Natural Language Processing and its Potential Future Applications

Aarushi Mishra

69460793

ESL 515 C

Cassandra Rosado

Fall 2017

Introduction

Since the innovation of computers, the technology industry has witnessed many advances in almost every field. Computer can be considered as the most helpful and intelligent innovation by human kind. It has not only rendered ease to our life but has led to discoveries that were once beyond our reach. It is almost impossible to name even a single field of development where computer is not used. Until 1950s, advancement in computer referred to improving the speed at which computations were done and reducing the size of the computer that was once elephantine in size. However, today the advancements have taken place that are beyond what a person might have expected 50 years ago. In 1956, the term *Artificial Intelligence* was used for the first time at Dartmouth College in a conference organised by John McCarthy, the man who is considered as the father of Artificial Intelligence (Buchanan, 2005). Since then the scientists and researchers have developed machines that in some or the other way illustrate characteristics of human intelligence. The first AI enabled machine could play Chess with a human as its opponent. Hence, the identified goal of AI is to generate expert systems that exhibit human intelligence i.e. learn, behave intelligently, demonstrate, explain, and advice its users, think like humans and much more.

Natural Language Processing, or NLP, is a discipline within Artificial Intelligence and is one of the most popular area of study today. Modern human being tends to like automated processes. Machine, any device that makes traditional process of accomplishing any task easier, has been a perpetual fascination for scientists and innovators. Perhaps, this fascination has led to revolutionary developments of machines. After attaining accuracy and efficiency, the next goal is to build machines that exhibit human intelligence. Natural Language Processing is the field of study that focuses on how humans and machines interact

with each other. According to Gill (2017), it is a subset of AI that aims to make this interaction possible through human intelligence i.e. by making machines learn and interpret the language used by humans. It is the union of characteristics of computer science, computational linguistics and artificial intelligence. This paper presents a brief introduction to Natural Language Processing and the steps through which this form of artificial intelligence can be attained. The main purpose is to understand the concepts of NLP and infer potential future applications.

Why do we need Natural Language Processing?

Natural Language Processing is one important aspect of Artificial Intelligence. This is because it renders machine the ability to communicate. Effective communication is always considered as the indispensable feature of any system and therefore, NLP plays a crucial role in creation of intelligent devices. According to Gill (2017), "The ultimate goal of NLP is to the fill the gap how the humans communicate (natural language) and what the computer understands (machine language)" (p. 1). Realising numerous possibilities of what NLP can achieve, it is now considered as an indispensable facet of Artificial Intelligence.

Appropriate use of NLP can lead to increased efficiency of humans in terms of time and accuracy. Unlike humans, computers can work relentlessly, and NLP can help to exploit this "non-human" feature to attain "human" goals. For example, automated speech and automated text writing can reduce the loads on human workers by a remarkable factor, perform tasks through written programs on large data sets in small fraction of time, automatic summarization and automatic translation for understanding complex phenomenon and non-familiar languages. NLP might not be a "must-have" but it is the future of intelligence devices. It is a crucial key in maximizing our efficiency so that other complex problems can be explained.

Phases of Natural Language Processing

The process of using Natural Language Processing for language analysis is not facile. To ensure the success of such a system, it is required that the process of language interpretation is done accurately. This is crucial because of the degree of ambiguity that exists in natural languages. In many languages, the meaning of a word is variable with respect to the context in which it is used. Such systems, which carry the ability to resolve ambiguity of language are complex in design and methods and following the structure and fundamental steps is the key to developing accurate intelligent machines.

The process of language analysis can be broken down into five stages as shown in the figure:

