A brief Introduction to Natural Language Processing

What is this NLP

According to Wikipedia, It is a subfield of Computer Science, Information Engineering and Artificial Intelligence, Concerned with the interaction between Computers and Human Languages

From a beginners perspective (i.e. my perspective), it is the INTERPRETER that is needed, to convey our expressions and thoughts (expressed in form of words) with appropriate meaning, for a computer program to understand and perform some actions accordingly

Look at how Jarvis(for the DC fans, it is Tony Stark’s “Alfred”, BUT with an Artificial Intelligence, so he can do more than just bringing breakfast in the morning) understands everything and even understands the vague instructions and can also read between the lines, AS during his first flight with the mark-II armour, the Iron Man asks Jarvis to check the weather and Air traffic control and listen to the ground control, instruction with which Jarvis understands that Tony is going to try to fly, so he cautions him that there are still terabytes of calculation needed before an actual flight.(which Tony disregards completely, and LUCKILY everything goes okay, or the movie would’ve sucked majorly!)

Of course Jarvis is fictional but our aim as data scientists is to make a computer system so accurate and efficient that it should understand not only the meaning of the words but also humour, sarcasm, sentiments, and intent. So in short to pass the Turing Test, we NEED Natural Language Processing.

But what IS a Natural Language

Paraphrasing the Wikipedia definition, it is the language that has been naturally established over the period of human evolution.

If only there was a conference in primitive ages where the leader said that “Okay you guys go to China and only use mandarin, here is the structure, grammar and all the words you can use.”

Even English, which is the one of the most spoken language across the world has borrowed many words from different languages, plethora of words originated from Latin and Greek words, and cultural influences change the usage of words altogether, isn’t that cool?(the word cool could mean fashionable or exciting or of course refer to the a thing’s temperature)

Most of today’s textual data is unstructured, there are no “on-the-nose” dialogue in real life, and we use gestures, pauses, the tonality of sentences, and with the understanding of the context of situation, the person next to us understands what we’re trying to convey.

\*\*GIF\*\*“YOU’RE TEARING ME APART LISA!!”

So let’s get back to processing such unstructured text data

What could we achieve if we could accurately find out, the meaning of the text?

1. We can classify and organise chucks of text, (Movie Reviews: Good or Bad based on user comments, no rating system will be needed, Google News collects news articles from different news sources under “ONE GOOGLE NEWS”)
2. We summarise a huge article (like this one, sorry!) into a paragraph using computer (make that computer read it and make it paraphrase what this bloke has been on about!)
3. Word/ phrase recommendation (Check LinkedIn it gives you a list of plausible and “Professional” replies in the chat window! I never run out of responses now! Wish they could make someone like this for Online Dating apps too! That’s a million dollar idea, right there!)
4. Machine Translation (translate the instruction or just translate to other language using a computer program(google translate))
5. Sentimental Analysis (recommend a sad-face emoji right after you type “Mom, I failed the exam, again!” and a unimpressed-face emoji when your mom says “You don’t say, Mr. GamerBoy1776”)
6. Speech recognition, with the help of a few additional libraries we can convert speech-to-text (Virtual assistants like Siri, Alexa and YouTube’s auto generated closed captions and many more)
7. Topic Segmentation, Separating articles to categories by understanding their content
8. Extracting the most useful Information from the given text (extracting OTP from a text message)
9. Chat bot, the hottest thing of all the NLP application (Currently majority of service industry use domain specific Chat bots for an initial interaction)
10. Advertisement Matching(If I could spy on all the customers, and find out what they like, I can show them targeted advertisements, not that I personally would do that, but If I want to, I could, with NLP! \*\*Evil Laughter\*\* \*\*MuHahahaha\*\*)

Two Main Components of NLP

NLU or Natural Language Understanding

As easy as it seems for us humans, understanding the nuances of natural language is the most difficult task in NLP. A great NLU will simplify 90% of your task. All humans should have a good NLU as well. As they say, you have to be a good listener to be a good speaker, or you have to be a good reader to be a good writer.

So the NLU does the reading and listening, it breaks down sentences on the basis of words and tries to analyse the input.

It also maps the given input to useful representation, based on the grammatical rules of the language, which so it makes it easier for us to understand what NLU is actually doing, how verbs, nouns, pronouns, adjectives such parts of speech are getting recognized by the program.

NLG or Natural Language Generation

Once your input is properly analysed and understood, IF and WHEN, the program has to respond to the user using natural language the NLG comes into picture.

It,

Plans the respond, i.e. first it finds the relevant content to respond.

Plans the sentences, i.e. constructs the sentences accordingly, with proper grammar and taking care of the tone of the sentence

Realizes the text, displays or delivers the text in a structured format

SO Let’s understand why NLU is the HARD PART:

English, which is a language of many, still has many flaws, but we as humans have the natural ability to understand what someone is trying to convey, based on the context or situation and non-verbal cues. Nonetheless, these flaws called as ambiguities. If we turn off some of our senses (mainly the common sense), we might get into some funny conversations.

There are three types of ambiguities,

1. Lexical Ambiguity (ambiguity of words)
2. Syntactical Ambiguity (ambiguity of the sentence)
3. Referential Ambiguity (ambiguity of the structure)

Let’s get right to details of each

1. Lexical Ambiguity

When words mean two different things entirely,

E.g. 1. We saw her duck (did she have a pet duck? Or did she just duck because the ball could’ve hit her in the head)

E.g. 2. Are you nuts? (There might be some nuts lying around and a crazy person might be talking to them, OR it might be an impulsive expression to tell the other person that whatever he is doing or saying is absolutely crazy)

Here is the Link to a resource if you want to get in the details from a grammatical perspective

1. Syntactical Ambiguity

When the sentences are such a way that it can cause ambiguous understanding,

E.g. 1. The other night, I shot an elephant in my pyjamas, how the hell did he get into my pyjamas, I’ll never know (This famous joke exploits the syntactical ambiguity, we initially think that the guy was in pyjamas, but our expectations are shattered when he says the next line, making us laugh)

E.g. 2. To enter the premises, you’ll need either a voucher or 10 bucks and your driving licence. (Do I only need a voucher? or if I don’t have it I’ll need 10bucks and the DL OR do I need one of these either a voucher or 10 bucks and I’ll have to show the DL along with any of those two), Gosh! It’s even hard to explain.

If(voucher):

Thou Shall Pass

Else if(10$ and DL):

Thou Shall Pass

Else:

Thou shall not pass

OR is it

If(voucher or 10$):

If(DL)

Pass

Else:

Nope

Hope you got this! So this is Syntactical Ambiguity!

1. Referential Ambiguity

Tends to happen when we use pronouns

E.g. The guard hit him in the head, *he* was crazy! (Was it the guard who was crazy? or was the guy who got hit in the head, crazy?)

So the NLU has to do a herculean task to get the right meaning out of the sentences.

Various Natural Language Processing Libraries

1. NLTK (Natural Language Tool Kit)
2. TextBlob
3. SpaCy
4. Gensim
5. Pattern
6. Stanford CoreNLP