An Interpretation of Lemmatization and Stemming in Natural Language Processing

Divya Khyani[1], Siddhartha B S[2], Niveditha N M[3],Divya B M[4]
[1]*BGS Institute of Technology,*[2,3,4]*Adichunchanagiri University-BGSIT*

Abstract:*This research paper aims to provide a general perspective on Natural Language processing, lemmatization, and Stemming. It focuses on building up a base that helps in attaining a general idea over the technology. It explains the concept of Natural Language Processing, its evolution over the years, its applications, its merits, and demerits. In addition to that, it also gives a brief idea about concepts such as Lemmatization and Stemming. It helps in understanding their working, the algorithms that come under these processes, and their applications. At last, this research provides the comparison of lemmatization and stemming, attempting to find which one is the best.*
***Keywords: Natural Language processing, lemmatization, and Stemming*.**

1. INTRODUCTION

Natural Language Processing commonly called NLP among data analysts is the capacity of machine code to recognize human language the way it is spoken i.e; their natural mother tongue such as Hindi, Marathi, Tamil, etc. It includes two types of algorithm-one which take human-produced text as input and the other which produces natural-looking test as output. It helps computers to understand, manipulate, and interpret human Language.[5] It helps the developers to perform tasks like translation, text summarizing, relationship extraction, speech recognition, etc.

Division of NLP - **Natural Language Generation** and **Natural Language Understanding**

Natural Language Generation: Producing meaningful phrases and sentences in the form of natural language. This involves text planning, sentence planning, and text realization.[6]

Natural Language Understanding: For a better understanding of the text. This involves mapping the given input in natural language into useful representations for better analysis of the language.

2. History of NLP

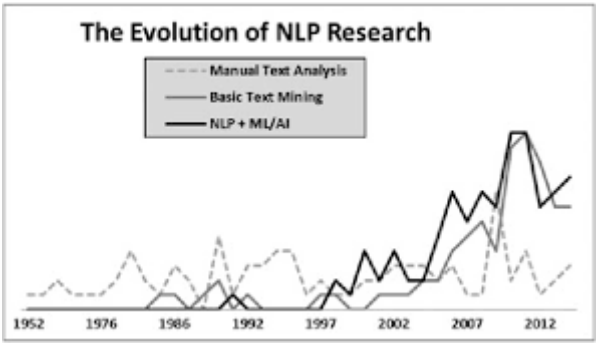

Figure 1: Evolution of NLP in research

The whole concept of NLP had emerged in the 1940's i. e; during the Second World War to convert one human language to another (Russia->English). In Fig 1 explains about the roadmap of NLP in research. In the 1950's Alan Turing wrote an article titled "**Computing Machinery and Intelligence**" now known as **Turing Test**. According to his statement, if a device could be considered as a part of a conversation with the support of a teleprinter and if it copies humans such
that there are no significant differences then the machine is recognized as capable of having thoughts. Later in 1957, Noam Chomsky published a book, Syntactic Structure where he created a style of grammar called "**Syntactic Structures**". The aim was to developa machine that can reciprocate the human brain, mainly in terms of thought process and conversation capability.

