# What Is NLP?

NLP: NLP stands for Natural Language Processing. It is a subfield of artificial intelligence (AI) and linguistics concerned with the interactions between computers and human (natural) languages. NLP focuses on enabling computers to understand, interpret, and generate human language in a way that is both meaningful and useful. It involves tasks such as text classification, sentiment analysis, language translation, named entity recognition, speech recognition, and text generation, among others. NLP techniques often involve machine learning and statistical methods to analyze and understand human language data.
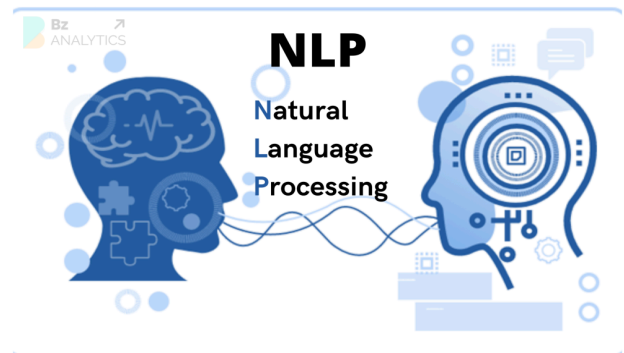
Examples of NLP:

1. **Sentiment Analysis:** NLP techniques are used to analyze text data from sources like social media, customer reviews, and surveys to determine the sentiment or opinion expressed. This helps businesses gauge public opinion about their products, services, or brand reputation.
2. **Language Translation:** NLP algorithms are employed in machine translation systems such as Google Translate or Microsoft Translator to translate text from one language to another automatically. These systems use statistical methods or neural networks to understand and generate translations accurately.
3. **Named Entity Recognition (NER):** NLP models can identify and classify named entities such as names of people, organizations, locations, dates, and other types of entities within text data. NER is useful in applications like information extraction, entity linking, and question answering systems.
4. **Text Summarization:** NLP techniques are used to automatically generate concise summaries of longer texts, such as news articles or research papers. Extractive summarization methods select important sentences or phrases from the original text, while abstractive summarization methods generate new sentences that convey the main points of the text.
5. **Chatbots and Virtual Assistants:** NLP is used in the development of conversational agents, chatbots, and virtual assistants that can understand and respond to natural language queries from users. These systems employ techniques like natural language understanding (NLU) and natural language generation (NLG) to interact with users in a human-like manner.
6. **Text Classification:** NLP models are trained to classify text documents into predefined categories or labels based on their content. Applications include spam detection in emails, sentiment analysis of social media posts, topic categorization of news articles, and identifying the intent of customer queries in customer support systems.
7. **Information Retrieval:** NLP techniques are used to retrieve relevant information from large collections of text documents in response to user queries. Search engines like Google use NLP algorithms to understand user queries and return the most relevant web pages based on their content and relevance to the query

Relation Between NLP and ML, DL
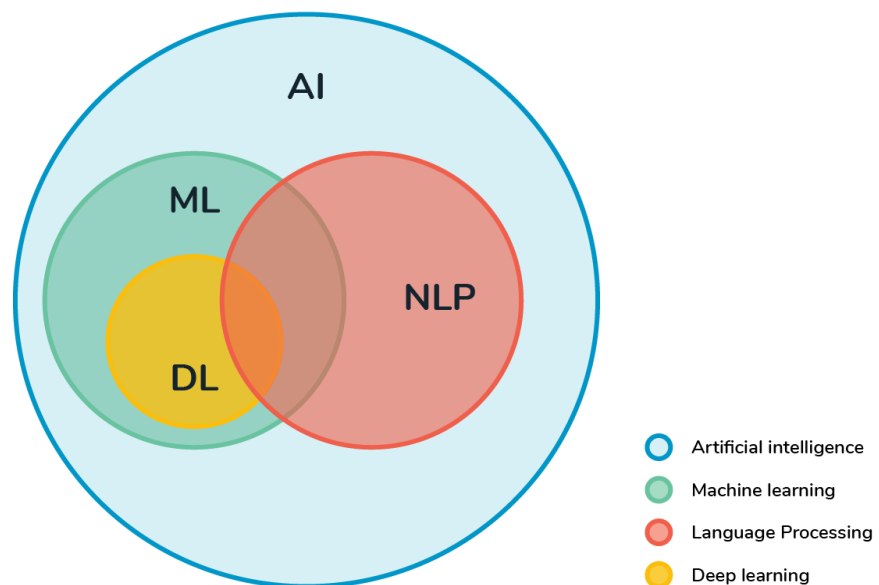
Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) are closely related fields within artificial intelligence (AI). Here's how they intersect:

- **NLP and ML:** Machine learning plays a significant role in NLP tasks. ML algorithms are used to process and analyze large amounts of language data, enabling tasks such as text classification, sentiment analysis, and named entity recognition. ML models learn patterns and relationships in language data, which are then used to make predictions or perform tasks without being explicitly programmed.

- **NLP and DL:** Deep learning is a subset of machine learning that utilizes neural networks with many layers to learn complex patterns in data. DL has revolutionized NLP by enabling the development of models known as neural networks that can process and understand language in a more sophisticated manner. Models like recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer architectures (e.g., BERT, GPT) have achieved remarkable success in various NLP tasks, including language translation, text generation, and sentiment analysis.

In summary, NLP relies heavily on both machine learning and deep learning techniques. Machine learning provides the foundational algorithms and methods for analyzing language data, while deep learning, as a subset of ML, offers more advanced neural network architectures capable of handling complex language tasks with high accuracy.

