What Is Natural Language Processing?

by Jason Brownlee

Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

The study of natural language processing has been around for more than 50 years and grew out of the field of linguistics with the rise of computers.

In this post, you will discover what natural language processing is and why it is so important.

After reading this post, you will know:

- What natural language is and how it is different from other types of data.
- What makes working with natural language so challenging.
- Where the field of NLP came from and how it is defined by modern practitioners.

Natural Language

Natural language refers to the way we, humans, communicate with each other.

Namely, speech and text.

We are surrounded by text.
Think about how much text you see each day:

- Signs
- Menus
- Email
- SMS
- Web Pages
- and so much more...

The list is endless.

Now think about speech.

We may speak to each other, as a species, more than we write. It may even be easier to learn to speak than to write.

Voice and text are how we communicate with each other.

Given the importance of this type of data, we must have methods to understand and reason about natural language, just like we do for other types of data.

**Challenge of Natural Language**

Working with natural language data is not solved.

It has been studied for half a century, and it is really hard.

*It is hard from the standpoint of the child, who must spend many years acquiring a language … it is hard for the adult language learner, it is hard for the scientist who attempts to model the relevant phenomena, and it is hard for the engineer who attempts to build systems that deal with natural language input or output. These tasks are so hard that Turing could rightly make fluent conversation in natural language the centerpiece of his test for intelligence.*

Natural language is primarily hard because it is messy. There are few rules.

And yet we can easily understand each other most of the time.

*Human language is highly ambiguous … It is also ever changing and evolving. People are great at producing language and understanding language, and are capable of expressing, perceiving, and interpreting very elaborate and nuanced meanings. At the same time, while we humans are great users of language, we are also very poor at formally understanding and describing the rules that govern language.*
From Linguistics to Natural Language Processing

Linguistics

Linguistics is the scientific study of language, including its grammar, semantics, and phonetics. Classical linguistics involved devising and evaluating rules of language. Great progress was made on formal methods for syntax and semantics, but for the most part, the interesting problems in natural language understanding resist clean mathematical formalisms.

Broadly, a linguist is anyone who studies language, but perhaps more colloquially, a self-defining linguist may be more focused on being out in the field.

Mathematics is the tool of science. Mathematicians working on natural language may refer to their study as mathematical linguistics, focusing exclusively on the use of discrete mathematical formalisms and theory for natural language (e.g. formal languages and automata theory).

Computational Linguistics

Computational linguistics 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. Computational linguistics is the study of computer systems for understanding and generating natural language. … One natural function for computational linguistics would be the testing of grammars proposed by theoretical linguists.

Large data and fast computers mean that new and different things can be discovered from large datasets of text by writing and running software.

In the 1990s, statistical methods and statistical machine learning began to and eventually replaced the classical top-down rule-based approaches to language, primarily because of their better results, speed, and robustness. The statistical approach to studying natural language now dominates the field; it may define the field.

Data-Drive methods for natural language processing have now become so popular that they must be considered mainstream approaches to computational linguistics. … A strong contributing factor to this development is undoubtedly the increase amount of available electronically stored data to which these methods can be applied; another
factor might be a certain disenchantment with approaches relying exclusively on hand-crafted rules, due to their observed brittleness.

The statistical approach to natural language is not limited to statistics per-se, but also to advanced inference methods like those used in applied machine learning.

… understanding natural language require large amounts of knowledge about morphology, syntax, semantics and pragmatics as well as general knowledge about the world. Acquiring and encoding all of this knowledge is one of the fundamental impediments to developing effective and robust language systems. Like the statistical methods … machine learning methods off the promise of automatic the acquisition of this knowledge from annotated or unannotated language corpora.

Statistical Natural Language Processing

Computational linguistics also became known by the name of natural language process, or NLP, to reflect the more engineer-based or empirical approach of the statistical methods.

The statistical dominance of the field also often leads to NLP being described as Statistical Natural Language Processing, perhaps to distance it from the classical computational linguistics methods.

I view computational linguistics as having both a scientific and an engineering side. The engineering side of computational linguistics, often called natural language processing (NLP), is largely concerned with building computational tools that do useful things with language, e.g., machine translation, summarization, question-answering, etc. Like any engineering discipline, natural language processing draws on a variety of different scientific disciplines.

Linguistics is a large topic of study, and, although the statistical approach to NLP has shown great success in some areas, there is still room and great benefit from the classical top-down methods.

Roughly speaking, statistical NLP associates probabilities with the alternatives encountered in the course of analyzing an utterance or a text and accepts the most probable outcome as the correct one. … Not surprisingly, words that name phenomena that are closely related in the world, or our perception of it, frequently occur close to one another so that crisp facts about the world are reflected in somewhat fuzzier facts about texts. There is much room for debate in this view.
Natural Language Processing

As machine learning practitioners interested in working with text data, we are concerned with the tools and methods from the field of Natural Language Processing.

We have seen the path from linguistics to NLP in the previous section. Now, let’s take a look at how modern researchers and practitioners define what NLP is all about.

In perhaps one of the more widely textbooks written by top researchers in the field, they refer to the subject as “linguistic science,” permitting discussion of both classical linguistics and modern statistical methods. 

The aim of a linguistic science is to be able to characterize and explain the multitude of linguistic observations circling around us, in conversations, writing, and other media. Part of that has to do with the cognitive size of how humans acquire, produce and understand language, part of it has to do with understanding the relationship between linguistic utterances and the world, and part of it has to do with understand the linguistic structures by which language communicates.

They go on to focus on inference through the use of statistical methods in natural language processing.

Statistical NLP aims to do statistical inference for the field of natural language. Statistical inference in general consists of taking some data (generated in accordance with some unknown probability distribution) and then making some inference about this distribution.

In their text on applied natural language processing, the authors and contributors to the popular NLTK Python library for NLP describe the field broadly as using computers to work with natural language data. We will take Natural Language Processing — or NLP for short — in a wide sense to cover any kind of computer manipulation of natural language. At one extreme, it could be as simple as counting word frequencies to compare different writing styles. At the other extreme, NLP involves “understanding” complete human utterances, at least to the extent of being able to give useful responses to them.

Statistical NLP has turned another corner and is now strongly focused on the use of deep learning neural networks to both perform inference on specific tasks and for developing robust end-to-end systems.
In one of the first textbooks dedicated to this emerging topic, Yoav Goldberg succinctly defines NLP as automatic methods that take natural language as input or produce natural language as output.

*Natural language processing (NLP) is a collective term referring to automatic computational processing of human languages. This includes both algorithms that take human-produced text as input, and algorithms that produce natural looking text as outputs.*