Part 8: Step by Step Guide to Master NLP – Useful Natural Language Processing Tasks

CHIRAG GOYAL — June 22, 2021

Beginner Data Science NLP Python Text Word Embeddings

This article was published as a part of the Data Science Blogathon

Introduction

This article is part of an ongoing blog series on Natural Language Processing (NLP). Up to part-7 of this series, we completed the most useful concepts in NLP. While going away in this series, let's first discuss some of the useful tasks of NLP so that you have much clarity about what you can do by learning the NLP. After this part, we will start our discussion on Syntactic and Semantic Analysis in detail including the concept of Grammar and Parsing, etc.

So, In this part of this blog series, we will discuss some of the very useful tasks of Natural Language Processing in a detailed manner.

This is part-8 of the blog series on the Step by Step Guide to Natural Language Processing.

Table of Contents

Text Classification

Text classification is one of the classical problems of NLP. This includes several examples from which some of them are mentioned below:

- Email Spam Identification,
- Topic classification of news,
- Sentiment classification

In simple words, text classification is defined as a technique to systematically classify a text object (document or sentence) in one of the fixed categories. This application becomes really helpful when we work with too large data for the purpose of organizing, information filtering, and storage of data.

Typically, a natural language classifier consists of the following two parts:

- Training
- Prediction

Firstly the text input is processed and from preprocessed text, we create the features. Then, we give these features to our machine learning models and after learning from these features, we used that model for the prediction of the new text.

But while making such applications, one has to keep in mind that the text classification models are heavily dependent on the quality and quantity of features, so while applying any machine learning model it is always a good practice to use more and more data to train it.

