




# Parabole

Natural language processing: Brute force or intelligent design

MD. ISHAQ



## Contents

Introduction .....	2
Breaking the sentence .....	3
Tagging the parts of speech (POS) and generating dependency graphs .....	4
Building the appropriate vocabulary .....	4
Linking different components of vocabulary .....	5
Clustering .....	5
Setting the context.....	6
Extracting semantic meanings .....	6
Extracting named entities (often referred to as Named Entity Recognition = NER) .....	7
Data Element Extraction .....	8
Conclusion.....	8
About us.....	9

## Introduction

In the early 1970's, the ability to perform complex calculations was placed in the palm of people's hands. The invention of the hand-held calculator allowed for significant time savings as these small devices "understood" numbers and evolved to possess the capability of performing more advanced calculations in Mathematics, Calculus and Geometry and Statistics. We see a similar path for highly developed NLP solutions for language and Parable is on a mission to lead the industry.

Natural-language processing (NLP) is a field of computer science and artificial intelligence being developed and used in the analysis of interactions between computers and humans, in particular the development of methods to program computers to successfully process large volumes of natural language data. Once considered fantastical, NLP in its various forms is now utilized in many facets of business, entertainment and social media and is being used increasingly, albeit with varying efficiency, within machine learning systems. Machines may be trained using a variety of complex and unstructured sources assembled under myriad contexts. In this age, where robots are engaged in conversations on a near-human level, NLP is increasingly automating operational processes ranging from the simple, like answering a question from the web, to the complex, such as processing gigabytes of unstructured data and calculating that data's contexts, generating terminologies and making implicit connections.

Interestingly, NLP works in a quite human way. When we speak to each other, usually the context or setting within which a conversation takes place is understood by both parties, and therefore the conversation is easily interpreted. There are, however, those moments where one of the participants may fail to properly explain an idea, conversely, the listener (the receiver of the information), may fail to understand the context of the conversation for any number of reasons. Similarly, machines can fail to comprehend the context of text unless properly and carefully trained.

For humans, learning in early childhood occurs in a consistent way; children interact with unstructured data and process that data into information. After amassing this information, we begin to analyze information to understand its implications in a given situation or the nuance of a given problem. We understand that at a certain point, we have a learned understanding of our life and environment.

After understanding implications, information can be used to solve a set of problems or life situations. Humans iterate through multiple scenarios to consciously or unconsciously

simulate whether a solution will be a success or failure. After practice with this unstructured data -> information -> knowledge, humans, hopefully, gain what we refer to as wisdom.

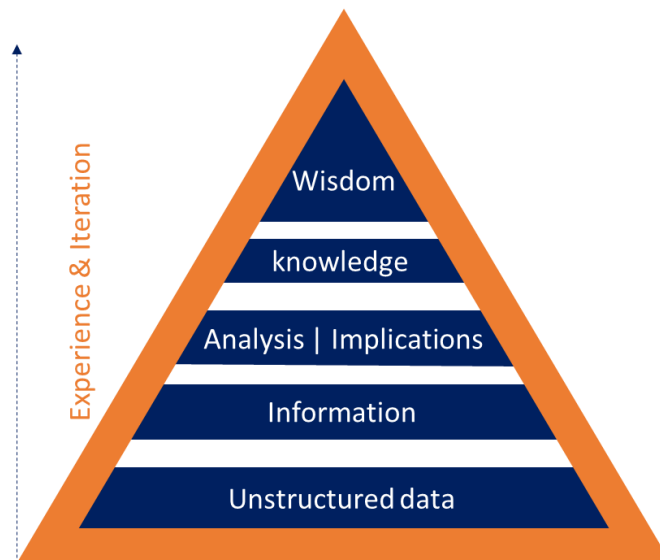


Figure 1: Information maturity

Cognitive machines learn by a similar method; initially, the machine translates unstructured textual data into meaningful terms, then identifies connections between those terms, and finally comprehends context. Many technologies conspire to process natural languages, the most popular of which are Stanford Core NLP, Spacy, and Apache NLTK, amongst others. Cognition involves machine recognition of a natural language, however, to be valuable, several challenges must be overcome, especially when the data within a system lacks consistency. While this inconsistency actually allows the machine to capture variety and subjectivity, it is not part of the initial phase of machine learning.