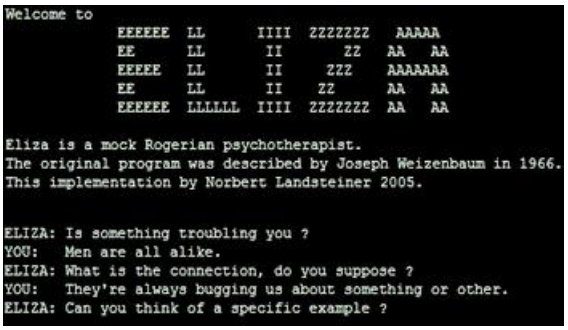

Figure 2: Example of ELIZA

ELIZA, a chatbot was designed by Joseph Weizenbaum to imitate a psychiatrist using refection techniques from 1964 to 1966 at the Artificial Intelligence Laboratory of MIT. Later in the 1960's, Fig 2 a US-based Research Council (NRC) created the **Automatic Language Processing Advisory Committee** (ALPAC) to evaluate the progress of NLP research. During 1966 a fall was observed since there were no computers capable to carry a basic conversation translation as per the reports of ALPAC. Thus, ML/NLP almost died due to a very low pace observed in the research.
With the need for Artificial Intelligence in the 1970s, NLP got a new life. LUNAR was developed by W.A Woods in 1978 which analyzed, compared, and evaluated chemical data on a lunar rock and soil composition, urn answered all the related questions. In the 1990s, the pace of growth of MT/NLP increased. More and more concepts came into existence such as word sense disambiguation, statistical language processing, information extraction, and automatic summarizing.
In 2001, **Yoshio Bengio** along with his support members introduced the first neutral language prototype using a feed-forward neural network where information moves only in one direction from input nodes to the output nodes with the help of hidden notes. In 2011, Apple gave in a huge development to the world via **SIRI's** invention. It was considered as very essential as it was the very first NLP/AI assistants which a general customer could interact with. Within Siri, the Automated Speech Recognition module enables the conversion of the customer's words into digitally understood theories.

3. Why and where it is used

Why: NLP is required when we want an intelligence system like a robot to perform as per our instructions. It helps computers communicate with humans in their language and scales other language-related tasks. It helps the computer to read the formatted text, hear the conversation, understand it, deduce emotions, and recognize the essential portions. Since human language is very complicated and varied, they tend to convey their thoughts through numerous methods in verbal fashion as well as in written format.[1] NLP is important because it helps in resolving ambiguity in languages and adds structure to the data for text analysis and speech recognition.

Everyday life example: When we type something on our phones, we tend to look at words suggested on the search engines based on what we entered in the search box and what are we presently concerned about.
Business life example: Some companies trying ways to advertise in a way such that more customers are attracted. They can use google to find common search terms that their users type when searching for the company product.

Where: It is used in numerous fields such as business, sports, art, health, marketing, education, politics, etc
It is used for Language Translation, voice responses, personal assistance, etc.NLP is also used to identify the unstructured data widely present and extract the natural Language so that it gets converted to a computer understandable format. Using NLP, multiple interactions can be carried out between machines and humans including humans talking to a machine, the machine captures some audio, processing, and converting to text format for a better understanding.

4. Applications

1. **Sentiment Analysis:** It is used to understand the replies received to the business-related messages posted on socialplatforms, also termed as opinion mining.[2] It is done by substituting some figures to the text-format as positive, negative, or neither of them keeping it neutral to recognize the emotion beneath the quoted words (happy, sad, angry, annoyed, disgusted, irritated,etc.)
2. **Chatbots:** They are a relief to consumers frustrated by the customer care call assistance. They are the immediate need of the hour since they offer real-time solutions for simple as well as complicated customer related issues. They are trusted by the consumers, and engineers as they help retain time, labor work, price, and provide a reliable approach to problems in addition to being very well-liked.
3. **Customer Service:** NLP helps to gain perspective into audience tastes, preferences, and mindsets. Taking an example of a customer's recorded call, it holds the emotions which he/she is experiencing at the moment when on that call.This can help in fulfilling their expectations in the future and understanding their present views, hence helps in receiving helpful feedback.
4. **Managing the advertisement:** NLP plays a major role in the positioning of advertisements in the appropriate place at the correct time and for the approachable audience. NLP matches the unique words in the formatted text and helps in reaching the appropriate customers.
5. **Market Intelligence**: NLP helps to follow and observe market progress documentation and pull out important data to build new techniques.
6. **Machine Translation:** It is defined as the process in which one source text or Language is converted into another one fulfilling the requirements for the market enhancement.[7]
7. **Helps in dealing with spams:** Spam filtering system will help in locating the spam data and filter it out. It can be made by using NLP functionality by considering the majorly found false-positive and false-negative issues.
8. **Automatic Summary:** This is yet another application widely needed because there is massive information flooding all around the net and is never-ending. It is the method to create a brief, appropriate gist of huge formatted-text documentations. This technique will provide us concise information in a small amount of time and with better accuracy.
9. **Question Answering:** It aims on developing systems that automatically find answers to the problems frequently asked by human beings in their mother tongue.[2] This can be done by syntax and semantic study of the queries.

