Contents

1 Introduction

1.1 What is Natural Language Processing

Natural language processing (NLP) refers to the branch of **computer science** and more specifically, the branch of **artificial intelligence** or AI concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

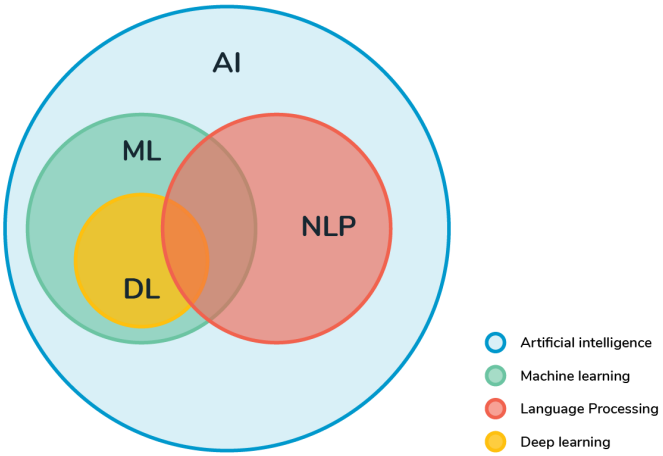

Figure 1: NLP is subfield of AI

NLP combines *computational linguistics* rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment.

1.2 Why NLP is so Important

In a world of Google and other search engines, shoppers expect to enter a phrase, or even an idea, into a search box and to instantly see personalized recommendations

that are clearly relevant to what it was they were meaning to discover. It's the sort of interaction that must go on at a speed and scale that can't be sustained by humans alone. Instead, doing right by consumers requires machines and systems

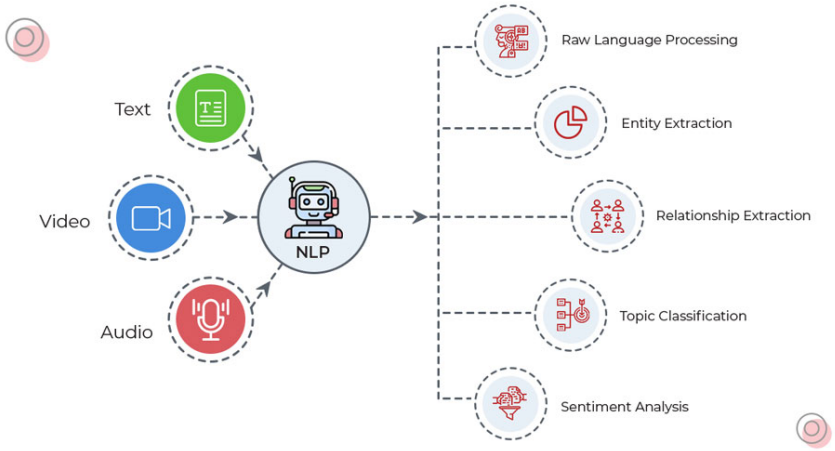

Figure 2: NLP functioning

that are constantly learning and developing insights into what customers mean and what they want.

1.3 Computational linguistics

It is the modern study of linguistics using the tools of computer science. Yesterday's linguistics may be today's computational linguist as the use of computational tools and thinking has overtaken most fields of study.

2 NLP Tasks

Human language is filled with ambiguities that make it incredibly difficult to write software that accurately determines the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammar and usage exceptions, variations in sentence structure—these just a few of the irregularities of human language that take humans years to learn, but that programmers must teach natural language-driven applications to recognize and understand accurately from the start, if those applications are going to be useful.

Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the following:

2.1 Speech recognition (aka speech-to-text)

It is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions. What makes speech recognition especially challenging is the way people talk—quickly, slurring words together, with varying emphasis and intonation, in different accents, and often using incorrect grammar.

2.2 Part of speech tagging

It is the process of determining the part of speech of a particular word or piece of text based on its use and context. Part of speech identifies 'make' as a verb in 'I can make a paper plane,' and as a noun in 'What make of car do you own?'

2.3 Word sense disambiguation

It is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that determine the word that makes the most sense in the given

context. For example, word sense disambiguation helps distinguish the meaning of the verb 'make' in 'make the grade' (achieve) vs. 'make a bet' (place).