Components of NLP In AI

Natural language generation (NLG)

NLG is a method of creating meaningful phrases and sentences (natural language) from data. It comprises three stages: text planning, sentence planning, and text realization.

- Text planning: Retrieving applicable content.
- Sentence planning: Forming meaningful phrases and setting the sentence tone.
- Text realization: Mapping sentence plans to sentence structures.
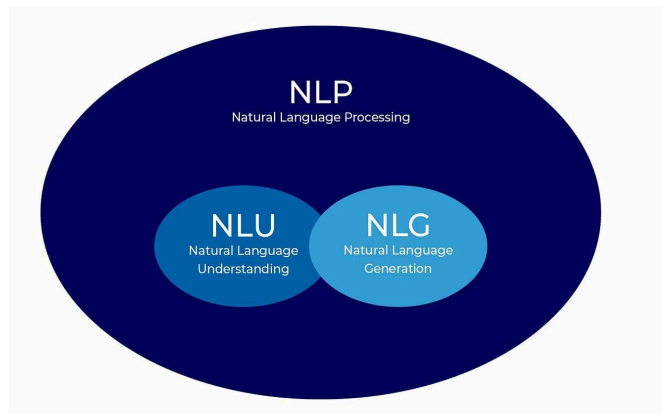
Chatbots, machine translation tools, analytics platforms, voice assistants, sentiment analysis platforms, and AI-powered transcription tools are some applications of NLG.

**Natural language understanding (NLU)**

NLU enables machines to understand and interpret human language by extracting metadata from content. It performs the following tasks:

- Helps analyze different aspects of language.
- Helps map the input in natural language into valid representations.

NLU is more difficult than NLG tasks owing to referential, lexical, and syntactic ambiguity.

- **Lexical ambiguity:** This means that one word holds several meanings. For example, "The man is looking for the match." The sentence is ambiguous as 'match' could mean different things such as a partner or a competition.
- **Syntactic ambiguity:** This refers to a sequence of words with more than one meaning. For example, "The fish is ready to eat." The ambiguity here is whether the fish is ready to eat its food or whether the fish is ready for someone else to eat. This ambiguity can be resolved with the help of the part-of-speech tagging technique.
- **Referential ambiguity:** This involves a word or a phrase that could refer to two or more properties. For example, Tom met Jerry and John. They went to the movies. Here, the pronoun 'they' causes ambiguity as it isn't clear who it refers to.

Pipeline of NLP In AI?

Natural Language Processing Pipeline
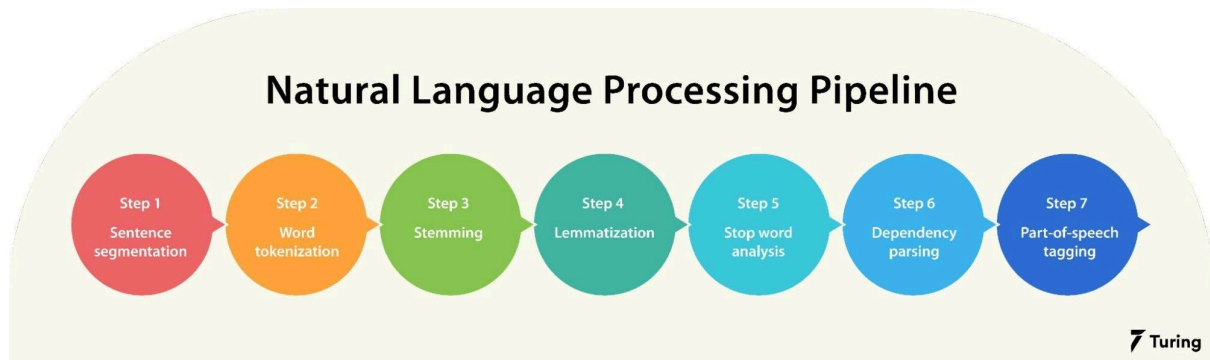
The NLP pipeline comprises a set of steps to read and understand human language.

Step 1: Sentence segmentation

Sentence segmentation is the first step in the NLP pipeline. It divides the entire paragraph into different sentences for better understanding. For example, "London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium."

After using sentence segmentation, we get the following result:

1. "London is the capital and most populous city of England and the United Kingdom."
2. "Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia."
3. "It was founded by the Romans, who named it Londinium."

Step 2: Word tokenization

Word tokenization breaks the sentence into separate words or tokens. This helps understand the context of the text. When tokenizing the sentence "London is the capital and most populous city of England and the United Kingdom", it is broken into separate words, i.e., "London", "is", "the", "capital", "and", "most", "populous", "city", "of", "England", "and", "the", "United", "Kingdom", ".".

Step 3: Stemming

Stemming helps in preprocessing text. The model analyzes the parts of speech to figure out what exactly the sentence is talking about.

Stemming normalizes words into their base or root form. In other words, it helps to predict the parts of speech for each token. For example, intelligently, intelligence, and intelligent. These words originate from a single root word 'intelligen'. However, in English there's no such word as 'intelligen'.

