

CHALLENGES IN NLP

By.Priyanshu Kamble

Artificial intelligence has become part of our everyday lives – Alexa and Siri, text and email autocorrect, customer service chatbots. They all use machine learning algorithms and Natural Language Processing (NLP) to process, “understand”, and respond to human language, both written and spoken.

Give this NLP sentiment analyzer a spin to see how NLP automatically understands and analyzes sentiments in text (Positive, Neutral, Negative).

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

- Contextual words and phrases and homonyms.
- Synonyms.
- Irony and sarcasm.
- Ambiguity.
- Errors in text or speech.
- Colloquialisms and slang.
- Domain-specific language.
- Low-resource languages.
- Lack of research and development.
- Contextual words and phrases and homonyms

The same words and phrases can have different meanings according to 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.

- 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.

So, for building NLP systems, it's important to include all of a word's possible meanings and all possible synonyms. Text analysis models may still occasionally make mistakes, but the more relevant training data they receive, the better they will be able to understand synonyms.

- 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.

- Ambiguity

Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.

- Lexical ambiguity

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

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 saw a boy through my binoculars or the boy had my binoculars with him

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.”

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.

- 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.

- 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.

- 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.

- 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.

- 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.