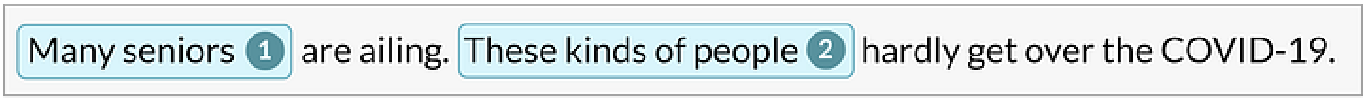
It’s an anaphoric expression where the pronoun (2) refers to more than one antecedent (1).



Edison and Tesla 1 were both inventors. They 2 were also the greatest rivals.

Coreferring noun phrases

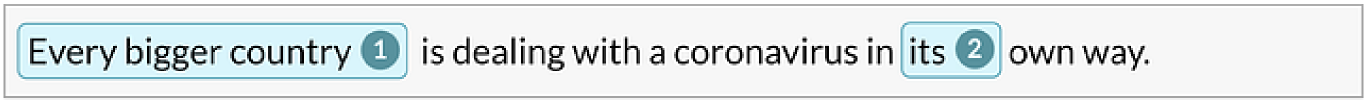
It’s also an anaphoric example of a situation in which the second noun phrase (2) is a reference to an earlier descriptive form of an expression (1).



Many seniors 1 are ailing. These kinds of people 2 hardly get over the COVID-19.

Presuppositions / bound variable

Some argue whether presupposition can be classified as a coreference (or any other “reference”) resolution type. That’s because a pronoun (2) is not exactly referential — in the sense that we can’t replace it with the quantified expression (1). However, after all the pronoun is a variable that is bound by its antecedent [3].



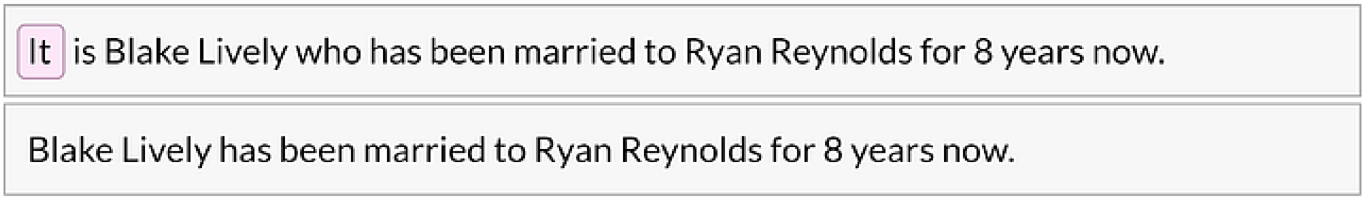
Every bigger country 1 is dealing with a coronavirus in its 2 own way.

Misleading pronominal references

There are also certain situations that can be misleading. It's when there is no relationship between a pronoun and other words in the text and yet the pronoun is there. While creating a CR algorithm we need to pay special attention to those kinds of references so it's good to know in what situations we come into contact with them.

Clefts

A cleft sentence is considered to be a complex expression which has a simpler, less deceptive substitution. It's a case where the pronoun "it" is redundant and we can easily come up with a sentence that has the same meaning but doesn't use the pronoun.

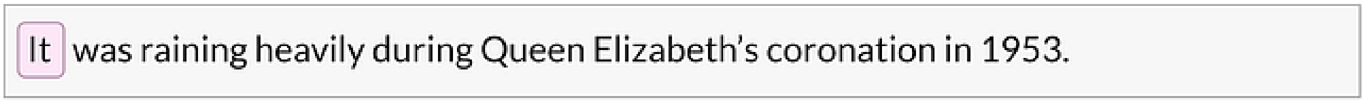


It is Blake Lively who has been married to Ryan Reynolds for 8 years now.

Blake Lively has been married to Ryan Reynolds for 8 years now.

Pleonastic "it"

This type of reference is very common in English so it requires an emphasis. The pronoun "it" doesn't refer to any other term but it is needed in the sentence in order to make up a grammatical expression.



It was raining heavily during Queen Elizabeth's coronation in 1953.

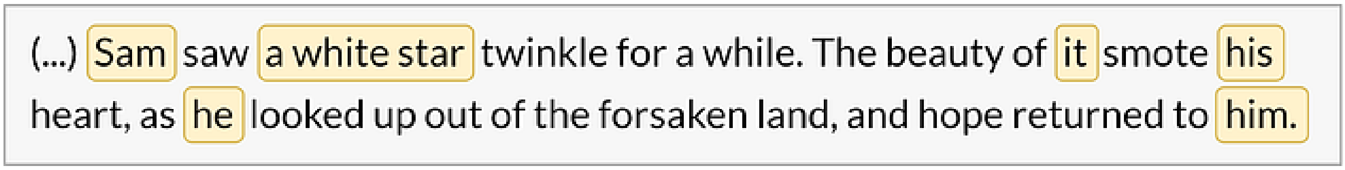
Steps for coreference resolution by example

It's always best to visualize an idea and provide a concrete example as opposed to just theorizing about a topic. What's more, we'll try to explain and give concrete examples of the most common terms, associated with coreference resolution that we may come across in articles and papers.