Image source: Medium.com

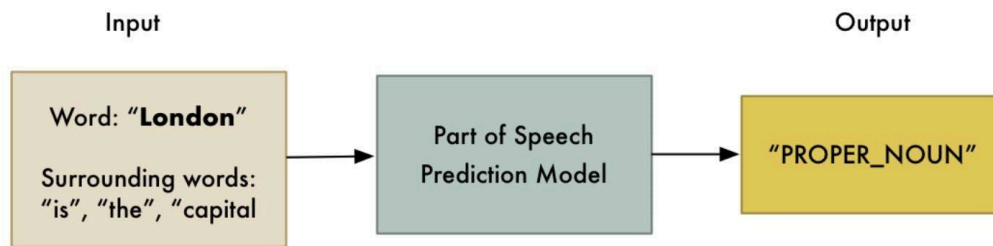The result will be:

Image source: Medium.com

Step 4: Lemmatization

Lemmatization removes inflectional endings and returns the canonical form of a word or lemma. It is similar to stemming except that the lemma is an actual word. For example, 'playing' and 'plays' are forms of the word 'play'. Hence, play is the lemma of these words. Unlike a stem (recall 'intelligen'), 'play' is a proper word.

Step 5: Stop word analysis

The next step is to consider the importance of each and every word in a given sentence. In English, some words appear more frequently than others such as "is", "a", "the", "and". As they appear often, the NLP pipeline flags them as stop words. They are filtered out so as to focus on more important words.

**Step 6: Dependency parsing**

Next comes dependency parsing which is mainly used to find out how all the words in a sentence are related to each other. To find the dependency, we can build a tree and assign a single word as a parent word. The main verb in the sentence will act as the root node.
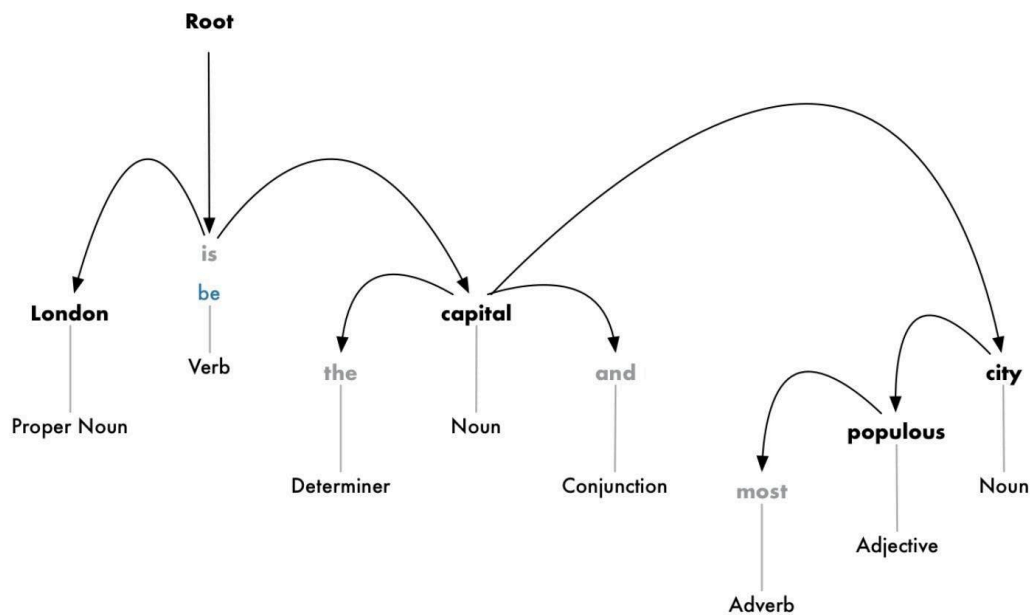
Step 7: Part-of-speech (POS) tagging

POS tags contain verbs, adverbs, nouns, and adjectives that help indicate the meaning of words in a grammatically correct way in a sentence.
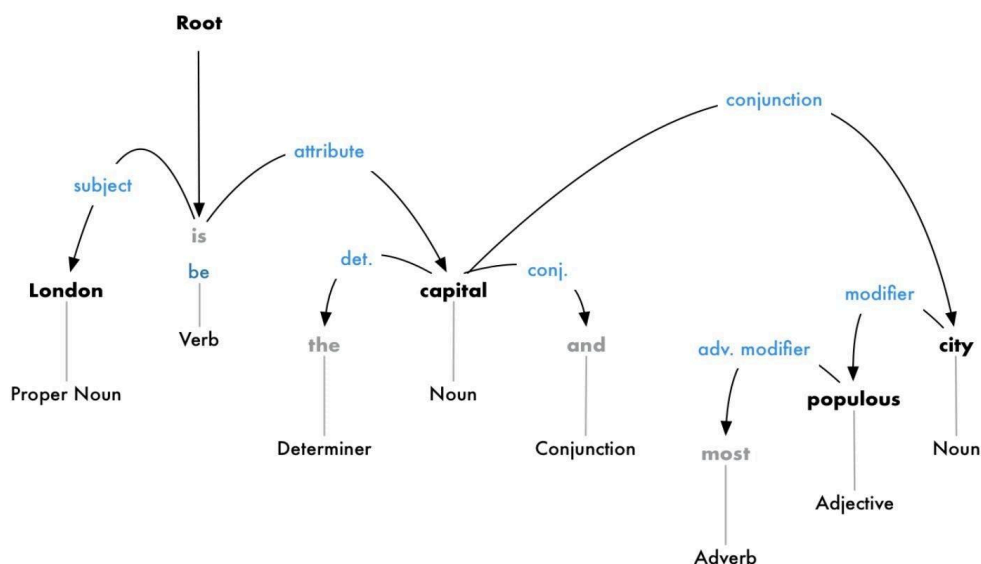
Natural language processing is used when we want machines to interpret human language. The main goal is to make meaning out of text in order to perform certain tasks automatically such as spell check, translation, for social media monitoring tools, and so on.

