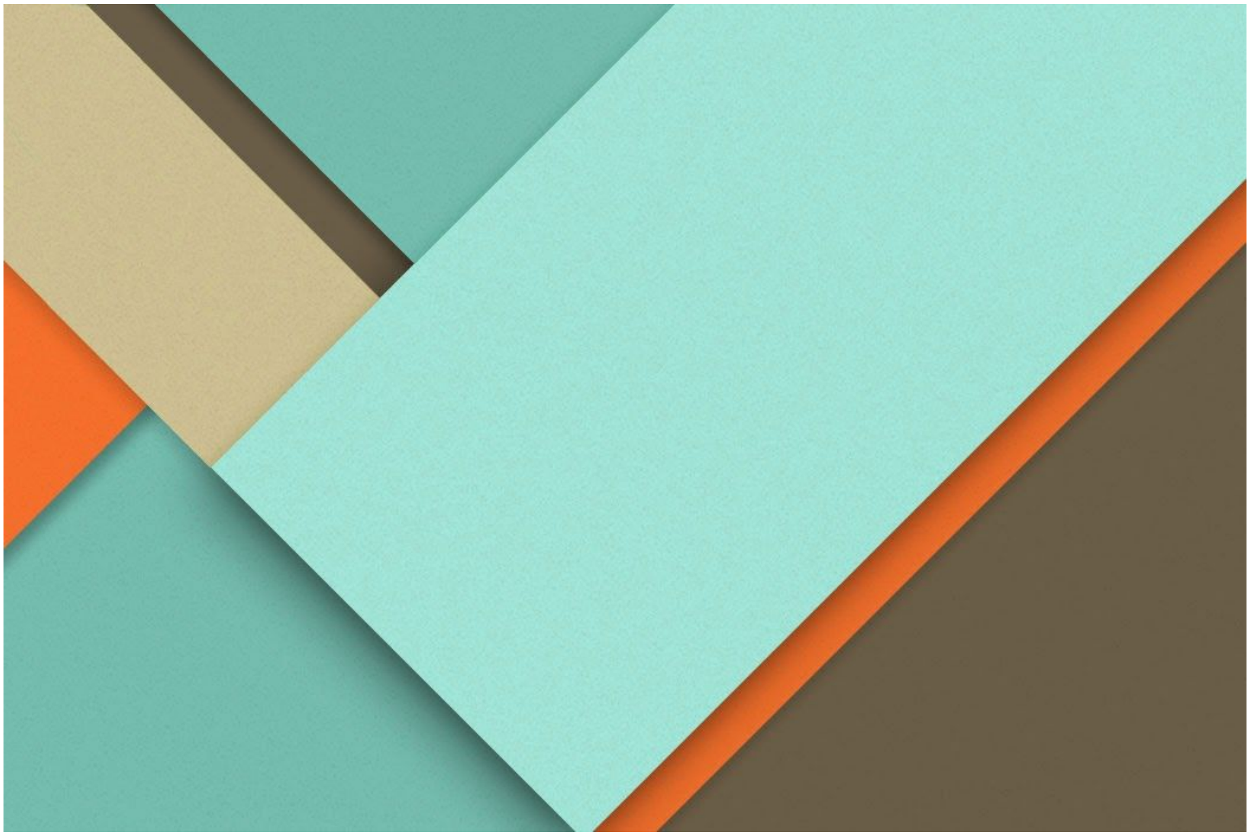

Natural Language Processing Assignment 1

13.08.2020

Anish Sachdeva
DTU/2K16/MC/013
Mathematics & Computing
Delhi Technological University (DTU)

Tweet Sentiment Analysis *(Morphology, Semantics)*

The following tweets have the same sentiment but the verbs used in them have different suffixes.

Yeah!! I am learning NLP 🎉

Yeah!! I will learn NLP 🎉

Both of the above tweets have the same user sentiment, but before we can analyse their sentiments we must first extract the words and remove stop words such as {*I*, *me* , *mine* , *we*...} etc. and after that the remaining words must be stemmed down.

So, both these tweets will be converted into [*Yeah*, *learn*, *NLP*, 🎉]

Similarly, this isn't just limited to analyzing tweets, but performing txt analysis of any kind be it on Facebook posts or on user reviews of a restaurant or just plain text sentiment analysis on Literature and books requires us to stem down the words just their roots and remove suffixes and prefixes for better analysis.