The first step in order to apply coreference resolution is to decide whether we would like to work with single words/tokens or spans.

But what exactly is a span? It's most often the case that what we want to swap or what we are swapping for is not a single word but multiple adjacent tokens. Therefore span is a whole expression. Another name for it you may come across is a mention. They are often used interchangeably.

In most state of the art solutions, only spans are taken into consideration. It is so since spans carry more information within them, while single tokens may not convey any specific details on their own.



(...) Sam saw a white star twinkle for a while. The beauty of it smote his heart, as he looked up out of the forsaken land, and hope returned to him.

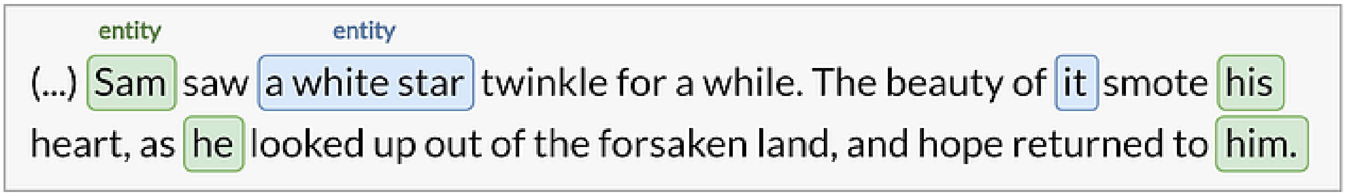
Step 1 — identify potential spans

The next step is to somehow combine the spans into groups.

As we can see in this great quote from J.R.R. Tolkien, there are several potential spans that could be grouped together. Here we have spans like "Sam" or "his" that have only a single token in them, but we also see the span "a white star" consisting of three consecutive words.

Combining items is referred to as clustering or grouping. It is, as its name suggests, a method of taking arbitrary objects and grouping them together into clusters/groups within which these items share a common theme. These can range from words in NLP, through movie categories on Netflix, to grouping food based on their nutritional values.

There are many ways one may group, but what's important is things in the same group should possess similar properties and be as different as possible from other clusters.


entity
entity

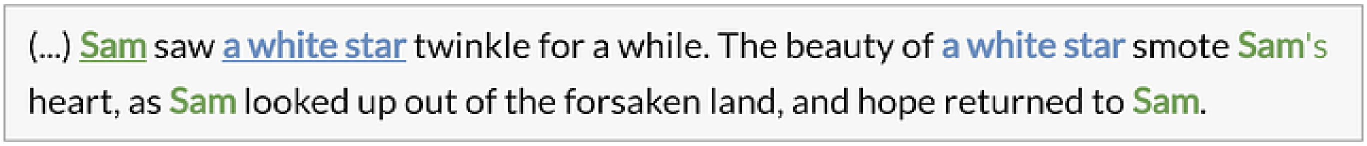
(...) Sam saw a white star twinkle for a while. The beauty of it smote his heart, as he looked up out of the forsaken land, and hope returned to him.

Step 2 — group spans

Here the "property" we are looking for is the spans referring to the same real-world entity.

The resulting groups are [Sam, his, he, him] as well as [a white star, it]. Notice that "Sam" and "a white star" are marked as entities. This is a crucial step in coreference resolution. We need to not only identify similar spans but also determine which one of them is, often referred to as, the real-world entity.

There is no single definition of a real-world entity but we will simply define it as an arbitrary object that doesn't need any extra context to clarify what it is, in our example: "Sam", or "a white star". On the other hand, "his" or "him" are not real-world entities, since they must be accompanied by additional background information.



(...) Sam saw a white star twinkle for a while. The beauty of a white star smote Sam's heart, as Sam looked up out of the forsaken land, and hope returned to Sam.

Step 3 — replace pronouns with real-world entities

As we can see [his, he, him] and [it] have been replaced with the real-world entities, from the corresponding groups — “Sam” and “a white star” respectively. As a result, we obtained a text without any pronouns while still being valid grammatically and semantically.

Summary

The aim of Coreference Resolution is to find, group and then substitute any ambiguous expressions with real-world entities they are referring to.

We’ve discussed a difference between coreference and anaphora resolution as well as shown and explained a couple of common problems associated with them. We’ve also managed to walk through the typical process of CR using an example.

By doing so, sentences become self-contained and no additional context is needed for the computer to understand their meaning. It won’t always be the case where we have well-defined entities but more often than not coreference resolution will lead to information gain.