## Breaking the sentence

Formally referred to as sentence boundary disambiguation (SBD), this breaking process is no longer difficult to achieve, but is nonetheless, a critical process, especially in the case of highly unstructured data. A breaking application should be intelligent enough to separate paragraphs into their appropriate sentence units, however, highly complex data might not always be available in easily recognizable sentence forms. This data may exist in the form of tables, graphics, notations, page breaks, etc., which need to be appropriately processed for the machine to derive meanings in the same way a human would approach interpreting text.

## Tagging the parts of speech (POS) and generating dependency graphs

People understand, to a greater or lesser degree, their native language; there is no need, other than for the formal study of that language, to further understand the individual parts of speech in a conversation or reading, as these have been learned in the past. In order for a machine to learn, it must understand formally, the fit of each word, i.e., how the word positions itself into the sentence, paragraph, document or corpus. In general, NLP applications employ a set of POS (parts of speech) tagging tools that assign a POS tag to each word or symbol in a given text. Subsequently, the position of each word in a sentence is determined by a dependency graph, generated in the same procedure. Those POS tags can be further processed to create meaningful single or compound vocabulary terms.

### Building the appropriate vocabulary

Using these POS tags and dependency graphs, a powerful vocabulary can be generated and subsequently interpreted by the machine in a way comparable to human understanding.

Consider the following paragraph:

“All employees are responsible for the management of risk, with the ultimate accountability residing with the Board. We have a strong risk culture, which is embedded through clear and consistent communication and appropriate training for all employees. A comprehensive risk management framework is applied throughout the group, with governance and corresponding risk management tools. This framework is underpinned by our risk culture and reinforced by the HSBC Values.” -HSBC annual report 2017

Sentences are generally simple enough to be parsed by a basic NLP program. But to be of real value, an algorithm should also be able to generate, at a minimum, the following vocabulary terms:

- > Employees
- > Management of risk
- > Ultimate accountability
- > Board
- > Strong risk culture
- > Clear and consistent communication
- > Appropriate training for all employees
- > Comprehensive risk management framework
- > Governance and corresponding risk management tools
- > Framework

### > Risk culture

Unfortunately, most NLP software applications do not result in creating a sophisticated set of vocabulary. At Parable, we have developed new processes and algorithms to elevate the value of NLP. The value can be directly seen in our Cognitive Text Analytics product - Jana™. Jana™ has raised the bar in document analytics by adding capabilities to further refine and improve the output for SME's in the banking, finance and insurance sectors.

### Linking different components of vocabulary

Recently, new approaches have been developed that can execute the extraction of the linkage between any two vocabulary terms generated from a document (or "corpus"). WordSpace, a vector-space based model, assigns vectors to each word in a corpus, those vectors ultimately capture each word's relationship to closely occurring words or set of words. But statistical methods like WordSpace are not sufficient to capture either the linguistics or the semantic relationships between pairs of vocabulary terms. In the above-stated example, "All employees are responsible for the management of risk, with the ultimate accountability residing with the Board", two vocabulary terms, "Board" and "management of risk" are connected with the Board having ultimate accountability, but since these two terms are statistically distant, the extent of the relationship bond between this pair cannot be ascertained, neither linguistically nor semantically. A more sophisticated algorithm will be needed to capture the relationship bonds that exist between coupled words.

## Clustering

Statistical approaches like Latent Dirichlet allocation (LDA) can model the document into various topics, however, this method requires an enormous corpus to properly train a machine. Others, including but not limited to; Brown Clustering, Naïve Bayesian Classification, TensorFlow, etc. can classify different vocabulary terms through supervised learning. But each of these systems require an enormous amount of training so that machines may learn the various classifications and regressions needed to provide a desired result.

