Applications of Natural Language Processing and Machine Learning Algorithms

Shweta Kharat
Santa Clara University

Abstract
This paper states what Natural Language Processing(NLP) is. It describes different applications of Natural language processing and Machine Learning(ML) algorithms that are used for implementing NLP. This paper also contains different tools that can be used for Natural Language Processing.

1. Introduction

Natural language processing is a part of computer science and artificial intelligence that deals with the interaction of computers and human spoken languages. It helps convert natural language to machine understandable structures which can then be use to gain knowledge from the data like grouping related data together, knowing sentiment behind the text etc. Natural language means the language used by human beings for communication. Natural language processing algorithm helps computers by creating structures of human languages that gives it the human ability to understand language. NLP involves Speech recognition, Natural language understanding or Natural language generation. Natural language understanding means knowing the syntax, semantics and discourse for a particular language. Syntax includes parsing, sentence breaking, word segmentation etc. Semantics needs to convert words to machine language. Discourse means to summarize the data.

2. Tools and Algorithms for NLP

Different tools that does Natural language processing for us are Natural Language Tool Kit(NLTK), Stanford CoreNLP, Apache open NLP, NLP4J, GATE NLP. These are java libraries that we can use to implement NLP except NLTK which is collection of NLP softwares for different programming languages. By using these libraries, we can implement application as simple as dividing sentence into noun, adjectives and verbs.

3. Applications of Natural Language Processing(NLP)

Filtering the spam, understand a language, classify text, extract patterns and information, question answering, study social media texts, voice recognition for human language or speech to text conversion are some applications of NLP. More specific applications are tweet vigilance, sentiment analysis, building disaster knowledge base, finding cancer patients using clinical notes, voice controlled home automation. Few of the applications are discussed below.

3.1. Tweet Analysis [9]

Tweet analysis includes finding word density or word count, finding relativity between two tweets, finding top words, finding popularity of a topic, finding sentiment of a tweet or comment or clustering users and tweets.

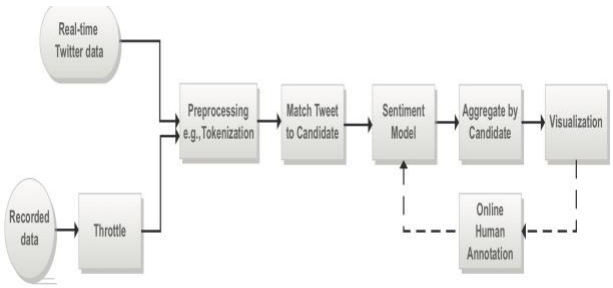
3.1.1. Twitter Vigilance [3]

Twitter Vigilance is a platform developed by university of Florence which collects and analyzes tweets in real time or daily basis. It uses metrics like part of speech for real time processing.

3.1.2. Twitter Sentiment Analysis [1][2]

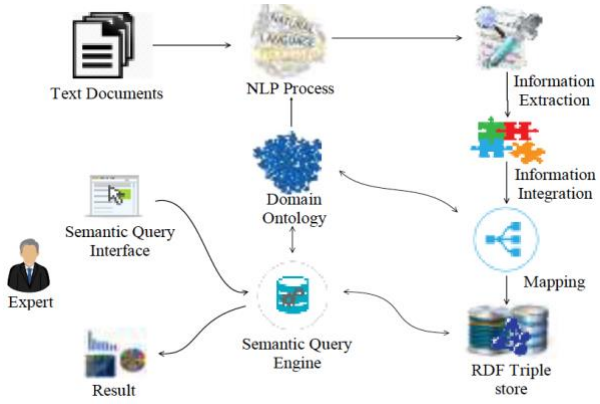
Twitter Sentiment Analysis is nothing but determining polarity of a tweet. We can

categorize tweets into positive, neutral or negative using sentiment analysis. SentiWordNet is an open resource that allows us to classify text into sentiments.
Example of sentiment analysis is A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle as shown in diagram below:



3.2. Building Disaster Knowledge Base [4]

Ontology and Natural language processing support is used to build a disaster knowledge base. The detailed process is as shown below:



It collects data from text documents and uses NLP to process the text which is in natural language format. Then the information is extracted and integrated to build mapping using domain ontology and saved in database. We can query data to get desired output.
Example: “In the year 2013, a flood disaster has occurred in the city Allahabad, Himachal Pradesh, India. In the disaster, 580 people were killed and over 2000 injured. The impact of this bad incident: property worth of nearly Rs.200 crores is damaged and more than 3500 families were homeless.”
Here, “a flood disaster” in sentence one, “the disaster” in sentence two are referred to the same data entity.