5. Advantages of NLP

- NLP includes automatic summary generation which provides us with a readable summary of text like newspaper also referred to as an automatic content generation.
- It helps in determining which words in a chunk of the text refers to the same object if multiple sentences are given, the process is termed as Co-reference Resolution.
- It is also adopted by various companies to improve the efficiency of Documenting, thus improving the accuracy and finding the relevant information from a large chunk of databases.
- Other merits may include a high rate of flexibility, structuring highly unstructured data, performing more language-based data comparison, and gives solutions to the problems regarding any concept which is related to the processing of natural language.
- No indexing is necessary in the case of NLP which makes this technology stand out.

6. Disadvantages of NLP

- Context expert knowledge such as sarcasm is one of the major disadvantages of NLP wherein the actual meaning conveyed has to be opposite to what it implies.
- The visual context is lost while implementation, it is very unpredictable and requires clarity to understand.
- It becomes very difficult to understand the generalized searches. There have also been problems observed when dealing with synonyms, the user is expected to not create chaos and think of his search term. NLP system performs a single specific task and does not have a user interface.

7. What is Stemming?

It is defined as the process which produces variants of a root/base word. In simple words, it reduces a base word to its stem word. We use stemming to shorten the look-up and normalize the sentences for a better understanding.

Example:

1.for the root word “like “
Stemming will include :
-”likes”
-”liked”
-”likely”
-”liking”
2. for the root word “chocolate”
-”chocolates”
-”chocolatey”
-”choco”
So basically, all the words are reduced to chocolate by the stemming algorithm.

7.1 Applications of Stemming:

- Information Retrieval systems like search engines
- Determining domain vocabularies in domain analysis

7.2 Stemming Algorithms:

Various algorithms help in performing Stemming.Some of the stemming Algorithms are:

7.2.1 Porter Stemmer Algorithm:
This stemming Algorithm is used to remove and replace suffixes of English words to make words simpler and efficient. NLTK has a **PorterStemmer** class with the help of which we can implement the Porter Stemmer Algorithm.[3] Using this class we can convert the given text-word to the resultant stem which in turn gives a shorter word having the same root meaning.
For Example racing----> race

```
import nltk
from nltk.stem import PorterStemmer
word_stemmer = PorterStemmer()
word_stemmer.stem('racing')
```

Output: ‘race’

7.2.2 Lancaster Stemming Algorithm:
It is named after the university “Lancaster” as it was developed there. It is a very common stemming technique. NLTK has **LancasterStemmer** class with the help of which we can implement the Lancaster Stemming Algorithm to perform stemming.
For Example: eats--->eat[8]

```
import nltk
from nltk.stem import LancatserStemmer
Lanc_stemmer = LancasterStemmer()
Lanc_stemmer.stem('eats')
```

Output: ‘eat’

7.2.3 Regular Expression stemming Algorithm:
With the help of the Regexp stemming algorithm, we can build our stemmer.
NLTK has a **RegexpStemmer** class with the help of which we can implement Regular Expression Stemmer algorithms. It takes a single expression as input and removes suffix and prefix that matches the expression.
Example:ingeat--->eat

```
import nltk
from nltk.stem import RegexpStemmer
Reg_stemmer = RegexpStemmer()
Reg_stemmer.stem('ingeat')
```

Output: ‘eat’

7.2.4 Snowball stemming Algorithm:
It is a very useful stemming algorithm.NLTK has a **SnowballStemme**r class with the help of which we can implement Snowball Stemmer algorithms. It supports 15 languages other than English. For using this class, we have to build a sub-form with the language name that we are making use of and then use the stem() method on its successful creation.
Example: Bonjoura--->Bonjour

```
import nltk
from nltk.stem import SnowballStemmer
French_stemmer = SnowballStemmer('french')
French_stemmer.stem ('Bonjoura')
```

Output: 'Bonjour'

8. What is Lemmatization?

It is the process of assembling the inflected parts of a word such that they can be recognized as a single element, called the word's lemma or it's vocabulary form.[3] This process is the same as stemming but it adds meaning to particular words. In simple words, it connects text with alike meanings to a single word.
It is defined as an algorithm technique of finding the lemma of a word which is a root word rather than a root stem.[4] It is based on the intended meaning the word is trying to convey.
Example:
1. rocks: rock
2. Corpora: corpus
3. Better: good

8.1 Applications of Lemmatization:

- Use in Biomedicine: processing of text related to biomedicine can be efficient by using specialized lemmatization and may increase the efficiency of data retrieval tasks.
- Used in comprehensive retrieval systems like search engines
- Used in compact indexing