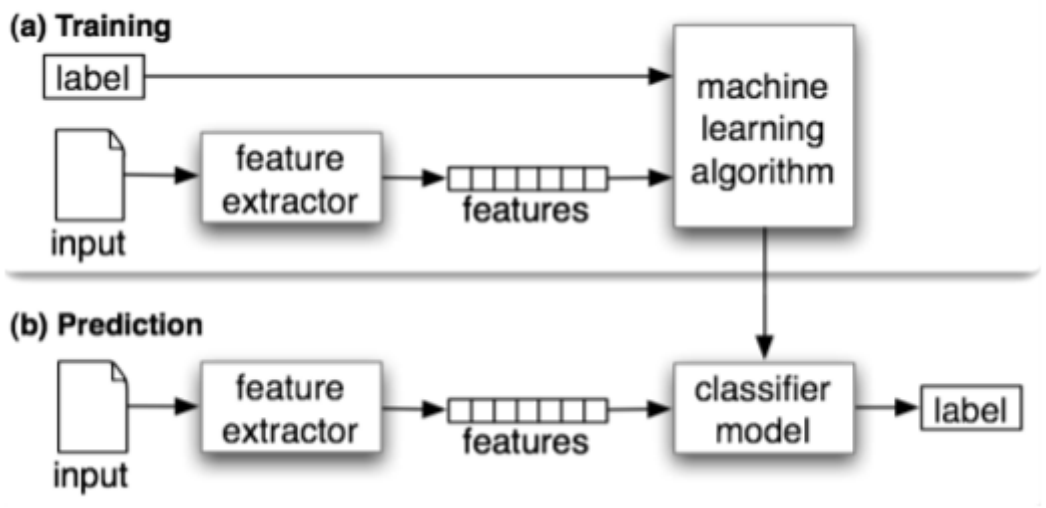


Image Source: Google Images

Sentiment Analysis

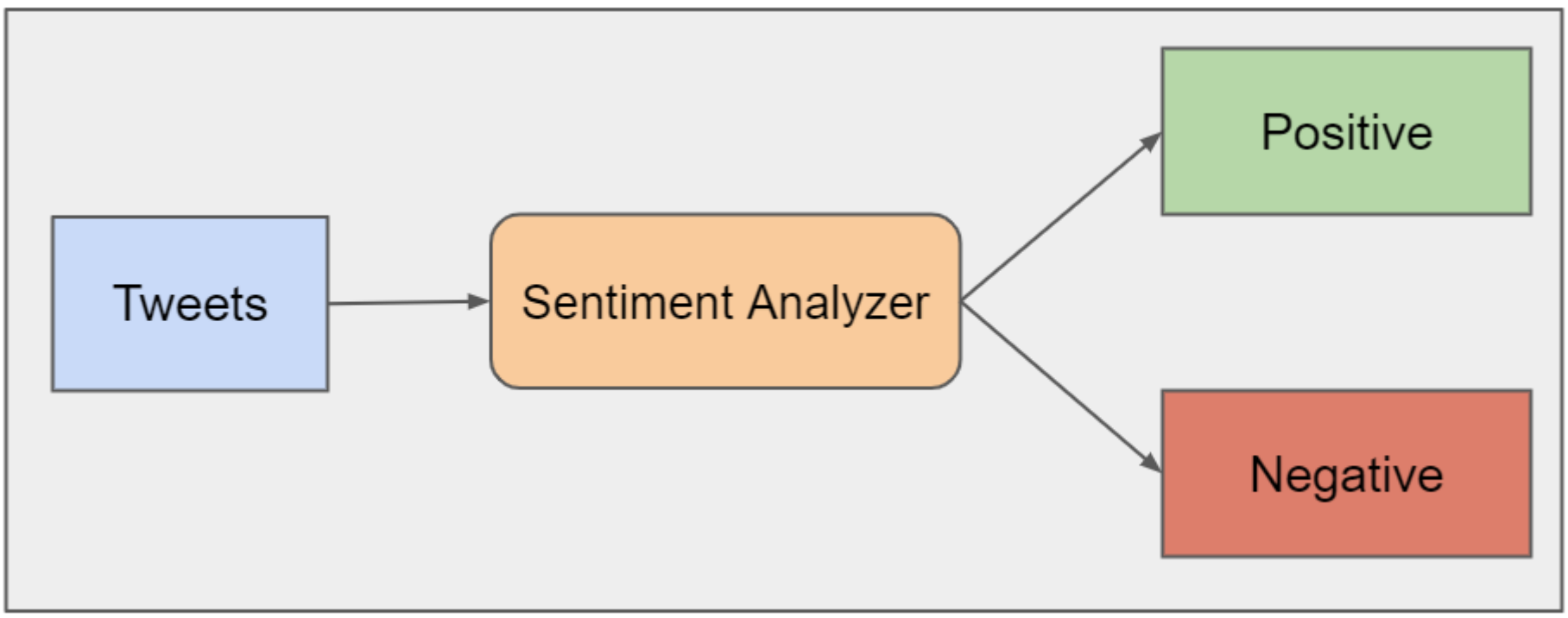


Image Source: Google Images

Sentiment Analysis is another important application of NLP. As the name suggests, sentiment analysis is used to identify the sentiments among several documents. This analysis also helps us to identify the sentiment where the emotions are not expressed explicitly.

Mostly Product based Companies like **Amazon** are using sentiment analysis to identify the opinion and sentiment of their customers online. It will help these big product-based companies to understand the thinking of customers about the products

So, with the help of sentiment analysis companies can judge their overall reputation from customer posts. In this manner, we can say that beyond determining simple polarity, sentiment analysis understands sentiments in context to help us better understand what is behind the expressed opinion.

