African Institute for Mathematical Sciences
South Africa

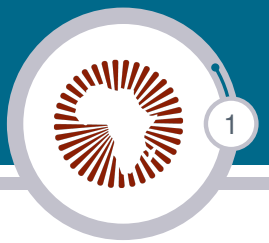
Natural Language Processing

Machine Learning has revolutionized Natural Language Processing

Salomon Kabongo
salomon@aims.ac.za

December 31, 2018

Content



Introduction
What is Natural Language Processing

Computers have long been able to defeat even the best human chess player, but are only recently matching some of the abilities of average human beings to recognize objects or speech.
The true challenge to artificial intelligence proved to be solving the tasks that are **easy** for people to perform but hard for people to **describe** formally—problems that we solve intuitively, that feel automatic, like **recognizing spoken words** or **faces in images** [4].

Definition

Natural Language Processing or **computer speech and language processing** or **human language technology** or **computational linguistic** is a subset of Artificial Intelligence that deals with enabling human-machine communication.

NLP is a phrase that is formed from 3 components - **natural** - as exists in nature, **language** - that we use to communicate with each other, **processing** - something that is done automatically.

Knowledge Base Approach vs Machine Learning
History

Historically language processing has been treated very differently in computer Science (*Natural Language Processing*), electrical engineering (*Speech Recognition*), linguistics (*Computational Linguistics*) and Cognitive Science(*Computational Psycholinguistics*). The earliest roots of the field date to the intellectually fertile period just after the second Wold War (1940's-1950's) [5].

Knowledge Base
Knowledge Base Approach

We introduced this presentation by talking about the defeat of the chess world champion Kasparov vs IBM's Deep Blue chess-playing system. This is one of simplest illustrations of the knowledge base approach, where a programmer can completely describe the brief rules of the a game to a computer ahead of time.
Because of how complex and usually difficult the wold is, most AI projects based on this approach failed, and lead researchers to think of another, more flexible, approach that can learn like a child.

The difficulties faced by systems relying on hard-coded knowledge suggest that AI systems need the ability to **acquire their own knowledge**, by extracting patterns from raw data. This capability is known as machine learning.
Introduction of machine learning allowed computers to tackle problems involving knowledge of the real world and make decisions that appeared impossible before.
For example a simple machine learning algorithm **naive Bayes** can separate legitimate e-mail from spam e-mail [4].

Natural Language Processing Sub-Field

Natural-language understanding

As a human, I understand English when someone talks to me, is that NLP? Yes! When done automatically, it is called **Natural Language Understanding (NLU)** [3].

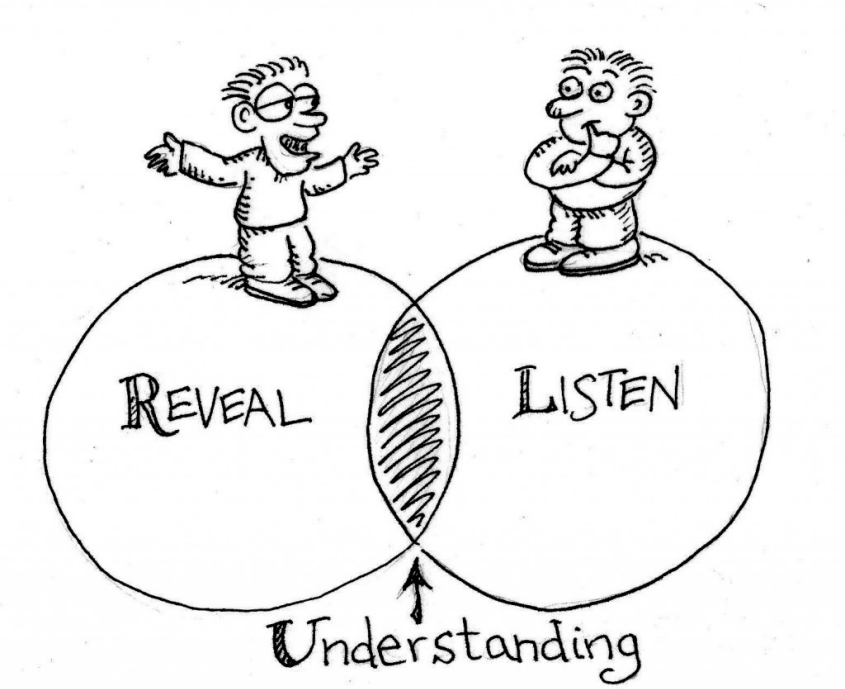


Figure: http://www.stuartduncan.name

Natural Language Processing Sub-Field
Machine Translation

I translated some Chinese to English for my friend, is that NLP? Yes, it is called **Machine Translation (MT)** when done automatically [3].

Figure: Image Credit: Google Cloud

Word Frequency

The idea is that the more often a word or term appears in a body of text, the more semantically significant it is.
In other words, the frequency of a word might tell us something about the meaning of the text.