Sentiment Analysis also makes use of semantics as the order in which words are arranged makes a big difference in the meaning of the sentence. Eg.

I love that I am understanding NLP & *I understand that I am loving NLP* both have the same words but their meanings are different because of the arrangement of words in them. This arrangement can be studied through semantic analysis of the text.

Web narrator (*Phonetics, Syntax, Semantics*)

Web narrators are a very important feature for people with disabilities or people that have vision impairments. The web narrator is a part of all major operating systems and web browsers and the web narrator reads the text on a website and narrates it to the user. To do so it must have syntax knowledge of the language so that it knows when to give pauses and the different pasues required for punctuation.

It must also be able to correctly pronounce words. To correctly pronounce words that have different pronunciations and different meanings in different contexts it must further understand the context of a sentence or structure. To understand the context the system must also be able to understand the semantics and use Word Sense disambiguation (WSD).

Another example of a narrator is not one specifically created for the web, but just a general world view narrator that receives an incoming 2 dimensional signal (video stream) and first using Convolutional Neural Networks identifies objects in the video stream and then using applications in Natural Language Processing narrates a description of those objects to a person.

Google Translate *(Semantics, Syntax, Phonetics, Pragmatics)*

Google translate is a very popular application that is used throughout the globe by millions of people on a daily basis. For Google translate to achieve such a high accuracy in translation from one language to the other, it needs to understand the context of the input message and then based on the context produce the desired translated message. To understand the context of a message Google translate needs to understand the semantics of the given string.

To then create a valid message in the translated language from this message whose context now Google understands, Google translate also needs to understand the syntax of the desired output language as the translated message must be in proper grammatical syntax.

Furthermore Google translate also offers that the user can play the translated message as an audio and to perform that Google Translate needs to understand the phonetics of the language.

Google Translate also offers users the ability to dictate text in a particular language and then to automatically convert the audio into text. To convert audio into text Google Translate must also have an engine that understands the pragmatics of the audio sample and can recognize individual words in the dialogue.

Siri/Alexa/Cortana/Google Assistant *(Pragmatics, Discourse, Phonetics, Syntax, Semantics)*

All of the above are commonly used popular voice assistants present on our smartphones and other devices in the house. For the assistant to understand the question that we ask it and then give an answer to the question it needs to be able to convert our voice to text and then further understand the meaning/context of what we are trying to say and what the intent is from the message.

To translate our voice from audio to text the virtual assistant needs to have a pragmatics engine and then to understand the meaning of our message and use the context that has been built up over several messages the virtual assistant needs to have a discourse engine.

It will also need to generate the result as a text output and for that it requires a semantic and syntactic engine as the produced output must be a valid text message the user can understand.

Once the virtual assistant has understood the internet of the user it can retrieve the information requested from some service and then convert the text information (or any other format) to speech.

To convert from text to speech the virtual assistant needs to understand the phonetics of the language and also the syntax to make proper punctuation sounds and pauses.

Chatbot *(Pragmatics, Discourse, Syntax, Semantics)*

Chatbots are present on every popular platform these days and every company has chatbots on their websites to assist their clients with easy questions and frequently asked questions e.g. a travel website might receive many queries such as when is my flight? My flight date? Flight PNR number? Which airport? Contact number for airlines? Etc.

To answer routine questions the chatbot needs to understand the intent of the users from the last message and the conversation and hence they require Pragmatic and Discourse.

Once the chatbots have retrieved the required information the chatbots need to convey it in a format the user will understand, basically valid text. To create valid etxt the chatbot needs knowledge of syntax and how to form logical valid grammatical structures and hence it requires syntactic knowledge. To generate the language the chatbot also needs a semantic checker and generator that generates semantically correct language.

Spell Checking & Grammatical Error Checking *(Semantics, Syntax)*

All popular rich text formatting applications like MS Word and Google Docs have an inbuilt spell checker and also an inbuilt grammatical error checker for grammatical mistakes.

The spell check is relatively a sipler task that only requires that every word be compared to a set of all valid words that are present in a language. If a word is found that doesn't exist in the set of all valid words that is flagged as an incorrect spelling word.