8.2 Implementation of lemmatization words using NLTK:

NLTK provides **WordNetLemmatizer** class which is a slim cover wrapped aroundthe **wordnetCorpus**. This class makes use of a function called M**orphy()** to the **WordNetCorpusReader** class to find a root word/lemma.

```
import nltk
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
lemmatizer.lemmatize('books')
```

Output: 'book'

9. Difference between Lemmatization and Stemming

- On one hand, Stemming techniques are implemented by chopping off from either the end or beginning of the word-text, keeping track of commonly used prefixes and suffixes that could be available in an infected word-text, and On the other hand, lemmatization considers the study of the word-texts intending to find something which adds meaning to it.
- Lemma is the foundation of all its inflected parts, and a stem isn't. This is why orderly dictionaries are a record of lemmas and not of stems.
- examples in support

  a. **Stemming:**
  Form=studies, suffix = -es, stem = studi
  Form=studying, suffix = -ing, stem = study

  b. **Lemmatization:**
  Form=studies, Meaningful Information=Third person,singular number,present tense of the verb study, lemma=study
  Form=studying, Morphological Infomation=Gerund of verb study,Lemma=study

10. Which one is best: lemmatization or stemming?

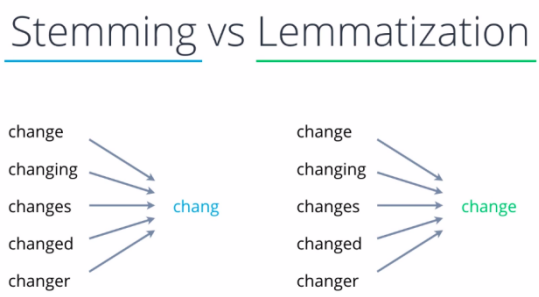

Figure 3: word stemming and Example of reduction in lemmatization.

Even though performing stemming is easier than the process lemmatization, the latter is proves to be the ultimate choice. Deep linguistics understanding is required to form the glossary that permits the algorithm to search for the meaningful part of the word in the process of lemmatization. Once this is done, the outcome will be more accurate. It makes use of vocabulary and morphological analysis of words to receive output free from derivative affixes.
On the other hand, stemming simply chops off the end of words irrespective of the fact that the output is conveyed any meaningful information.

CONCLUSION:

Processing of Natural Language is the area of study which focuses on making the communication between the language spoken by humans and machines possible. It helps to understand the text, enabling machines to recognize how a particular human talks in various languages concerning the place they reside. This human-like computer communication enables existent world implementations such as automatic text summary generation, the study of emotions, pulling out the relationship, and many more. Lemmatization and Stemming play a major role in text and natural language processing. Both of them create the base part of the inflected word-texts. The difference lies in the fact that stem is not a real word-text whereas lemma is a real language text format.

Acknowledgments

These should be brief and placed at the end of the text before the references.

REFERENCES

[1] Gelbukh, "Natural language processing," Fifth International Conference on Hybrid Intelligent Systems (HIS'05), Rio de Janeiro, Brazil, 2005, pp. 1 pp.-, DOI:10.1109/ICHIS.2005.79.
[2] Khurana, Diksha & Koli, Aditya & Khatter, Kiran & Singh, Sukhdev. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges
[3] Balakrishnan, Vimala & Ethel, Lloyd-Yemoh. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. Lecture Notes on Software Engineering. 2. 262-267. 10.7763/LNSE.2014.V2.134.
[4] Halácsy, P. (2006). Benefits of deep NLP-based Lemmatization for Information Retrieval. *CLEF*.
[5] Joseph, Sethunya & Sedimo, Kutlwano & Kaniwa, Freeson & Hlomani, Hlomani & Letsholo, Keletso. (2016). Natural Language Processing: A Review. Natural Language Processing: A Review. 6. 207-210.
[6] Sun, Shiliang & Luo, Chen & Chen, Junyu. (2016). A Review of Natural Language Processing Techniques for Opinion Mining Systems. Information Fusion. 36. 10.1016/j.inffus.2016.10.004.
[7] Krishna Prakash Kalyanathaya, D. Akila, and P. Rajesh. Advances in Natural Language Processing – A Survey of Current Research Trends, Development Tools, and Industry Applications. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5C, February 2019.
[8]https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html