Fighting Spam

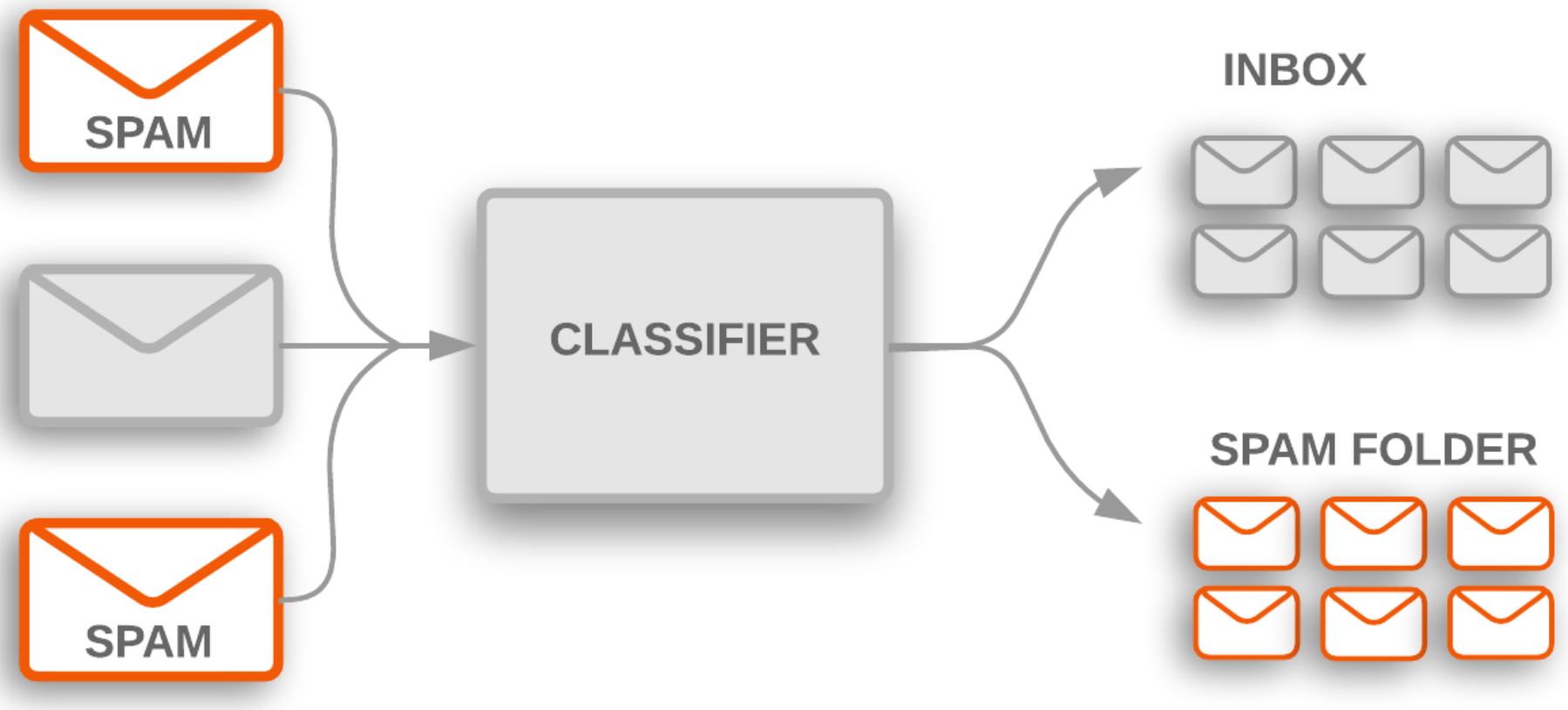


Image Source: Google Images

In today's digital era, one of the most common problems is unwanted emails. This makes Spam filters more important, as it is the first line of defense against this problem.

A spam filtering system can be developed by using NLP functionality by considering the major false-positive and false-negative issues.

Existing NLP models for spam filtering

Some of the existing NLP models used for spam filtering are as follows:

N-gram Modeling

An N-Gram model is defined as an N-character slice of a longer string. In this model, we used several N-grams of different lengths simultaneously in processing and detecting spam emails.

To know more about N-Gram, refer to our previous articles.

Word Stemming

Usually, in Spammers and generators of spam emails, there is a change in one or more characters of attacking words in their spam so that they can breach content-based spam filters. Due to this reason, we can say that content-based filters are not useful if they cannot understand the meaning of the words or phrases in the email.

So, In order to eliminate such problems in spam filtering, we developed a rule-based word-stemming technique, that can match words that look alike and sound alike.

This method for make spam filters has now become a widely-used technology. In this, we measured the incidence of the words in an email against their typical occurrence in a database of spam and ham(not spam) email messages using a statistical technique.

Text Matching or Similarity

Matching text objects to find similarity is one of the important areas of NLP. Some of the important applications of text matching are as follows:

- Automatic Spelling Correction,
- Data de-duplication,
- Genome analysis, etc.

Based on the requirement, there is a number of text-matching techniques available but in this article, we describe only the important techniques in a detailed manner:

Levenshtein Distance

In between two strings, the Levenshtein distance is defined as the minimum number of edits required to transform one string into the other, with the allowable edit operations such as

- Insertion,
- Deletion, or
- Substitution of a single character.

