Introduction to NLP

Introduction to Natural Language Processing

Natural language processing (NLP) is the intersection of computer science, linguistics and machine learning. The field focuses on communication between computers and humans in natural language and NLP is all about making computers understand and generate human language. Applications of NLP techniques include voice assistants like Amazon’s Alexa and Apple’s Siri, but also things like machine translation and text-filtering.

WHAT IS NATURAL LANGUAGE PROCESSING?

Natural language processing studies interactions between humans and computers to find ways for computers to process written and spoken words similar to how humans do. The field blends computer science, linguistics and machine learning.

Natural language processing has heavily benefited from recent advances in machine learning, especially from deep learning techniques. The field is divided into the three parts:

Speech recognition—the translation of spoken language into text.

Natural language understanding—a computer’s ability to understand language.

Natural language generation—the generation of natural language by a computer.

2. Why Natural Language Processing Is Difficult

Human language is special for several reasons. It is specifically constructed to convey the speaker/writer’s meaning. It is a complex system, although little children can learn it pretty quickly.

Another remarkable thing about human language is that it is all about symbols. According to Chris Manning, a machine learning professor at Stanford, it is a discrete, symbolic, categorical signaling system. This means we can convey the same meaning in different ways (i.e., speech, gesture, signs, etc.) The encoding by the human brain is a continuous pattern of activation by which the symbols are transmitted via continuous signals of sound and vision.

Understanding human language is considered a difficult task due to its complexity. For example, there are an infinite number of different ways to arrange words in a sentence. Also, words can have several meanings and contextual information is necessary to correctly interpret sentences. Every language is more or less unique and ambiguous. Just take a look at the following newspaper headline “The Pope’s baby steps on gays.” This sentence clearly has two very different interpretations, which is a pretty good example of the challenges in natural language processing.

3. Syntactic and Semantic Analysis

Syntactic analysis (syntax) and semantic analysis (semantic) are the two primary techniques that lead to the understanding of natural language. Language is a set of valid sentences, but what makes a sentence valid? Syntax and semantics.

Syntax is the grammatical structure of the text, whereas semantics is the meaning being conveyed. A sentence that is syntactically correct, however, is not always semantically correct. For example, “cows flow supremely” is grammatically valid (subject—verb—adverb) but it doesn’t make any sense.

SYNTACTIC ANALYSIS

Syntactic analysis, also referred to as syntax analysis or parsing, is the process of analyzing natural language with the rules of a formal grammar. Grammatical rules are applied to categories and groups of words, not individual words. Syntactic analysis basically assigns a semantic structure to text.

For example, a sentence includes a subject and a predicate where the subject is a noun phrase and the predicate is a verb phrase. Take a look at the following sentence: “The dog (noun phrase) went away (verb phrase).” Note how we can combine every noun phrase with a verb phrase. Again, it’s important to reiterate that a sentence can be syntactically correct but not make sense.

SEMANTIC ANALYSIS

The way we understand what someone has said is an unconscious process relying on our intuition and knowledge about language itself. In other words, the way we understand language is heavily based on meaning and context. Computers need a different approach, however. The word “semantic” is a linguistic term and means “related to meaning or logic.”

Semantic analysis is the process of understanding the meaning and interpretation of words, signs and sentence structure. This lets computers partly understand natural language the way humans do. I say this partly because semantic analysis is one of the toughest parts of natural language processing and it’s not fully solved yet.

Speech recognition, for example, has gotten very good and works almost flawlessly, but we still lack this kind of proficiency in natural language understanding. Your phone basically understands what you have said, but often can’t do anything with it because it doesn’t understand the meaning behind it. Also, some of the technologies out there only make you think they understand the meaning of a text. An approach based on keywords or statistics or even pure machine learning may be using a matching or frequency technique for clues as to what the text is “about.” But, because they don’t understand the deeper relationships within the text, these methods are limited.

4. Natural Language Processing Techniques for Understanding Text

Let’s look at some of the most popular techniques used in natural language processing. Note how some of them are closely intertwined and only serve as subtasks for solving larger problems.