Term Frequency (**TF**) and Inverse Document Frequency (**IDF**)

This approach use a more sophisticated measure of word importance.
Now let's look at the inverse document frequency. This is a measure of the relative number of documents within which the term appears. It's calculated as the log of total documents divided by the number of documents containing the term.
And finally, we just multiply TF by IDF to work out the overall importance of each term to the documents in which they appear.

Natural Language Processing Sub-Field
Text Processing II

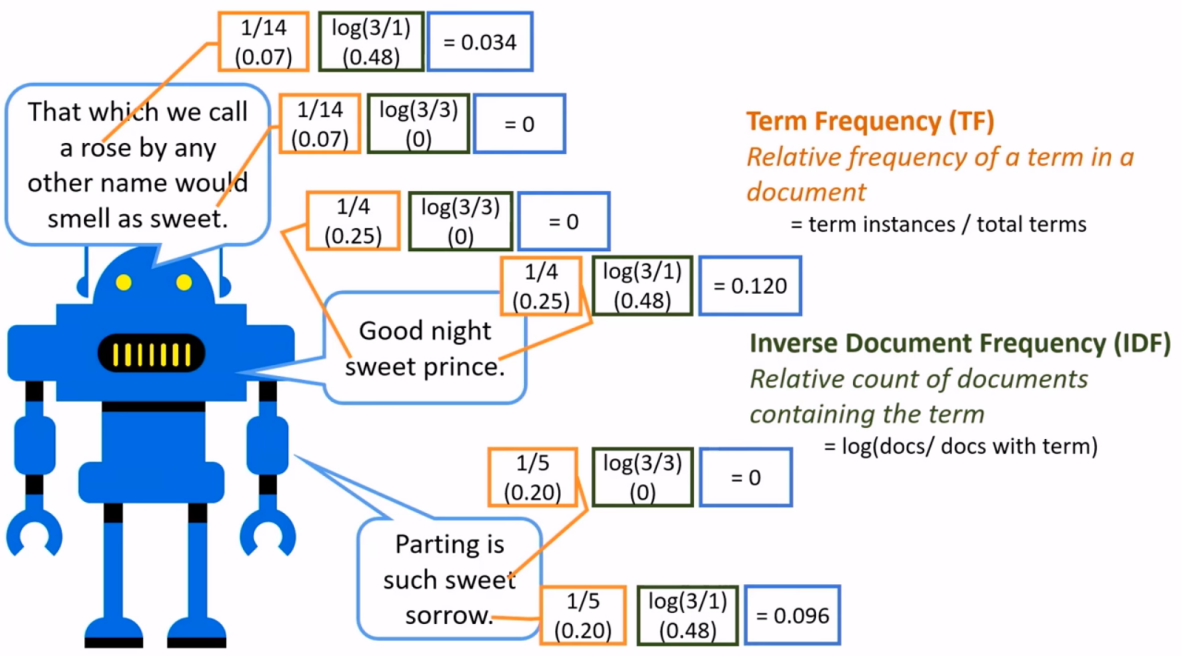


Figure: Microsoft, AI

Stemming or Lemmatization : There are sometimes words that are very similar, they're from the same root, stemming consist of judging those words as the same.

Natural Language Processing Sub-Field
Others

- Text Categorization

Natural Language Processing Sub-Field
Others

- Text Categorization
- Information Retrieval

Natural Language Processing Sub-Field
Others

- Text Categorization
- Information Retrieval
- Speech Recognition, and many others

Practical Example
TF-IDF (using python)

In this short tutorial, we will :

- Applying Term Frequency (TF) technique to a part of John F. Kennedy's talk (*We choose to go to the Moon*)

Practical Example
TF-IDF (using python)

In this short tutorial, we will :

- Applying Term Frequency (TF) technique to a part of John F. Kennedy's talk (*We choose to go to the Moon*)
- Applying Inverse Document Frequency (IDF) [1] on 3 different documents
  1. John F. Kennedy
  2. Abraham Lincoln talk
  3. Aims South Africa (About)

Open Questions I
Further study

- Translation : using world vectors between two languages can end poorly [2].