Application Of NLP

Benefits of NLP

Natural language processing has gained so much attention because of its practical importance. Later in this article, we will demonstrate some of the applications of NLP, but for now, let us concentrate on its main benefits, which include:

- **Large-scale analysis:** NLP allows us to process vast amounts of text documents of all kinds very quickly, which is impossible to do manually.
- **Accuracy:** NLP tools get better with experience, so their accuracy will only increase with time. Manual analysis, on the other hand, may lead to mistakes. Humans may be unattentive, especially while working with repetitive tasks.
- **Improved user experience:** Natural language processing allows for the automation of many routine tasks. It can sort trouble tickets, categorize customer feedback, and even communicate with customers. NLP can optimize website search engines, give better recommendations, or moderate user-generated content.
- **Better market understanding:** Machine learning and NLP, in particular, allow us to better search for and analyze relevant data. Businesses can use social media comments, customer reviews, trends, statistics to improve their services, adjust their supply, prices, etc.
- **Automation and efficiency:** Tasks like translation or text summarization can be fully automated. This, in turn, saves time and money. Some natural language processing tools make our daily tasks more comfortable, for example, grammar checkers.

These are some very general benefits of NLP. We will better understand its power and importance with examples.

NLP Use Cases in the Real World

Natural Language Processing is used in multiple ways. Here are some of them.

Machine translation

Everyone has used Google Translate at some point. But have you noticed how much it has improved over the last few years? Not only does it give you translations, but it also chooses the correct word form and offers synonyms when possible. Incredible, isn't it? Another interesting example of machine translation is the program TIDES. It is funded by the U.S. Defense Advanced Research Projects Agency, and its main goal is to automate text translation and summarization. The program is of great interest to the U.S. military in the Middle East. The focus is on understanding local languages. It will make the communication between military members and civilians as quick as possible.
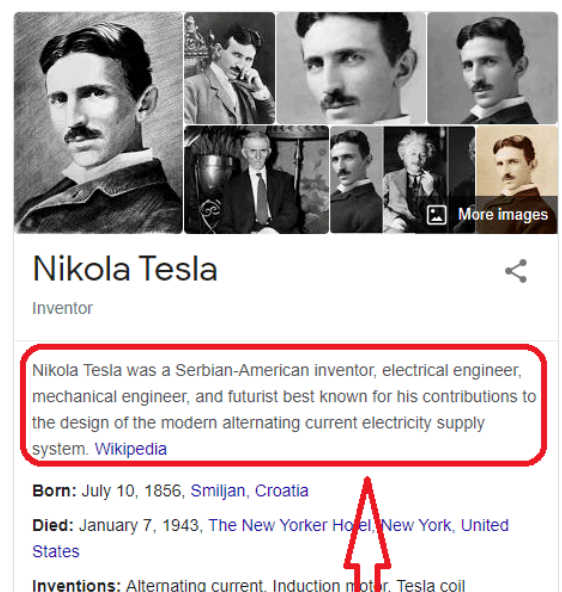
Application of NLP_(continue)
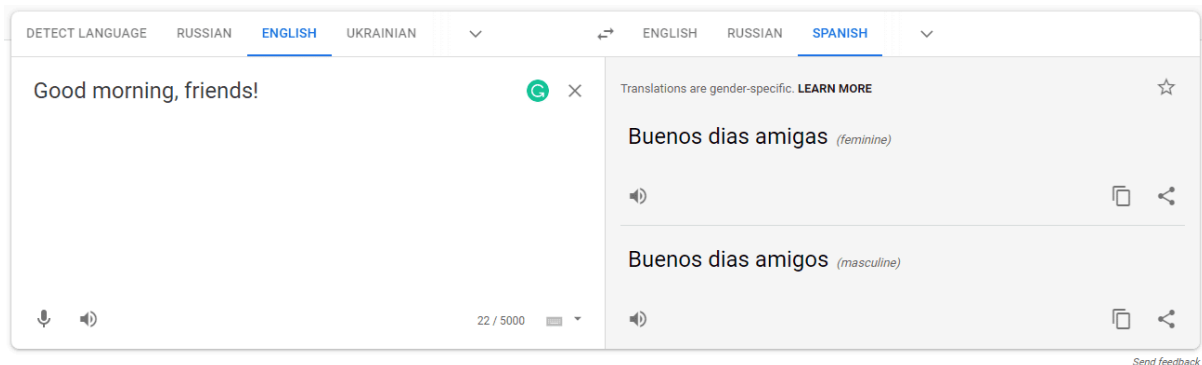
[Google Translate](#) can offer different translation options to choose from

Text summarization

Text summarization is a process of extracting the most important parts of the text, making it shorter and more explicit. Text summarization is extremely useful when there is no time or possibility to work with the entire text. Natural language processing algorithms will determine the most relevant phrases and sentences and present them as a summary of the text. We have all seen automatic text summarization in action, even if we did not realize it. One exciting application of text summarization is a Wikipedia article's description. Any time we enter our query, if there is a Wikipedia article about it, Google will show one or two sentences describing the entity we are looking for.

[Google uses text summarization](#) to show a brief description of the searched entity.

Another example would be news summarization. A typical American newspaper publishes a [few hundred](#) articles every day. There are [more than a thousand](#) such newspapers in the U.S., which yield hundreds of thousands of items daily. An enormous quantity, isn't it? Not a single human being can process such a massive amount of information. But machines can. And it is precisely NLP that makes it possible to analyze all of this news and extract the most important events.

Application of NLP_(continue)

Text classification

Text classification is used to assign an appropriate category to the text. As you may have seen, articles on news websites are often divided into categories. Such categorization is usually done automatically with the help of text classification algorithms.