2.4 Sentiment analysis

Some attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text. Basically understanding human emotions.

2.5 Natural language generation

It is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language.

3 NLP use cases

Natural language processing is the driving force behind machine intelligence in many modern real-world applications. Here are a few examples:

3.1 Spam detection

You may not think of spam detection as an NLP solution, but the best spam detection technologies use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more. Spam detection is one of a handful of NLP problems that experts consider 'mostly solved' (although you may argue that this doesn't match your email experience).

3.2 Machine translation

Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another. Effective translation has to capture accurately the meaning and tone of the input language and translate it to text with the same meaning and desired impact in the output language. Machine translation tools are making good progress in terms of accuracy. A great way to test any machine translation tool is to translate text to one language and then back to the original.

3.3 Social media sentiment analysis

NLP has become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

3.4 Text summarization

Text summarization uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text. The best text summarization applications use semantic reasoning and natural language generation (NLG) to add useful context and conclusions to summaries.

3.5 Virtual assistants and chatbots

Virtual assistants such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments. Chatbots perform the same magic in response to typed text entries. The best of these also learn to recognize contextual clues about human requests and use them to provide even better responses or options over time.

4 NLP tools and approaches

With the help of modern computer science tools and technology. NLP can be done easily. Lot of high level library and framework are available publically to be used.

4.1 Python and the Natural Language Toolkit (NLTK)

The Python programing language provides a wide range of tools and libraries for attacking specific NLP tasks. Many of these are found in the Natural Language Toolkit, or NLTK, an open source collection of libraries, programs, and education resources for building NLP programs. The NLTK includes libraries for many of the NLP tasks listed above, plus libraries for subtasks, such as sentence parsing, word segmentation, stemming and lemmatization (methods of trimming words down to their roots), and tokenization (for breaking phrases, sentences, paragraphs and passages into tokens that help the computer better understand the text). It also includes libraries for implementing capabilities such as semantic reasoning, the ability to reach logical conclusions based on facts extracted from text.

4.2 Statistical NLP, machine learning, and deep learning

The earliest NLP applications were hand-coded, rules-based systems that could perform certain NLP tasks, but couldn't easily scale to accommodate a seemingly endless stream of exceptions or the increasing volumes of text and voice data. Enter statistical NLP, which combines computer algorithms with machine learning and deep learning models to automatically extract, classify, and label elements of text and voice data and then assign a statistical likelihood to each possible meaning of those elements. Today, deep learning models and learning techniques based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) enable NLP systems that 'learn' as they work and extract ever more accurate meaning from huge volumes of raw, unstructured, and unlabeled text and voice data sets.

5 NLP Challenges

NLP is a powerful tool with huge benefits, but there are still a number of Natural Language Processing limitations and problems:

5.1 Contextual words and phrases and homonyms

The same words and phrases can have different meanings according the context of a sentence and many words – especially in English – have the exact same pronunciation but totally different meanings. For example: I ran to the store because we ran out of milk. Can I run something past you real quick? The house is looking really run down. These are easy for humans to understand because we read the context of the sentence and we understand all of the different definitions. And, while NLP language models may have learned all of the definitions, differentiating between them in context can present problems. Homonyms – two or more words that are pronounced the same but have different definitions – can be problematic for question answering and speech-to-text applications because they aren't written in text form. Usage of their and there, for example, is even a common problem for humans.

5.2 Synonyms

Synonyms can lead to issues similar to contextual understanding because we use many different words to express the same idea. Furthermore, some of these words may convey exactly the same meaning, while some may be levels of complexity (small, little, tiny, minute) and different people use synonyms to denote slightly different meanings within their personal vocabulary.

5.3 Irony and sarcasm

Irony and sarcasm present problems for machine learning models because they generally use words and phrases that, strictly by definition, may be positive or negative,

but actually connote the opposite. Models can be trained with certain cues that frequently accompany ironic or sarcastic phrases, like *"yeah right," "whatever,"* etc., and word embeddings (where words that have the same meaning have a similar representation), but it's still a tricky process.