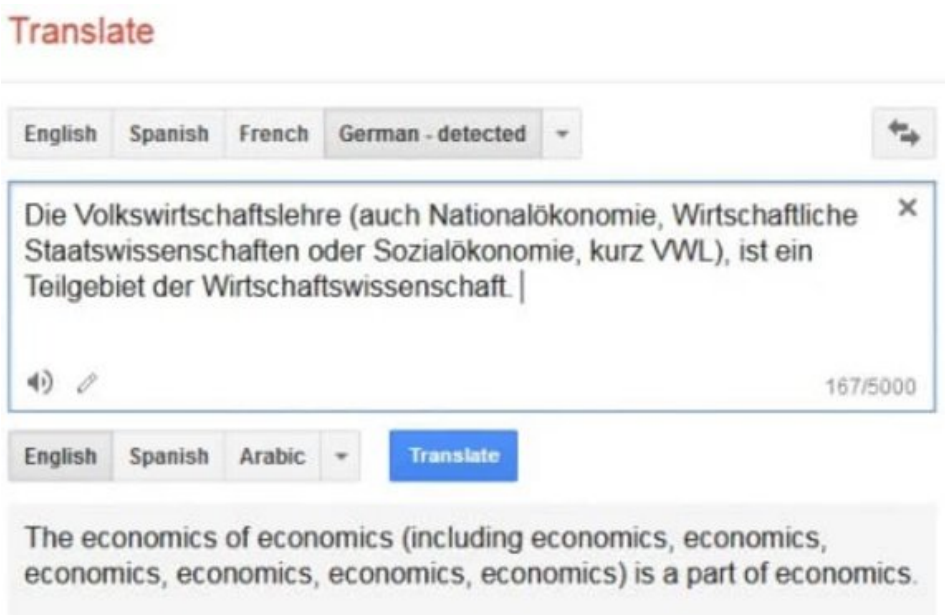
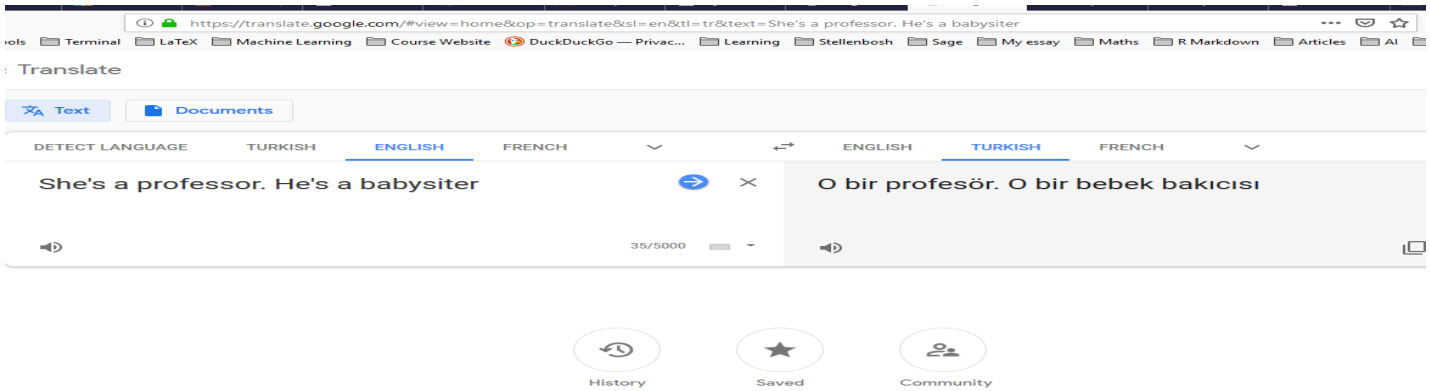
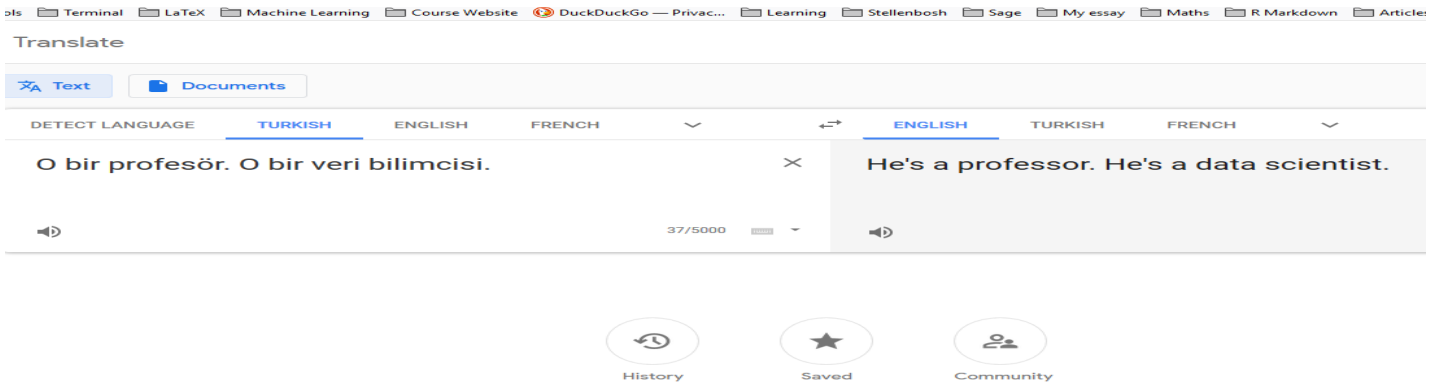


Figure: The German language has many different words related to economics, and they are all simply closest to the english vector for economics

Open Questions II
Further study

- Language can contain its own prejudices and unfair treatment towards groups. When we then train word vector on these prejudiced texts, our word vectors will likely reflect those problems [2].





References

[1] Online course : "Introduction to Artificial Intelligence (AI)".

[2] Online course : "NLT Fundamentals In Python".

[3] What is natural language processing (NLP)? accessed 19 December 2018.

[4] I. Goodfellow.
Deep learning.
Adaptive computation and machine learning. 2016.

[5] D. Jurafsky.
Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.
Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J., 2nd ed. edition, 2009.


TUASAKIDILA