NATURAL LANGUAGE PROCESSING TECHNIQUES

- Parsing
- Stemming
- Text Segmentation
- Named Entity Recognition
- Relationship Extraction
- Sentiment Analysis

PARSING

What is parsing? According to the dictionary, to parse is to “resolve a sentence into its component parts and describe their syntactic roles.”

That actually nailed it but it could be a little more comprehensive. Parsing refers to the formal analysis of a sentence by a computer into its constituents, which results in a parse tree showing their syntactic relation to one another in visual form, which can be used for further processing and understanding.

Below is a parse tree for the sentence “The thief robbed the apartment.” Included is a description of the three different information types conveyed by the sentence.

The letters directly above the single words show the parts of speech for each word (noun, verb and determiner). One level higher is some hierarchical grouping of words into phrases. For example, “the thief” is a noun phrase, “robbed the apartment” is a verb phrase and when put together the two phrases form a sentence, which is marked one level higher.

But what is actually meant by a noun or verb phrase? Noun phrases are one or more words that contain a noun and maybe some descriptors, verbs or adverbs. The idea is to group nouns with words that are in relation to them.

A parse tree also provides us with information about the grammatical relationships of the words due to the structure of their representation. For example, we can see in the structure that “the thief” is the subject of “robbed.”

With structure I mean that we have the verb (“robbed”), which is marked with a “V” above it and a “VP” above that, which is linked with a “S” to the subject (“the thief”), which has a “NP” above it. This is like a template for a subject-verb relationship and there are many others for other types of relationships.

STEMMING

Stemming is a technique that comes from morphology and information retrieval which is used in natural language processing for pre-processing and efficiency purposes. It's defined by the dictionary as to "originate in or be caused by."

Basically, stemming is the process of reducing words to their word stem. A "stem" is the part of a word that remains after the removal of all affixes. For example, the stem for the word "touched" is "touch." "Touch" is also the stem of "touching," and so on.

You may be asking yourself, why do we even need the stem? Well, the stem is needed because we're going to encounter different variations of words that actually have the same stem and the same meaning. For example:

I was taking a ride in the car.

I was riding in the car.

These two sentences mean the exact same thing and the use of the word is identical.

Now, imagine all the English words in the vocabulary with all their different fixations at the end of them. To store them all would require a huge database containing many words that actually have the same meaning. This is solved by focusing only on a word's stem. Popular algorithms for stemming include the Porter stemming algorithm from 1979, which still works well.

TEXT SEGMENTATION

Text segmentation in natural language processing is the process of transforming text into meaningful units like words, sentences, different topics, the underlying intent and more. Mostly, the text is segmented into its component words, which can be a difficult task, depending on the language. This is again due to the complexity of human language. For example, it works relatively well in English to separate words by spaces, except for words like “icebox” that belong together but are separated by a space. The problem is that people sometimes also write it as “ice-box.”

NAMED ENTITY RECOGNITION

Named entity recognition (NER) concentrates on determining which items in a text (i.e. the “named entities”) can be located and classified into predefined categories. These categories can range from the names of persons, organizations and locations to monetary values and percentages.

For example:

Before NER: *Martin bought 300 shares of SAP in 2016.*

After NER: *[Martin]Person bought 300 shares of [SAP]Organization in [2016]Time.*

RELATIONSHIP EXTRACTION

Relationship extraction takes the named entities of NER and tries to identify the semantic relationships between them. This could mean, for example, finding out who is married to whom, that a person works for a specific company and so on. This problem can also be transformed into a classification problem and a machine learning model can be trained for every relationship type.

SENTIMENT ANALYSIS

With sentiment analysis we want to determine the attitude (i.e. the sentiment) of a speaker or writer with respect to a document, interaction or event. Therefore it is a natural language processing problem where text needs to be understood in order to predict the underlying intent. The sentiment is mostly categorized into positive, negative and neutral categories.