Phonetic Matching

A Phonetic matching algorithm takes a keyword as input (such as a person's name, location name, etc) and generates a character string that identifies a set of words that are (roughly) phonetically similar. Some of the very useful application or examples of this are:

- For searching large text Corpuses(corpora),
- Correcting spelling errors,
- Matching relevant names, etc.

The two main algorithms which we can use for the above purpose are as follows:

- Soundex
- Metaphone

Flexible String Matching

A complete text matching system includes different algorithms pipelined together to compute a variety of text variations. Regular expressions are really helpful for this purpose as well. Some other common techniques include are

- Exact string matching,
- Lemmatized matching,
- Compact matching (takes care of spaces, punctuation, etc).

Cosine Similarity

When the text is represented as vector notation, a general cosine similarity can be applied in order to measure vectorized similarity. Cosine similarity provides the closeness among two texts.

Machine Translation

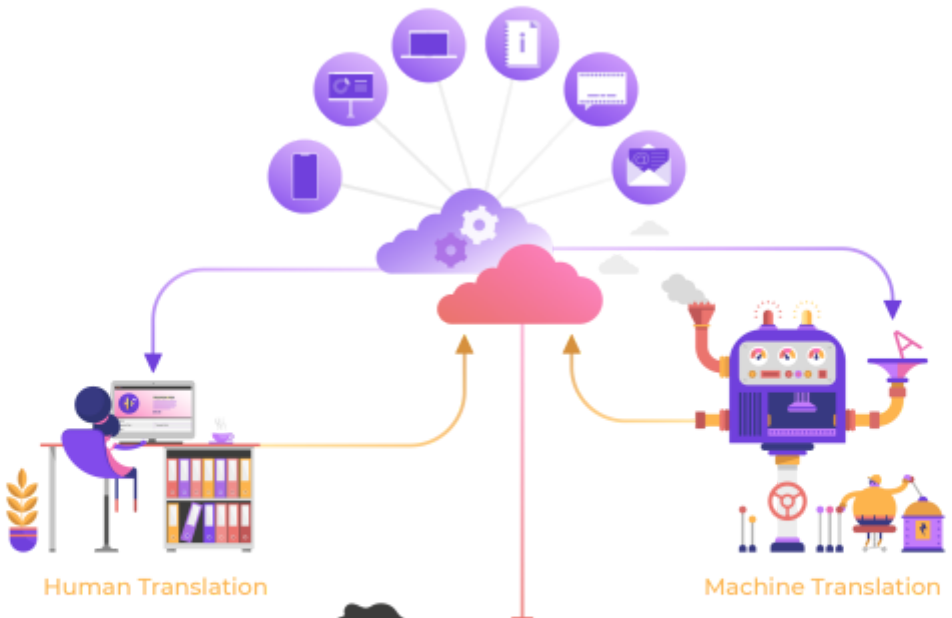


Image Source: Google Images

Machine Translation is an automatic system that translates text from one human language to another by taking care of grammar, semantics, and information about the real world, etc.

In simple words, Machine Translation is the process of translating one source language or text into another language.

Let's understand the following flowchart to understand the process of machine translation:

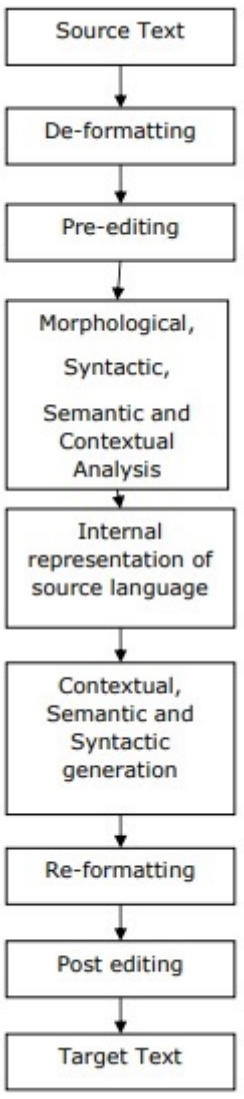


Image Source: Google Images

Types of Machine Translation Systems

Mainly, there are two different types of machine translation systems.