## Setting the context

One of the most important and challenging tasks in the entire NLP process is to train a machine to derive context from a discussion within a document. Consider the following two sentences:

“I enjoy working in a bank.”

“I enjoy working near a river bank.”

The context of these sentences is quite different. There are several methods today to train a machine to understand the differences between sentences. One established method is to build a knowledge graph where both possibilities would occur based upon statistical calculations, when a new document is under observation, the machine would refer to the knowledge graph to determine the setting before proceeding. One challenge in building the knowledge graph is domain specificity. Knowledge graphs cannot, in a practical sense, be made to be universal. In the example above “enjoy working in a bank” suggests “work, or job, or profession”, while “enjoy near a river bank” is just any type of work or activity that can be performed near a river bank. Two sentences with totally different contexts in different domains might confuse the machine if forced to rely solely on knowledge graphs. It is critical to enhance the methods used with a probabilistic approach in order to derive context and proper domain choice.

## Extracting semantic meanings

Linguistic analysis of vocabulary terms might not be enough for a machine to correctly apply learned knowledge. To successfully apply learning, a machine must understand further, the semantics of every vocabulary term within the context of the documents. By way of example, consider two sentences:

“Under US GAAP, gains and losses from AFS assets are included in net income.”

“Under IFRS, gains and losses from AFS assets are included in comprehensive income.”

Both sentences have the context of gains and losses in proximity to some form of income, but the resultant information needed to be understood is entirely different between these sentences due to differing semantics. It is a combination, encompassing both linguistic and semantic methodologies and built within Parabolé’s cognitive analytics that allows the machine to truly understand the meanings within a selected text.

Extracting named entities (often referred to as Named Entity Recognition = NER)

The next big challenge is to successfully execute NER, which is essential when training a machine learning engine to distinguish between simple vocabulary and named entities. In many instances, these entities are surrounded by dollar amounts, places, locations, numbers, time, etc., it is critical to make and express the connections between each of these elements, only then may a machine fully interpret a given text.



## Data Element Extraction

“The recent developments in technology have enabled the stock price of Apple to rise by 20% to \$168 as at Feb 20, 2018 from \$140 in Q3 2017.”

Think of this sentence broken down into the following structure:

Table 1: Data presentation

	Context					Data		
Data Element	Date	Named Entity	Classification	Process   Event   System   Component	Action	%age	Number	Money
Stock Price	Feb 20, 2018	Apple	Investment	Recent developments in technology	Rise	[by] 20%		[to] \$168
Stock Price	Q3'2017	Apple	Investment	Recent developments in technology	Rise			[from] \$140

This breakdown would be extremely difficult to achieve through linguistics as sentences possess unique structures and authors write with varying styles. Linguistics is, however, valuable as an initial approach to extract data elements from documents as a precursor to the work done within the semantic layer. This layer understands the relationship between data elements and their values and surroundings have to be machine-trained as well in order to suggest a modular output in a given format.

We, at Parabole have been carefully analyzing the challenges for NLP like those mentioned in this paper. Parabole’s engineers, developers, and ontologist’s efforts are focused on the production of the highest quality output that augments the textual analysis of any reader, allowing decisions to be made more quickly and efficiently using NLP. We invite you to see our Cognitive Text Analytics at <http://www.parabole.ai>

## Conclusion

In this article we endeavor to provide the reader with a brief introduction to automated ontology and how it tackles the challenge of enterprise scale and its ‘applicability within mainstream business use cases. Even though it is impossible to fully describe each of the different methods and algorithms thoroughly, it should provide an overview of the current progress in the field of automated ontology creation.

## About us

Parabole is a Princeton, NJ based cognitive analytics company, automating the creation of enterprise knowledge from unstructured sources of information. We provide our clients with platform to solve risk, finance and regulatory compliance-related challenges that depend on the knowledge-data interchange. We accomplish this by delivering a range of bespoke applications in the areas of credit, market, liquidity risk and data governance domains.

To learn more, visit [www.parabole.ai](http://www.parabole.ai) OR reach out to [info@mindparabole.com](mailto:info@mindparabole.com)