This is only the first article in the series concerning coreference resolution and natural language processing. In the next one, we will show the pros and cons of the biggest deep learning solutions that we’ve tested ourselves and finally decided to implement in our system.

For more articles like this take a look at NeuroSYS Blog.

Part 2 - Most popular coreference resolution frameworks

References

[1]: Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu (July 2020) *Anaphora and coreference resolution: A review*

[2]: Sharid Loaiciga, Liane Guillou, Christian Hardmeier (September 2017) *What is it? Disambiguating the different readings of the pronoun 'it'*

[3]: Stanford lecture (CS224n) by Christopher Manning (2019)

Data Science NLP Deep Learning Machine Learning Artificial Intelligence

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 Paweł Mielniczuk

18 Followers · Writer for Towards Data Science

Following

More from Paweł Mielniczuk and Towards Data Science

Paweł Mielniczuk in Towards Data Science

Elasticsearch — introduction to key concepts

5 essential steps to start working with Elasticsearch for NLP

16 min read · Sep 16, 2022

151

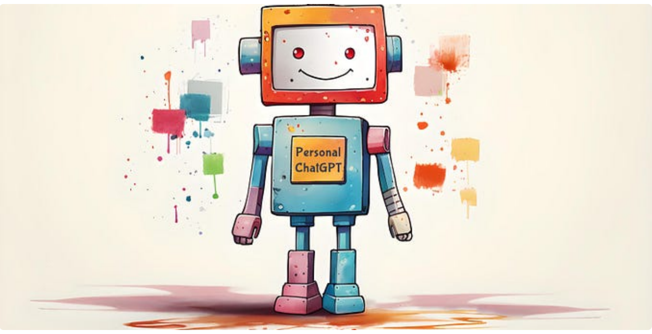
Antonis Makropoulos in Towards Data Science

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



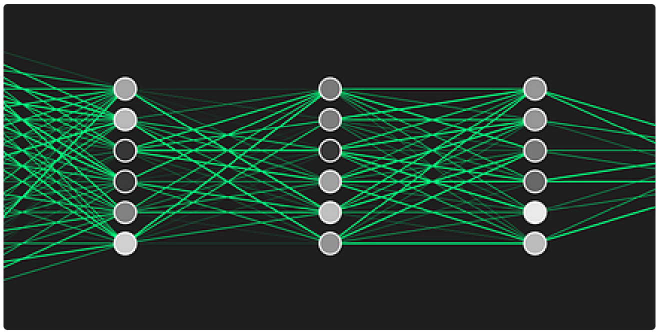
Robert A. Gonsalves in Towards Data Science

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



Callum Bruce in Towards Data Science

How to Program a Neural Network

A step-by-step guide to implementing a neural network from scratch

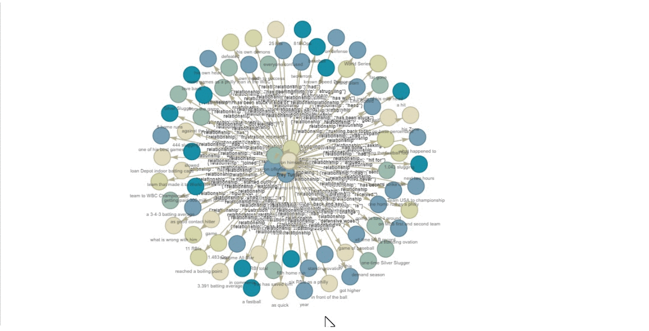
· 14 min read · Sep 24

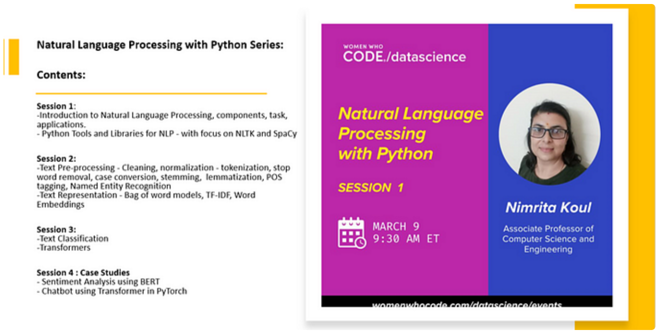
470 1

See all from Paweł Mielniczuk

See all from Towards Data Science

Recommended from Medium





Natural Language Processing with Python Series:

Contents:

Session 1:
-Introduction to Natural Language Processing, components, task, applications.
- Python Tools and Libraries for NLP - with focus on NLTK and SpaCy

Session 2:
-Text Pre-processing - Cleaning, normalization - tokenization, stop word removal, case conversion, stemming, lemmatization, POS tagging, Named Entity Recognition
-Text Representation - Bag of word models, TF-IDF, Word Embeddings

Session 3:
-Text Classification
-Transformers

Session 4 : Case Studies
- Sentiment Analysis using BERT
- Chatbot using Transformer in PyTorch

WOMEN WHO CODE./datascience

Natural Language Processing with Python

SESSION 1

MARCH 9
9:30 AM ET

Nimrita Koul

Associate Professor of Computer Science and Engineering

womenwhocode.com/datascience/events

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

Nimrita Koul

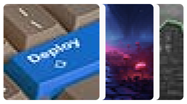
NLP with Python Part 2 NLTK

This is the second article in the series of articles on Natural Language Processing...