5.4 Ambiguity

Even for humans this sentence alone is difficult to interpret without the context of surrounding text. POS (part of speech) tagging is one NLP solution that can help solve the problem, somewhat. Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.

5.4.1 Lexical ambiguity

A word that could be used as a verb, noun, or adjective.

5.4.2 Semantic ambiguity

The interpretation of a sentence in context. For example: I saw the boy on the beach with my binoculars. This could mean that I

5.4.3 Syntactic ambiguity

In the sentence above, this is what creates the confusion of meaning. The phrase with my binoculars could modify the verb, "saw," or the noun, "boy."

5.5 Errors in text and speech

Misspelled or misused words can create problems for text analysis. Autocorrect and grammar correction applications can handle common mistakes, but don't always understand the writer's intention. With spoken language, mispronunciations, different accents, stutters, etc., can be difficult for a machine to understand. However, as

language databases grow and smart assistants are trained by their individual users, these issues can be minimized.

5.6 Colloquialisms and slang

Informal phrases, expressions, idioms, and culture-specific lingo present a number of problems for NLP – especially for models intended for broad use. Because as formal language, colloquialisms may have no "dictionary definition" at all, and these expressions may even have different meanings in different geographic areas. Furthermore, cultural slang is constantly morphing and expanding, so new words pop up every day. This is where training and regularly updating custom models can be helpful, although it oftentimes requires quite a lot of data.

5.7 Domain-specific language

Different businesses and industries often use very different language. An NLP processing model needed for healthcare, for example, would be very different than one used to process legal documents. These days, however, there are a number of analysis tools trained for specific fields, but extremely niche industries may need to build or train their own models.

5.8 Low-resource languages

AI machine learning NLP applications have been largely built for the most common, widely used languages. And it's downright amazing at how accurate translation systems have become. However, many languages, especially those spoken by people with less access to technology often go overlooked and under processed. For example, by some estimations, (depending on language vs. dialect) there are over 3,000 languages in Africa, alone. There simply isn't very much data on many of these languages.

5.9 Lack of research and development

Machine learning requires A LOT of data to function to its outer limits – billions of pieces of training data. The more data NLP models are trained on, the smarter they become. That said, data (and human language!) is only growing by the day, as are new machine learning techniques and custom algorithms. All of the problems above will require more research and new techniques in order to improve on them. Advanced practices like artificial neural networks and deep learning allow a multitude of NLP techniques, algorithms, and models to work progressively, much like the human mind does. As they grow and strengthen, we may have solutions to some of these challenges in the near future.

6 Products based on NLP

6.1 Amazon's Alexa

Figure 3: Variant alexa products by amazon

Alexa is Amazon's all-knowing, interactive voice assistant. Available on Amazon's lineup of Echo speakers, smart thermostats, soundbars, lamps and lights, and right on your phone through the Alexa app, Alexa can do quick math for you, launch your favorite playlists, check news and weather, and control many of your home's smart products. In this guide, we explain where Alexa comes from, exactly how Alexa works, where Alexa gets her name, and more.

6.2 Apple's Siri

Siri, Apple's personal digital assistant, uses machine learning and natural speech to answer questions, return relevant search information, perform actions and more.