[News categorization]() allows readers to find articles related to a specific topic quickly

You can also encounter text classification in product monitoring. Suppose you are a business owner, and you are interested in what people are saying about your product. In that case, you may use natural language processing to categorize the mentions you have found on the internet into specific categories. You may want to know what people are saying about the quality of the product, its price, your competitors, or how they would like the product to be improved.

Sentiment analysis

Sentiment analysis is a sub-task of text classification. Its primary purpose is to divide the text into two categories: positive and negative. For example, if the company wants to know its customers' opinions, they might collect a set of tweets and conduct sentiment analysis. Because there are [millions of tweets]() created every day, it would be physically impossible to manually perform such a task. Therefore, NLP is used.
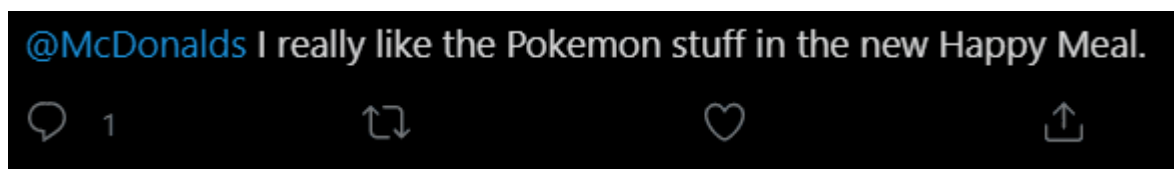
NLP algorithms would recognize such a positive comment and let the company know what its customers think
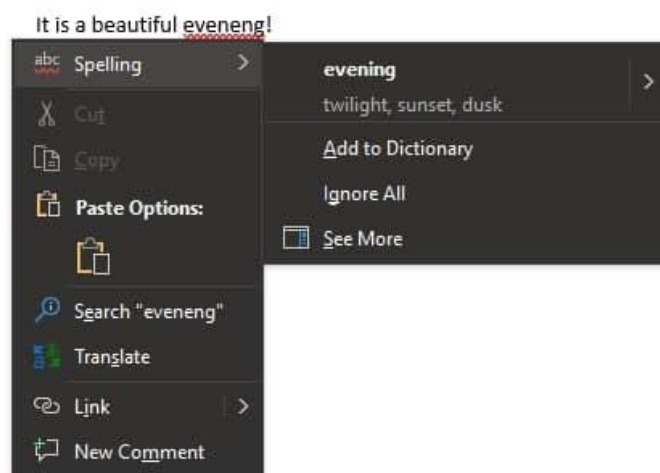
Application of NLP(continue)

Language modeling

Language modeling refers to predicting the probability of a sequence of words staying together. In layman's terms, language modeling tries to determine how likely it is that certain words stand nearby. This approach is handy in spelling correction, text summarization, handwriting analysis, machine translation, etc. Remember how Gmail or Google Docs offers you words to finish your sentence? This is language modeling in action.

Gmail provides the most common endings as you type

Another famous example is Microsoft Word's spelling checker. It calculates the probability of a word appearing in a sentence. It then gives you recommendations on correcting the word and improving the grammar.

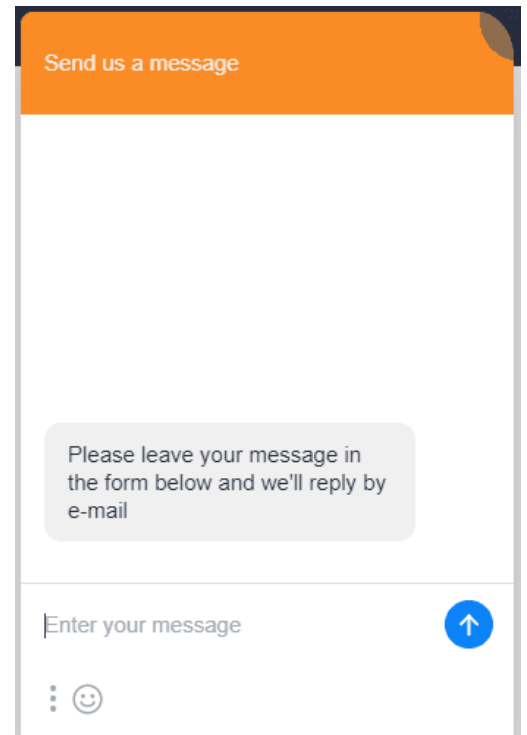Microsoft Word can find incorrect spelling and propose options to correct it

Application of NLP_(continue)

Question answering

Question answering is a subfield of NLP, which aims to answer human questions automatically. The most famous use case is chatbots. Many websites use them to answer basic customer questions, provide information, or collect feedback. It makes customer support much easier and quicker.

Chatbots make customer support quick and efficient and often help clients with their questions

