Introduction to Natural Language Processing

The essence of Natural Language Processing lies in making computers understand the natural language. That's not an easy task though. Computers can understand the structured form of data like spreadsheets and tables in the database, but human languages, texts, and voices form an unstructured category of data, and it becomes difficult for the computer to understand it, and there is the need for Natural Language Processing.

There's a lot of natural language data out there in various forms and it would get very easy if computers can understand and process that data. We can train the models in accordance with expected output in different ways. Humans have been writing for thousands of years, there are a lot of literature pieces available, and it would be great if we make computers understand that. But the task is never going to be easy. Various challenges are floating out there like understanding the correct meaning of the sentence, correct Named-Entity Recognition(NER), correct prediction of various parts of speech, and coreference resolution(the most challenging thing in my opinion).

Computers can't truly understand the human language. If we feed enough data and train a model properly, it can distinguish and try categorizing various parts of speech(noun, verb, adjective, supporter, etc…) based on previously fed data and experiences. If it encounters a new word it tried making the nearest guess which can be embarrassingly wrong few times.

It's very difficult for a computer to extract the exact meaning from a sentence. For example – The boy radiated fire like vibes. The boy had a very motivating personality or he actually radiated fire? As you see over here, parsing English with a computer is going to be complicated.

There are various stages involved in training a model. Solving a complex problem in Machine Learning means building a pipeline. In simple terms, it means breaking a complex problem into a number of small problems, making models for each of them and then integrating these models. A similar thing is done in NLP. We can break down the process of understanding English for a model into a number of small pieces.

It would be really great if a computer could understand that San Pedro is an island in Belize district in Central America with a population of 16, 444 and it is the second largest town in Belize. But to make the computer understand this, we need to teach computer very basic concepts of written language.

So let's start by creating an NLP pipeline. It has various steps which will give us the desired output(maybe not in a few rare cases) at the end.

Step #1: Sentence Segmentation

Breaking the piece of text in various sentences.

> ****Input 😗 *** San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America. According to 2015 mid-year estimates, the town has a population of about 16, 444. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.
>
> ****Output 😗 *** San Pedro is a town on the southern part of the island of Ambergris Caye in the 2.Belize District of the nation of Belize, in Central America.
> According to 2015 mid-year estimates, the town has a population of about 16, 444.
> It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.
> For coding a sentence segmentation model, we can consider splitting a sentence when it encounters any punctuation mark. But modern NLP pipelines have techniques to split even if the document isn't formatted properly.

Step #2: Word Tokenization

Breaking the sentence into individual words called as tokens. We can tokenize them whenever we encounter a space, we can train a model in that way. Even punctuations are considered as individual tokens as they have some meaning.

> ****Input 😗 *** San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America. According to 2015 mid-year estimates, the town has a population of about 16, 444. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.
>
> ****Output 😗 *** 'San Pedro', ' is', 'a', 'town' and so.

Step #3: Predicting Parts of Speech for each token

Predicting whether the word is a noun, verb, adjective, adverb, pronoun, etc. This will help to understand what the sentence is talking about. This can be achieved by feeding the tokens(and the words around it) to a pre-trained part-of-speech classification model. This model was fed a lot of English words with various parts of speech tagged to them so that it classifies the similar words it encounters in future in various parts of speech. Again, the models don't really understand the 'sense' of the words, it just classifies them on the basis of its previous experience. It's pure statistics.

The process will look like this:

> Input : Part of speech classification model
>
> Output : Town – common noun
>
> Is – verb

> The – determiner

And similarly, it will classify various tokens.

Step #4: Lemmatization

Feeding the model with the root word.

For example –

> There's a Buffalo grazing in the field.
>
> There are Buffaloes grazing in the field.

Here, both Buffalo and Buffaloes mean the same. But, the computer can confuse it as two different terms as it doesn't know anything. So we have to teach the computer that both terms mean the same. We have to tell a computer that both sentences are talking about the same concept. So we need to find out the most basic form or root form or lemma of the word and feed it to the model accordingly.

In a similar fashion, we can use it for verbs too. 'Play' and 'Playing' should be considered as same.

Step #5: Identifying stop words

There are various words in the English language that are used very frequently like 'a', 'and', 'the' etc. These words make a lot of noise while doing statistical analysis. We can take these words out. Some NLP pipelines will categorize these words as stop words, they will be filtered out while doing some statistical analysis. Definitely, they are needed to understand the dependency between various tokens to get the exact sense of the sentence. The list of stop words varies and depends on what kind of output are you expecting.

Step 6.1: Dependency Parsing

This means finding out the relationship between the words in the sentence and how they are related to each other. We create a parse tree in dependency parsing, with root as the main verb in the sentence. If we talk about the first sentence in our example, then 'is' is the main verb and it will be the root of the parse tree. We can construct a parse tree of every sentence with one root word(main verb) associated with it. We can also identify the kind of relationship that exists between the two words. In our example, 'San Pedro' is the subject and 'island' is the attribute. Thus, the relationship between 'San Pedro' and 'is', and 'island' and 'is' can be established.

Just like we trained a Machine Learning model to identify various parts of speech, we can train a model to identify the dependency between words by feeding many words. It's a complex task though. In 2016, Google released a new dependency parser Parsey McParseface which used a deep learning approach.

Step 6.2: Finding Noun Phrases

We can group the words that represent the same idea. For example – It is the second-largest town in the Belize District and largest in the Belize Rural South constituency. Here, tokens 'second', 'largest' and 'town' can be grouped together as they together represent the same thing 'Belize'. We can use the output of dependency parsing to combine such words. Whether to do this step or not completely depends on the end goal, but it's always quick to do this if we don't want much information about which words are adjective, rather focus on other important details.

Step #7: Named Entity Recognition(NER)

San Pedro is a town on the southern part of the island of Ambergris Caye in the 2. Belize District of the nation of Belize, in Central America.
Here, the NER maps the words with the real world places. The places that actually exist in the physical world. We can automatically extract the real world places present in the document using NLP.

If the above sentence is the input, NER will map it like this way:

> San Pedro – Geographic Entity
>
> Ambergris Caye – Geographic Entity
>
> Belize – Geographic Entity
>
> Central America – Geographic Entity

NER systems look for how a word is placed in a sentence and make use of other statistical models to identify what kind of word actually it is. For example – 'Washington' can be a geographical location as well as the last name of any person. A good NER system can identify this.

Kinds of objects that a typical NER system can tag:

> People's names.
>
> Company names.
>
> Geographical locations
>
> Product names.
>
> Date and time.
>
> Amount of money.
>
> Events

Step #8: Coreference Resolution:

San Pedro is a town on the southern part of the island of Ambergris Caye in the Belize District of the nation of Belize, in Central America. According to 2015 mid-year estimates, the town has a population of about 16, 444. It is the second-largest town in the Belize District and largest in the Belize Rural South constituency.

Here, we know that ‘it’ in the sentence 6 stands for San Pedro, but for a computer, it isn’t possible to understand that both the tokens are same because it treats both the sentences as two different things while it’s processing them. Pronouns are used with a high frequency in English literature and it becomes difficult for a computer to understand that both things are same.