Figure 4: Siri running on iPhone

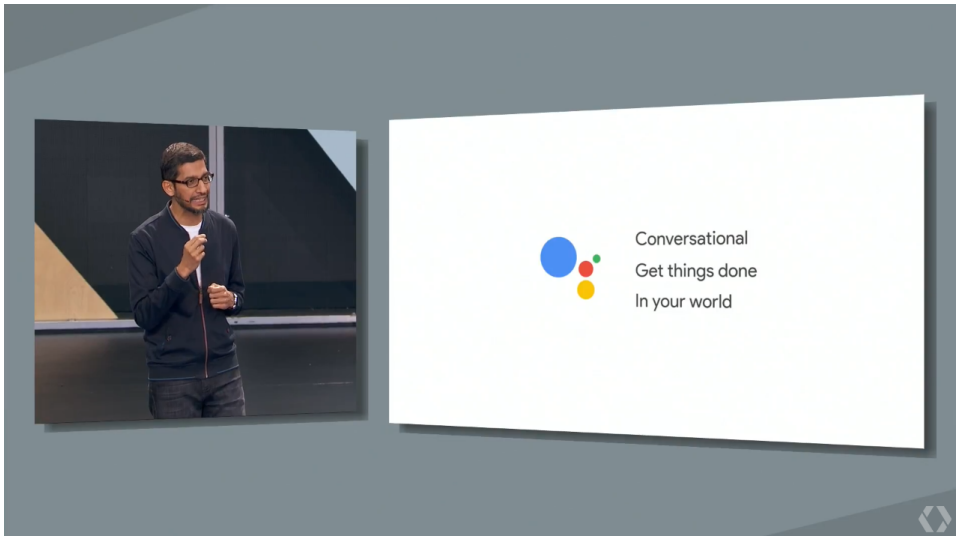

Figure 5: Sundar Pichai (CEO of google) introducing google assistant

6.3 Google voice assistant

Google Assistant offers voice commands, voice searching, and voice-activated device control, letting you complete a number of tasks after you've said the "OK Google" or "Hey Google" wake words. It is designed to give you conversational interactions. Google Assistant will: Control your devices and your smart home

7 Lojban

Lojban is a carefully constructed spoken language. It has been built for over 50 years by dozens of workers and hundreds of supporters.

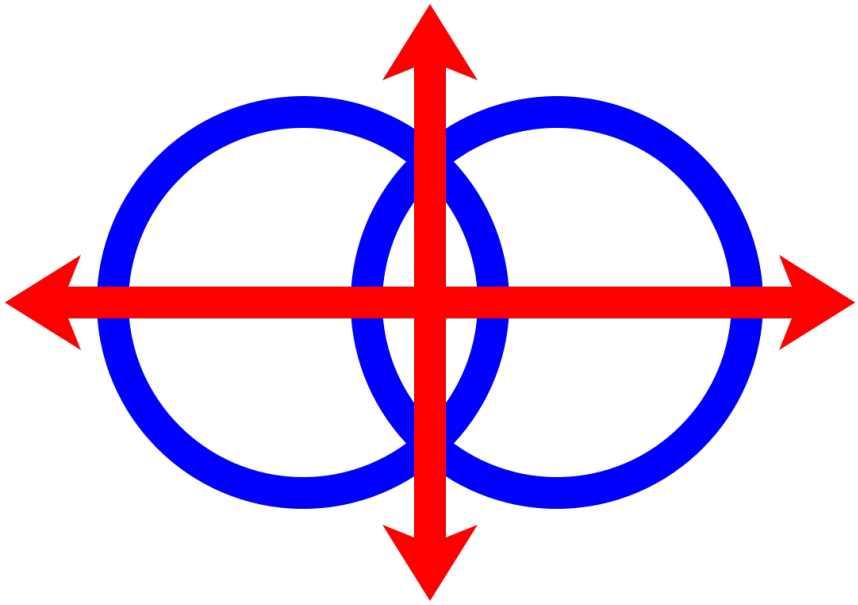

Figure 6: Lojban logo

- Lojban's grammar is based on simple rules, and its linguistic features are inspired by predicate logic.
- Lojban allows the expression of nuances in emotion using words called attitudinals, which are like spoken emoticons. ue marks that you're surprised; ba'u marks that you're exaggerating.
- You can be as vague or detailed as you like when speaking lojban. For example, specifying tense (past, present or future) or number (singular or plural) is optional when they're clear from context.
- Lojban is machine parsable, so the syntactic structure and validity of a sentence

is unambiguous, and can be analyzed using computer tools.

- There is a live community of speakers expanding the lojban vocabulary day by day.

7.1 Lojban means different things to different people

- a linguistic curiosity - a test-bed for language experimentation
- a challenging way to expand their minds or discipline their thoughts
- a new perspective on languages
- an entertaining medium to communicate with friends or create art
- a domain for exploring the intersection of human language and software

8 Conclusion

9 References

- https://en.wikipedia.org/wiki/Natural-language-processing
- https://www.ibm.com/cloud/learn/natural-language-processing
- https://machinelearningmastery.com/natural-language-processing
- https://mw.lojban.org/index.php?title=Lojban&setlang=en-US
- https://machinelearningmastery.com/natural-language-processing