Multilingual Machine Translation System

These systems produce translations between any pair of languages. They can be either uni-directional in nature or bi-directional in nature.

Approaches to Machine Translation (MT)

Let's now discuss some of the important approaches to Machine Translation.

Direct Approach

It is the oldest approach of Machine Translation, so it is less popular. The systems that use this approach are capable of translating the source language directly to the target language. Such systems are bi-lingual and uni-directional in nature.

Interlingua Approach

The systems that use the Interlingua approach translate firstly Source language to an intermediate language, known as Interlingua (IL), and then translate Interlingua to Target Language. Below is the Machine Translation Pyramid which helps us to understand the Interlingua approach:

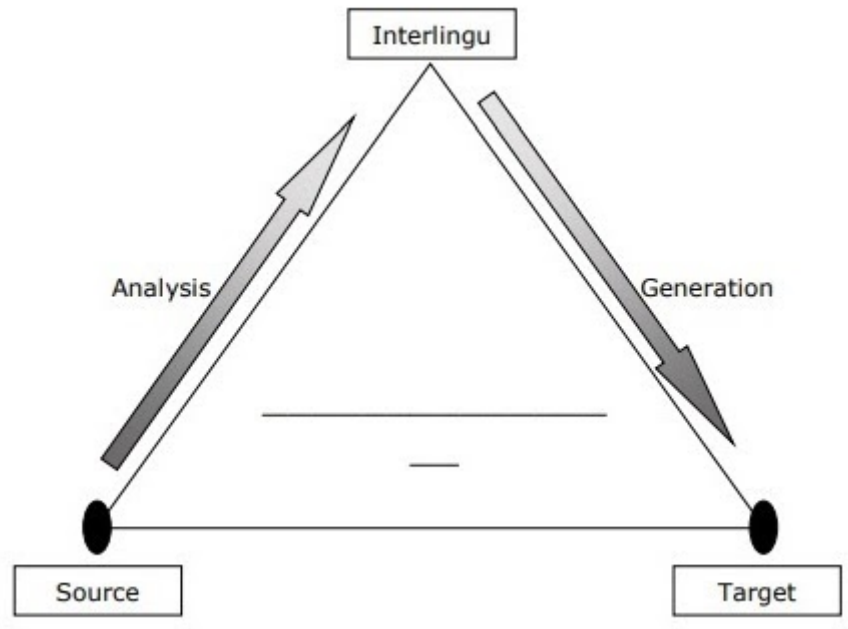


Image Source: Google Images

Transfer Approach

There are three stages involved in this approach of Machine translation:

- In the first stage, source language texts are converted to abstract Source Language -oriented representations.
- In the second stage, Source Language-oriented representations are converted into equivalent target language-oriented representations.
- In the third stage, the final text is produced.

Empirical MT Approach

This is an emerging approach for Machine Translation. Basically, this approach uses a large amount of raw data in the form of parallel corpora. Here, the raw data includes text and its translations. The following machine translation techniques used this approach:

- Analogybased,
- Example-based,
- Memory-based.