3.3. Machine Learning Algorithms

Simple algorithms like Naive Bayes algorithm or maximum Entropy can be used for NLP. But the problem is we just get 50% accuracy. We can improve accuracy up to 80% by training the model with negative examples or failure test cases. Other algorithms include Hidden Markov Model(HMM) and Conditional Random Fields(CRF).

3.3.1 Hidden Markov Model(HMM)

HMM creates sequence of tokens that gives us the information about the sequence of states called pattern theory which comes under grammar induction category. It is mostly used for speech and handwriting recognition in NLP.

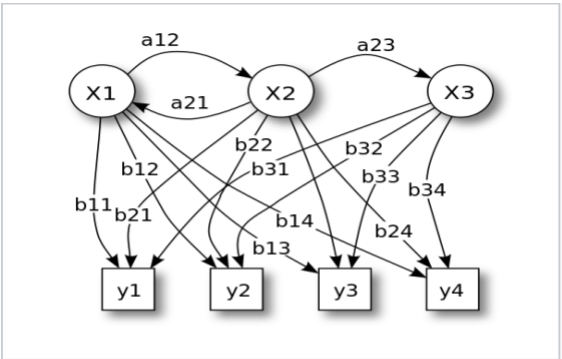


Figure 1. Probabilistic parameters of a hidden Markov model (example)
X — states
y — possible observations
a — state transition probabilities
b — output probabilities

3.3.2 Conditional Random Fields(CRF)

CRF is a type of statistical modeling method which is applied in pattern recognition and machine learning. It is used for structured prediction. CRFs categorizes as sequence modeling that is you solve the problem sequentially. CRF takes sequence of tokens as inputs by parsing the data and predict their tables sequentially.

4. Conclusion

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. To apply ML techniques to various NLP problems, we convert the unstructured text which is in the form of natural language into a structured format. Application of NLP mostly consists of text analysis or speech analysis for languages spoken by humans.

5. References

[1] *Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis using NLP Techniques by Monisha Kanakaraj and Ram Mohana Reddy Guddeti in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*

[2] *Twitter sentiment classification using stanford NLP by Shital Anil Phand, Jeevan Anil Phand in 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*

[3] *Twitter Vigilance: a Multi-User platform for Cross-DomainTwitter Data Analytics, NLP and Sentiment Analysis by Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Imad Zaza in 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation.*

[4] *Ontology and NLP Support for Building DisasterKnowledge Base by Sunitha Abburu and Suresh BabuGolla in Proceedings of the 2nd International Conference on Communication and Electronics Systems (ICCES 2017)IEEE Xplore Compliant - Part Number:CFP17AWO-ART, ISBN:978-1-5090-5013-0*

[5] *Closing the Loop - Finding Lung Cancer Patients using NLP byBipin Karunakaran∗, Debdipto Misra†, Kyle Marshall‡, Dhruv Mathrawala§ and Shravan Kethireddy in 2017 IEEE International Conference on Big Data (BIGDATA)*

[6] *Breast cancer staging using Natural Language Processing by Johanna Johnsi Rani G; Dennis Gladis; Marie Therese Manipadam; Gunadala Ishitha2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*

[7] *VOICE CONTROLLED HOME AUTOMATIONSYSTEM USING NATURAL LANGUAGE PROCESSING(NLP) AND INTERNET OF THINGS (IoT) by Mrs. Paul Jasmin Rani1, Jason Bakthakumar, Praveen Kumaar.B, Praveen Kumaar.U andSanthosh Kumar in 2017Third International Conference on Science Technology Engineering & Management (ICONSTEM)*

[8] *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining by Andrea Esuli and Fabrizio Sebastiani*

[9] *Twitter Part-Of-Speech Tagging Using Pre-Classification Hidden Markov Model by Shichang Sun, HongboLiu, Hongfei Lin, Ajith Abraham in 2012 IEEE International Conference on Systems, Man, and CyberneticsOctober 14-17, 2012, COEX, Seoul, Korea*

Acknowledgement

I would like to thank Professor Alex Sumarsono for his valuable guidance and support.