5 min read · Apr 5

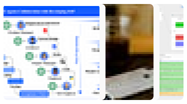
2

Lists



Predictive Modeling w/ Python

20 stories · 452 saves

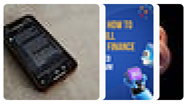


Natural Language Processing

669 stories · 283 saves

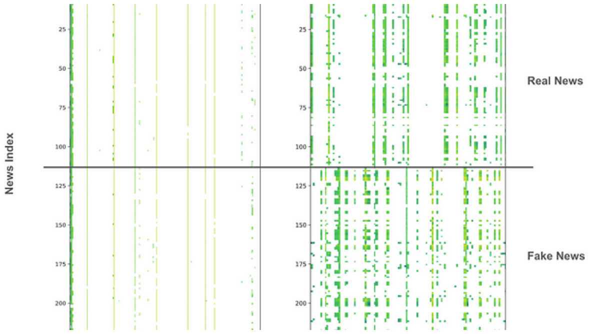
Practical Guides to Machine Learning

10 stories · 519 saves



ChatGPT prompts

24 stories · 459 saves

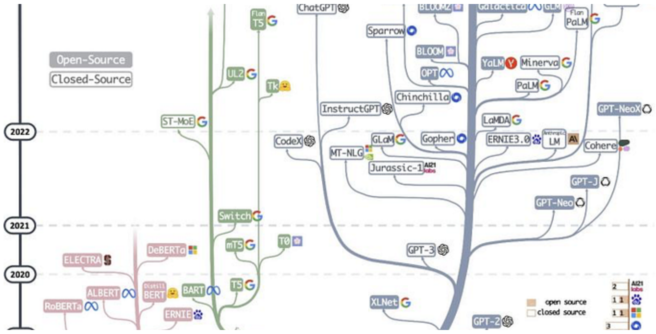


Sherry Wu in Stanford CS224W GraphML Tutorials

Spread No More: Twitter Fake News Detection with GNN

By Li Tian, Sherry Wu, Yifei Zheng as part of the Stanford CS224W course project.

11 min read · May 16



Haifeng Li

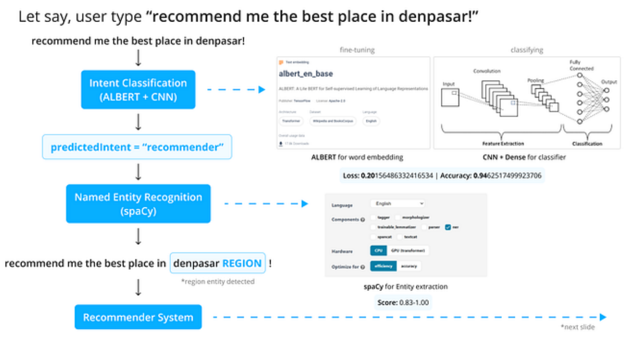
A Tutorial on LLM

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

15 min read · Sep 14

1

372



Let say, user type “recommend me the best place in denpasar!”

recommend me the best place in denpasar!

Intent Classification (ALBERT + CNN)

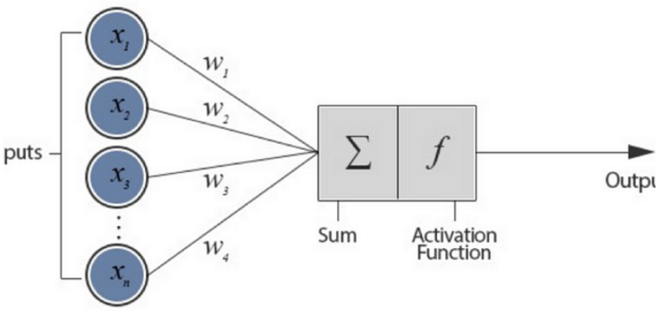
predictedIntent = “recommender”

Named Entity Recognition (spaCy)

recommend me the best place in denpasar REGION !

*region entity detected

Recommender System



Muhamad Luthfi Rey

Building a Smart Chatbot with Intent Classification and Named…

Indonesia attracts travelers worldwide with its vibrant culture, breathtaking landscapes, an…

7 min read · Jul 14

22

Francesco Franco

Introduction to Neural Networks

With Python implementation

12 min read · Sep 14

48

See more recommendations