Coreference Resolution

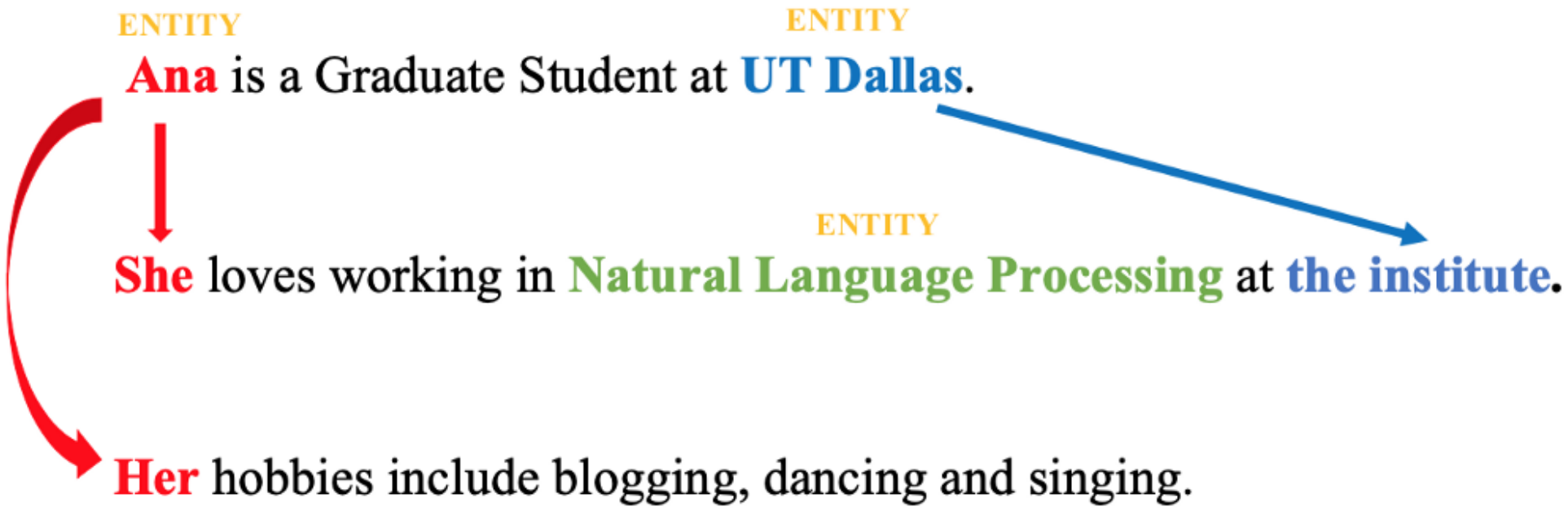


Image Source: Google Images

It is the process of finding relational links among the words (or phrases) within the sentences. Consider the following sentence:

```
”Chirag went to Kshitiz’s office to see the new pen. He looked at it for an hour.“
```

After observing the above sentence, humans can easily figure out that “he” denotes Chirag (and not Kshitiz), and that “it” denotes the pen (and not Kshitiz’s office).

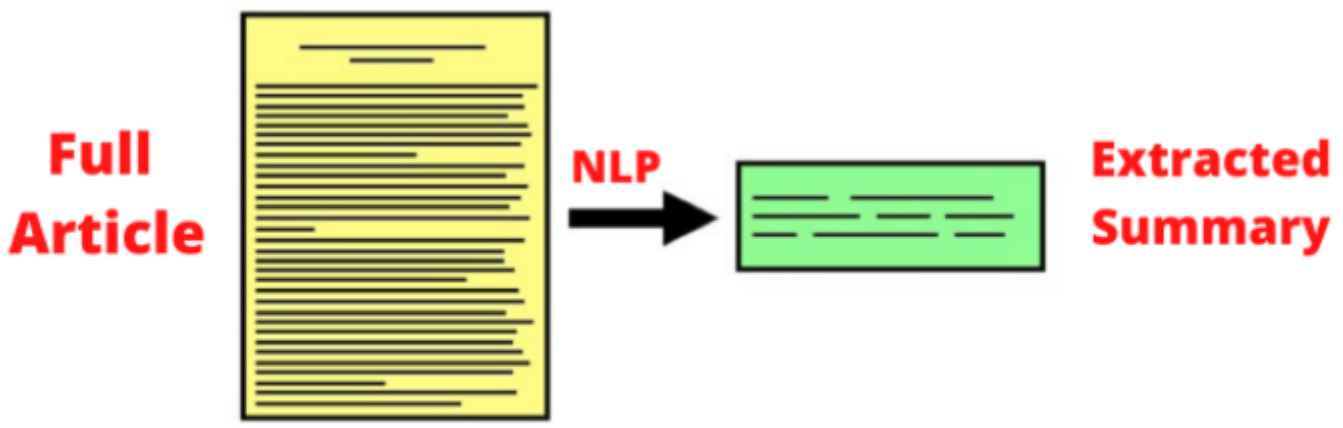
So, Coreference Resolution is the component of NLP that does this job automatically.

It is used in many applications including:

- Automatic Text Summarization,
- Question answering,
- Information extraction, etc.

For commercial purposes, Stanford CoreNLP provides a python wrapper.