Checking for correct grammar is much harder. Checking for grammar requires an engine that can understand the context of the sentence and also understand the syntax and grammar rules of a language. For this we require semantics and syntax engines.

Automatic Summarization *(Semantics, Pragmatics, Discourse)*

Summarization is a very helpful process and traditionally this is something that humans have always done on large pieces of information and text. But, using advanced NLP techniques we can accomplish this.

For a machine to summarize a large given corpus it needs to understand what are the important things and what information is superfluous and can be removed. For this, the engine needs to understand the context and meaning of the corpus and requires semantic analysis and needs to perform word sense Disambiguation (WSD). It also needs a Pragmatics and Discourse engine to understand the meaning of the entire corpus and then make intelligent extraction decisions.

Text Classification *(Pragmatics, Discourse)*

Text classification is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure, and categorize large corpi of data. Suppose we distribute documents in certain categories. A new document arrives, and it is necessary to determine to which category it belongs. By using NLP, text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.

To perform this task the text classifier needs to understand the context and meanings of different documents then group them accordingly. To understand the context is a task for pragmatics and discourse.

Analyzing Language Roots *(Morphology)*

Many words in the modern english language have originated or have been derived from French, Italian etc. Such as petit, piano, grand from French. And many words from the romantic languages such as French, Italian and Spanish have originated from Latin.

We can use morphological analysis on words to find where the word has originated from and what are the roots of a word. Using these roots we can see when in history a word crossed a language barrier and entered a new language.

E.g. bel is the latin root that signifies war and this has now entered english with several words like parabellum, bellicose etc.

Autocorrect & Autocomplete *(Semantics, Pragmatics)*

For performing autocorrect, the NLP engine needs to identify errors and then using semantics needs to autocorrect it with the correct version of the text. In Autocomplete the nlp engine needs to understand the context using pragmatics and then create a valid text stream using semantics and suggest that to the user.

Social Media Monitoring *(Morphology, Semantics)*

More and more people these days have started using social media for posting their thoughts about a particular product, policy, or matter. These could contain some useful information about an individual's likes and dislikes. Hence analyzing this unstructured data can help in generating valuable insights. Natural Language Processing comes to rescue here too.

Today, various NLP techniques are used by companies to analyze social media posts and know what customers think about their products. Companies are also using social media monitoring to understand the issues and problems that their customers are facing by using their products. Not just companies, even the government uses it to identify potential threats related to the security of the nation.

To accomplish this task the NLP engine needs to clean the data first and that requires many small steps such as removing stop words and punctuation marks and then stemming the rods. Stemming is reducing the word down to it's root and this is part of *Morphology*.

Then analysis needs to be run on the cleaned corpus and this analysis is semantic analysis that uses word disambiguation.

Targeted Advertising *(Morphology, Semantics)*

Similar to the social media monitoring mentioned above. Organizations can use content available on social media to extract useful information about the client/user and then use this information to specifically target the user with highly personalized advertisements.

Once again the data parsing step involves stemming and Morphology and the information extraction step is semantic text analysis.

Spam Filtering *(Morphology, Semantics)*

Spam filtering is a very common and widely used application of NLP, where incoming emails pass through a spam filter and are marked either as spam or not spam. To understand whether an email is spam or not, we need to perform semantic analysis on the email text and then using previously classified spam mails we are able to classify the given email as spam or not. This requires parsing of data which covers stemming (morphology) and then running a semantic analysis to extract key information points from the mail.

These key information points are then compared using a classification algorithm and are classified as either spam or not.

Question Answering Systems *(Syntax, Semantics)*

Another main application of natural language processing (NLP) is question-answering. Search engines put the information of the world at our fingertips, but they are still lacking when it comes to answer the questions posted by human beings in their natural language. We have big tech companies like Google working in this direction.

Question-answering is a Computer Science discipline within the fields of AI and NLP. It focuses on building systems that automatically answer questions posted by human beings in their 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 the valid answers can be generated by the system. The exact answers can be generated by doing syntax and semantic analysis of the questions. Lexical gap, ambiguity and multilingualism are some of the challenges for NLP in building a good question answering system.