With the use of sentiment analysis, for example, we may want to predict a customer’s opinion and attitude about a product based on a review they wrote. Sentiment analysis is widely applied to reviews, surveys, documents and much more.

If you’re interested in using some of these techniques with Python, take a look at the Jupyter Notebook about Python’s natural language toolkit (NLTK) that I created. You can also check out my blog post about building neural networks with Keras where I train a neural network to perform sentiment analysis.

5. Benefits of Natural Language Processing

Now that we’ve learned about how natural language processing works, it’s important to understand what it can do for businesses.

ENHANCED DATA ANALYSIS

While NLP and other forms of AI aren’t perfect, natural language processing can bring objectivity to data analysis, providing more accurate and consistent results.

FASTER INSIGHTS

With the Internet of Things and other advanced technologies compiling more data than ever, some data sets are simply too overwhelming for humans to comb through. Natural language processing can quickly process massive volumes of data, gleaning insights that may have taken weeks or even months for humans to extract.

INCREASED EMPLOYEE PRODUCTIVITY

NLP handles mundane tasks like sifting through data sets, sorting emails and assessing customer responses. With these repetitive responsibilities out of the way, workers are freed up to focus on more complex and pressing matters.

HIGHER-QUALITY CUSTOMER EXPERIENCE

In the form of chatbots, natural language processing can take some of the weight off customer service teams, promptly responding to online queries and redirecting customers when needed. NLP can also analyze customer surveys and feedback, allowing teams to gather timely intel on how customers feel about a brand and steps they can take to improve customer sentiment.

6. NLP Use Cases

Keeping the advantages of natural language processing in mind, let’s explore how different industries are applying this technology.

CUSTOMER SERVICE

While NLP-powered chatbots and callbots are most common in customer service contexts, companies have also relied on natural language processing to power virtual assistants. These assistants are a form of conversational AI that can carry on more sophisticated discussions. And if NLP is unable to resolve an issue, it can connect a customer with the appropriate personnel.

MARKETING

Gathering market intelligence becomes much easier with natural language processing, which can analyze online reviews, social media posts and web forums. Compiling this data can help marketing teams understand what consumers care about and how they perceive a business' brand.

HUMAN RESOURCES

Recruiters and HR personnel can use natural language processing to sift through hundreds of resumes, picking out promising candidates based on keywords, education, skills and other criteria. In addition, NLP's data analysis capabilities are ideal for reviewing employee surveys and quickly determining how employees feel about the workplace.

E-COMMERCE

Natural language processing can help customers book tickets, track orders and even recommend similar products on e-commerce websites Teams can also use data on customer purchases to inform what types of products to stock up on and when to replenish inventories.

FINANCE

In finance, NLP can be paired with machine learning to generate financial reports based on invoices, statements and other documents. Financial analysts can also employ natural language processing to predict stock market trends by analyzing news articles, social media posts and other online sources for market sentiments.

INSURANCE

Insurance companies can assess claims with natural language processing since this technology can handle both structured and unstructured data. NLP can also be trained to pick out unusual information, allowing teams to spot fraudulent claims.

EDUCATION

NLP-powered apps can check for spelling errors, highlight unnecessary or misapplied grammar and even suggest simpler ways to organize sentences. Natural language processing can also translate text into other languages, aiding students in learning a new language.

HEALTHCARE

Healthcare professionals can develop more efficient workflows with the help of natural language processing. During procedures, doctors can dictate their actions and notes to an app, which produces an accurate transcription. NLP can also scan patient documents to identify patients who would be best suited for certain clinical trials.

MANUFACTURING

With its ability to process large amounts of data, NLP can inform manufacturers on how to improve production workflows, when to perform machine maintenance and what issues need to be fixed in products. And if companies need to find the best price for specific materials, natural language processing can review various websites and locate the optimal price.

CYBERSECURITY

IT and security teams can deploy natural language processing to filter out suspicious emails based on word choice, sentiment and other factors. This makes it easier to protect different departments from spam, phishing scams and other cyber attacks. With its ability to understand data, NLP can also detect unusual behavior and alert teams of possible threats.