Automatic Text Summarization



Convert Article to Summary

Image Source: Google Images

Text Summarization: Given a text article or paragraph, summarize it automatically to produce the most important and relevant sentences in order.

In this digital era, the most valuable thing is data or information. Now, the question that comes to mind is:

'NO', as the information is overloaded and our access to knowledge and information far exceeds our capacity to understand it. So, we are in serious need of automatic text summarization and information as the flood of information over the internet is not going to stop.

So, In simple words, we can say that text summarization is the technique to create a short, and accurate summary of longer text documents. It will help us to extract the relevant information in less amount of time. Therefore, NLP plays an important role in developing an automatic text summarization.

Question-Answering

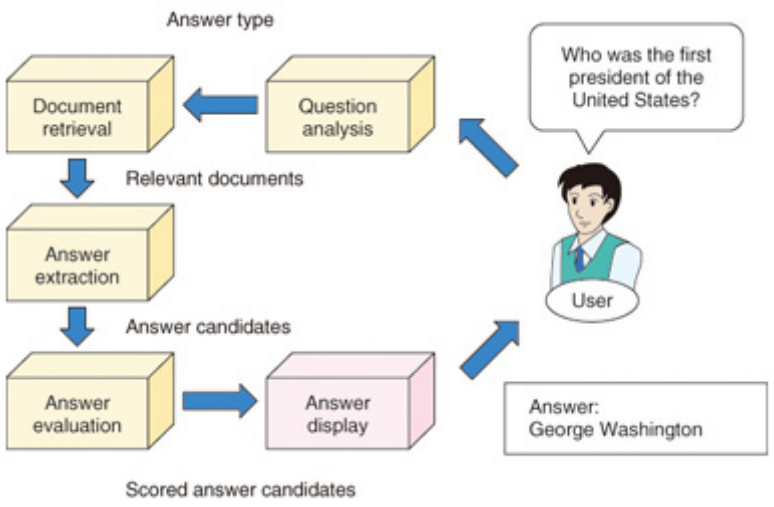


Image Source: Google Images

Question-answering is another important task of NLP. Search engines put the information of the world at our fingertips, but they are still lacking when it comes to answering the questions that are asked by human beings in their own natural language. Since this is a good application to work with, so Big tech companies like **Google, IBM, Microsoft**, are also working in this direction.

In the Computer Science discipline, Question-answering is within the fields of AI and NLP. The aim of this type of application is to build systems that automatically answer the questions asked by human beings in their own natural language. A computer system that understands the natural language has the capability of a program system to translate the sentences written by humans into an internal representation so that valid answers can be generated by the system. It generates exact answers by doing syntax and semantic analysis of the questions. But there are some challenges that the NLP faced while building a good question answering system such as,

- Lexical gap,
- Ambiguity,
- Multilingualism, etc.

Other Important NLP tasks

Some other important tasks of NLP are as follows:

Natural Language Generation and Understanding

Natural Language Generation is the process of converting information from computer databases or semantic intents into a language that is easily readable by humans.

Natural Language Understanding is the process of converting chunks of text into more logical structures that are easier for computer programs to manipulate.

Optical Character Recognition

Document to Information

This includes parsing textual data that is present in documents such as websites, files, and images to an analyzable and clean format.

This ends our Part-8 of the Blog Series on Natural Language Processing!

Other Blog Posts by Me

You can also check my previous blog posts.

Previous Data Science Blog posts.

LinkedIn

Here is my Linkedin profile in case you want to connect with me. I'll be happy to be connected with you.

Email

For any queries, you can mail me on Gmail .

End Notes

Thanks for reading!

I hope that you have enjoyed the article. If you like it, share it with your friends also. Something not mentioned or want to share your thoughts? Feel free to comment below And I'll get back to you. 😉

The media shown in this article are not owned by Analytics Vidhya and are used at the Author's discretion.

Part 2: Step by Step Guide to NLP - Knowledge Required to Learn NLP

How to Perform Basic Text Analysis without Training Dataset

Top 8 Python Libraries For Natural Language Processing (NLP) in 2021

blogathon NLP python text classification

About the Author

CHIRAG GOYAL

Our Top Authors