Another fascinating application of question answering is robotics. Look at those incredible robots from Boston Dynamics. They are truly breathtaking, and they are becoming more and more complex every year. They can do many different things, like dancing, jumping, carrying heavy objects, etc. But can they be considered intelligent? According to the Turing test, a machine is deemed to be smart if, during a conversation, it cannot be distinguished from a human, and so far, several programs have successfully passed this test. ELIZA, PARRY, Eugene, and Cleverbot are some of the examples. All these programs use question answering techniques to make a conversation as close to human as possible. We can only hope that we will be able to talk to machines as equals in the future.

```
Welcome to
        EEEEEE  LL      IIII   ZZZZZZ   AAAAA
        EE      LL       II        ZZ  AA   AA
        EEEEE   LL       II       ZZZ  AAAAAAA
        EE      LL       II      ZZ    AA   AA
        EEEEEE  LLLLLL  IIII  ZZZZZZ    AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

ELIZA was the first natural language processing computer program to pass the Turing test

Application of NLP_(continue)

Video games

Video games are gaining more and more popularity. They are an essential aspect of our lives (at least, for some of us), and it is fascinating to watch the evolution of games caused by AI. In particular, natural language processing is used to generate unique conversations and create exceptional experiences. We are not limited to a preprogrammed set of phrases anymore. Our game may develop in any direction thanks to natural language processing.

Also, let's take AI Dungeon as an example. It is a text adventure game that uses natural language processing models to generate unlimited content. You can play infinitely long and always face different challenges and situations. All the characters, stories, conversations, and plots are unique. How cool is that?

> The next morning, you wake up to find that the city has been taken over by orcs. The orcs are led by a large orc named Orgrim Doomhammer.
>
> *> Hide under the bed*
>
> You quickly get under the bed, but not before hearing some noises from outside. You hear a couple of orcs talking about how they will take it from here.
>
> *> Start making annoying noises in hopes that they'll just go away*
>
> You make an irritating noise and then run off. They don't seem to notice you

AI Dungeon allows each player to create his/her individual scenario

This is not a complete list of NLP use cases; there are many more. However, these are the most widely known and commonly used applications, and they show how powerful and exciting natural language processing can be.

NLP Achievements

What has NLP Achieved So Far?

Natural language processing appeared 70 years ago when Alan Turing proposed his famous test as a machine intelligence criterion. It was believed at that time that machines would become extremely smart and powerful in the near future. Scientists claimed: such problems as machine translation would be solved within several years.

However, as we now know, these predictions did not come to life so quickly. Machines are still pretty limited. There is still much work to do and many difficulties to overcome. But it does not mean that natural language processing has not been evolving. On the contrary, it has achieved incredible heights. NLP was revolutionized by the development of neural networks in the last two decades, and we can now use it for tasks we could not even imagine before. Here are some of the most exciting things we can do with natural language processing:

Search engine

Google now processes more than 3.5 billion searches per day. An enormous number, isn't it? And it is rising steadily. But what is truly interesting is HOW Google does that. There are several advanced NLP models used to process all the queries. The most famous one is BERT. It was developed in 2018, and it is already helping the search better understand one in 10 searches in the U.S. only.

An example of how BERT improves the query's understanding is the search "2019 brazil traveler to usa need a visa". Earlier it was not clear to the computer whether it is a Brazilian citizen who is trying to get a visa to the U.S. or an American – to Brazil. The algorithm was just matching keywords. On the other hand, BERT takes into account every word in the sentence and can produce more accurate results. The word "to," in this case, tells the model a correct destination
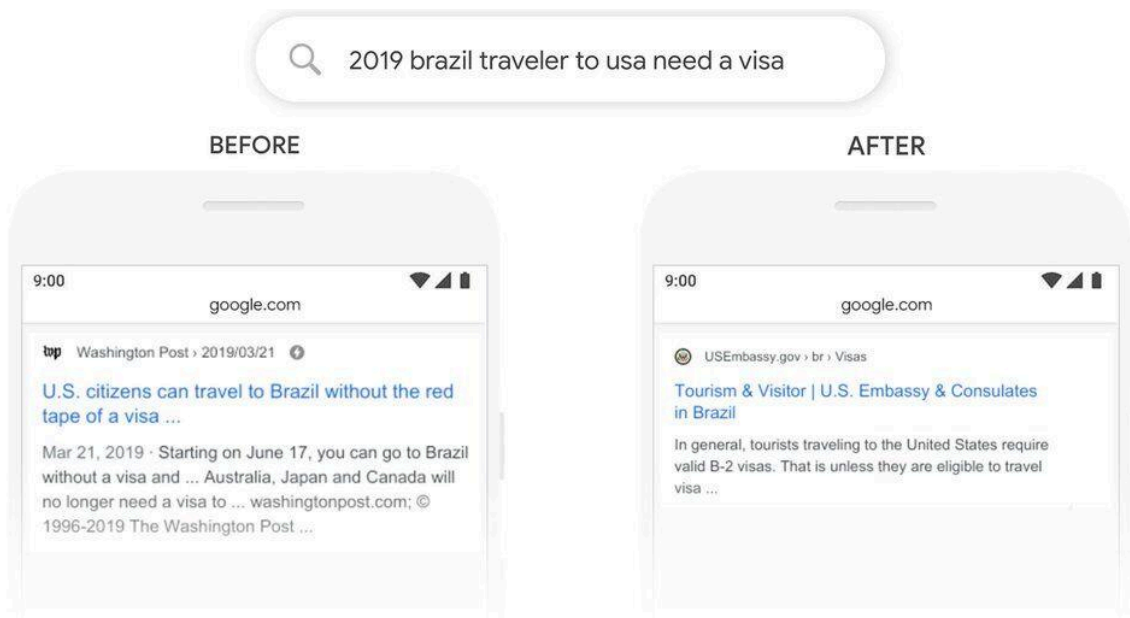
NLP Achievements(continue)

BERT can provide much more relevant results by analyzing the sentence as a whole instead of just matching keywords

Text generation

Natural language processing can be used not only for routine tasks but also for creative ones. Many believe that it is impossible to replicate human creativity and ingenuity, but the GPT-3 model from OpenAI has proven them wrong. GPT-3 can write code, propose use cases for random objects, create memes, and much more.

GPT-3 is capable of accomplishing a wide variety of tasks due to its vast amount of parameters that can be tuned for any problem.
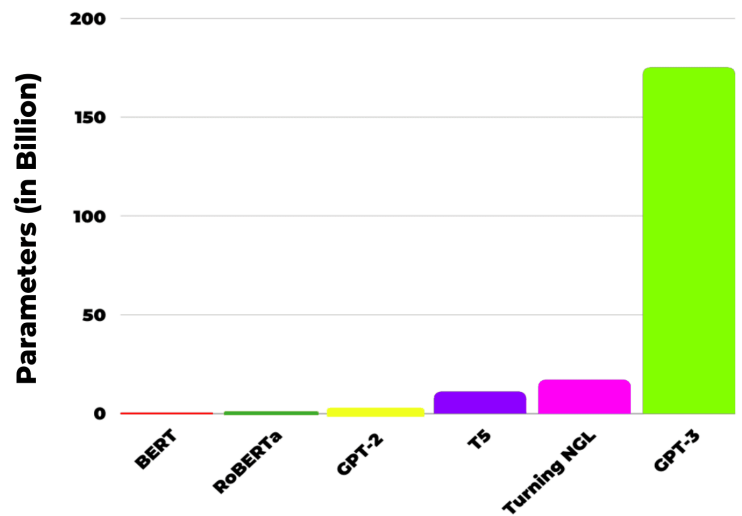
Text correction

There are many advanced tools for grammar checking besides the one used by Microsoft Word – for example, Grammarly. Apart from the usual spelling correction, it can check your style and tone. It can propose synonyms and word rearrangements, control for the level of formality, and even detect tautology.

Grammarly can check your grammar and suggest improvements

Virtual assistants

Siri, Amazon Alexa, Google Assistant – they are everywhere: at home, at work, and even on your phone. Virtual assistants may be useful and amusing, but they are also the result of NLP evolution. Virtual assistants are different from the above applications. They also use speech recognition – a remarkable feature that makes our lives easier.
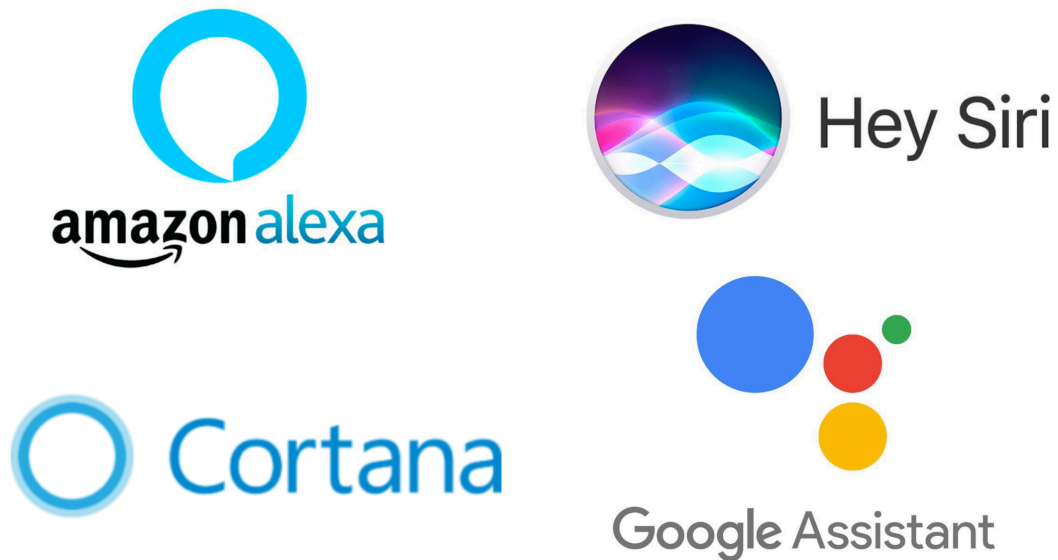
Amazon Alexa, Google Assistant, Siri, and Cortana – the four most popular virtual assistants

These are just a few examples that demonstrate the power of natural language processing. But it can go even further! So, what is it that lies ahead?

Future Of NLP

What is the Future of NLP?

The primary point of natural language processing is to make computers able to understand human language. NLP scientists will try to create models with even better performance and more capabilities.

We can probably expect these NLP models to be used by everyone and everywhere – from individuals to huge companies. Natural language processing is likely to be integrated into various tools and services, and the existing ones will only become better.

Let's take virtual assistants as an example. They are limited to a particular set of questions and topics and the moment. The smartest ones can search for an answer on the internet and reroute you to a corresponding website. However, virtual assistants get more and more data every day, and it is used for training and improvement. We can anticipate that programs such as Siri or Alexa will be able to have a full conversation, perhaps even including humor.

Another promising field is text generation. Nowadays, computers can already generate quite complex texts, even songs. But this is not the limit, and future models will be able to do more. Perhaps, write a poem or a fairy tale. Everything is possible!

Let's take one more example: automatic translation. Imagine how cool it would be to go abroad without worrying about the language. You will just use your phone to translate anything you need. And yes, it is already possible, but the accuracy will only increase!

What are the Challenges Natural Language Processing has to Overcome?

But, as it always is, some problems need to be solved to achieve all of that. They can be divided into two categories: issues related to data and problems related to the language itself.

NLP: Problems With Data

Low-resource languages

Some languages are very poorly presented, as they are not widely used. This concerns, first of all, African languages. There are thousands of them, but only a few people know them. And, of course, those people do not use any NLP technologies so far.
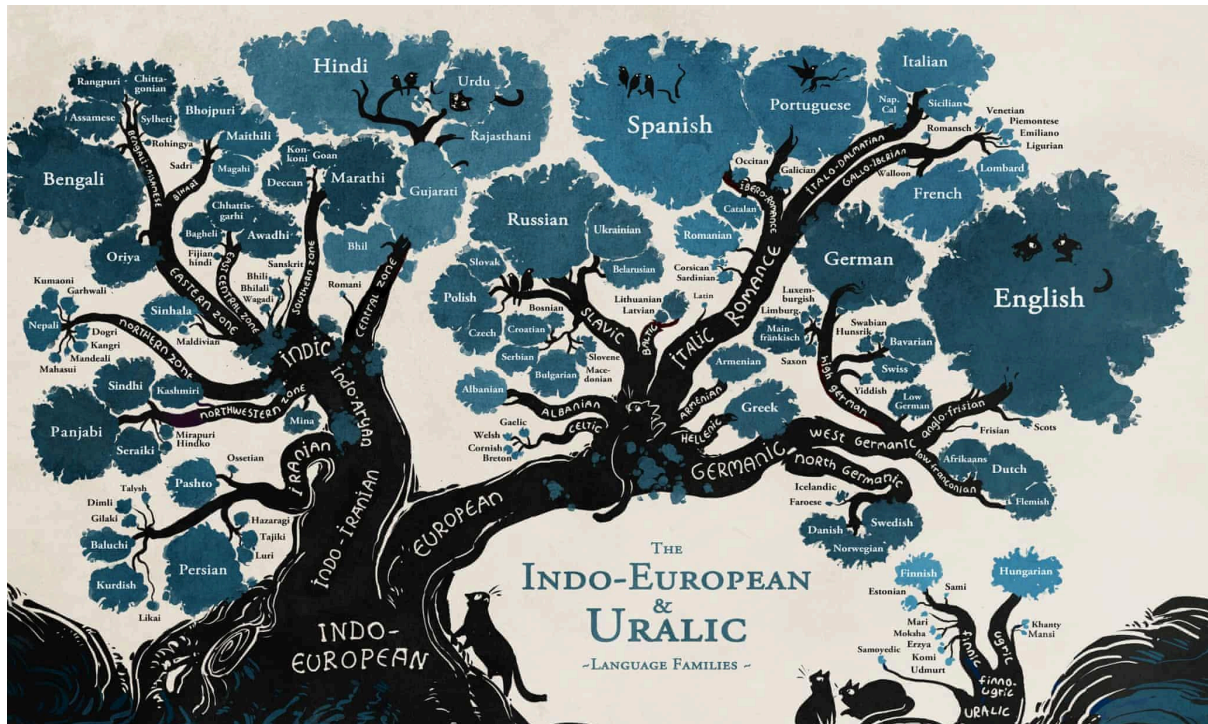
Young, H. (2015, January 23). A language family tree in pictures. *The Guardian.*
There are thousands of local languages in the world; many of them are used by a minimal number of people

Large amounts of data

The amount of data generated daily is genuinely enormous. To process and analyze it, we need time and power. Massive data centers are built for this purpose, but the best natural language processing models still train for weeks and take a lot of space, and these models are becoming more prominent all the time.

Low quality

A considerable portion of data generated is not of a particular quality. It needs some preprocessing, which takes time and money. It is, therefore, relatively expensive and time-consuming to work with text data.

Evaluation

Due to vast amounts of data, it is usually not labeled, making it challenging to test its performance.

Data labeling

Data labeling is one of the biggest challenges NLP data scientists face. Most of the text data generated are unlabeled. So it has to be done manually. In some cases, it is an easy task. For example, an entirely accurate sentiment analysis may be done by already existing models. But in others, it is complicated and time-consuming. An example is NER (or Named Entity Recognition), which is the task of extracting named entities (names, organizations, locations, etc.) from unstructured texts. It requires concentration and attention. Otherwise, essential data may be missed.

NLP: Language Challenges

The following are among the significant challenges NLP scientist have to deal with while working with a language:

- **Context**
  Same words may have different meanings depending on the context, while some words may have similar pronunciation but different meanings. This creates trouble for speech-to-text systems and, later, NLP models.
- **Synonyms**
  There are words with the same meanings, but some of them are very rarely used. Such terms may not be recognized as synonyms during the analysis.
- **Emotions**
  Human language has many complex mechanisms that are difficult for machines. For example, sarcasm or irony is not likely to be recognized by the natural language processing model in most cases.

This tweet obviously contains sarcasm, but the NLP model is more likely to detect it as a positive one

Domain-specific language

Vocabulary across industries may differ quite a lot. Thus, it is challenging to create a universal model. Usually, it is a better idea to train models on data related to one particular industry.

There are other issues, such as ambiguity and slang, that create similar challenges. The main point is that the human language is a very complex and diversified mechanism. It varies greatly across geographical regions, industries, ages, types of people, etc. It is, therefore, quite challenging to analyze a language as a whole.

At the moment, scientists can quite successfully analyze a part of a language concerning one area or industry. There is still a long way to go until we will have a universal tool that will work equally well with different languages and accomplish various tasks.

Conclusion

Natural language processing, while having its limitations, is still potent. It makes our lives much more comfortable and can bring significant benefits to businesses. There are many applications already in use. Among them: automatic translators, grammar checking tools, virtual assistants, etc. But it is anticipated that this number will only increase.

It will undoubtedly take some time, as there are multiple challenges to solve. But NLP is steadily developing, becoming more powerful every year, and expanding its capabilities.

Sources:
- ➔ turing.com
- ➔ chatGPT
- ➔ Data Science UA