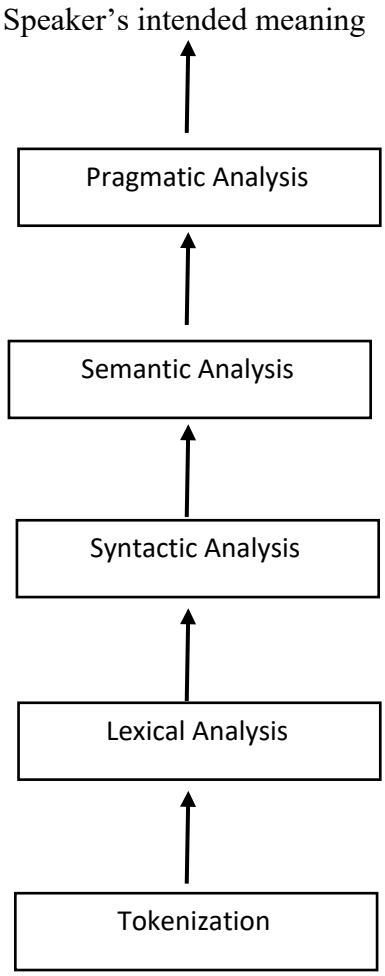

Figure 1.1: The stages of analysis in processing natural language

From "Handbook of natural language processing (Vol. 2)," by F. J. Damerau and N. Indurkhya, 2010, *CRC Press, Taylor and Francis Group, Boca Raton, FL*. ISBN 978-1420085921. Copyright 2010 by *Taylor and Francis Group, LLC*

1. Text Preprocessing

To initiate the analysis of a natural language, it is first necessary to breakdown the language into the smallest unit of that language and then interpret the meaning. According to Damerau and Indurkhya (2010), text preprocessing refers to converting the raw text into

"linguistically meaningful units". One of the important stages of text preprocessing is text segmentation where input language is decomposed into its component words and sentences. It can be achieved by word segmentation and sentence segmentation.

Word Segmentation or Tokenization is the processes of identifying the start and end of each word and hence recognise each word as the smallest, indivisible, meaningful unit of the language called token (Adi et al., 2017). Sentence segmentation refers to the process of identifying the start and end of a sentence. Sentences are separated by punctuation marks in most of the languages (Damerau and Indurkhya, 2010). The purpose of this stage to define sentence boundaries. This process is also referred as sentence boundary recognition.

2. Lexical Analysis

Words are the building blocks of any natural language and carry significant units of information. For natural language based intelligent systems, it is necessary to include operations or steps that register a word's structural and substantial properties. This is done in the process called Lexical Analysis. According to Damerau and Indurkhya (2010), lexical analysis is the process of connecting each word with its meaning given in a dictionary with the help of labels. Thus, it helps to understand a word's structure and other related properties. The challenge is to identify the appropriate meaning in the given context since a single word may have multiple meanings. The tokens, that were identified in the previous stage, are then classified into grammatical classes (e.g. verb or noun) and then it is associated with a meaning. For example, verb deliver represents a set of words {delivers, deliver, delivering, delivered}. It is the task of the lexical analyser to interpret the correct association between the meaning and the given word (Kolomiyets and Moens, 2011).

3. Syntactic Analysis

Analysing the meaning of the words is not enough. Arrangement of these words matter to make sense of the information. A group of words much be arranged so that they are grammatically correct and morphologically appropriate. According to Damerau and Indurkhya (2010), syntactic analysis refers to the process of checking a sentence for its grammatical correctness and infer meaning of the sentence. For example, consider these two sentences, “Martha goes to party” and “Party goes to Marth”. A syntactic analyser will reject the second sentence because it does not make any sense to the reader or the listener. Therefore, arrangement of words and other semantic rules play a crucial role in rendering the intended meaning.

4. Semantic Analysis

Semantic Analysis is the process of drawing exact meaning of the sentences. In this stage, the language is processed to figure out the sense of the words that rest of the world can relate. Broadly, it refers to understanding the language in terms of correct usage. It includes complex tasks such as finding synonyms of the words, developing question-answering systems, translation of one natural language to other and word sense disambiguation. This is the most difficult and complex stage as deriving the right meaning is a gargantuan challenge. According to Damerau and Indurkhya (2010), one of the significant problems is that of abstract nouns. For example, exact equivalent of English words such as rights, security, dialogue, trauma and experience do not exist in many other languages. The variety of languages being spoken all over the world is too high. It is strenuous to collect data for all the languages.

Semantic analysis is followed by the process of pragmatic analysis. This analysis requires application of knowledge based on real world scenarios. The words uttered are re-

interpreted to figure out what they mean (Adi et al., 2017). Understanding the context and purpose of the uttered words is the main aim purpose of this stage.

Natural Language Generation

Natural Language Generation is different from Natural Language Processing as during language processing, the computer reads a language. On the other hand, in language generation, computer writes a language. According to Damerau and Indurkhya (2010), "Natural Language Generation is the process by which thought is rendered into language". A generator in language generation is essentially a computer program that starts with the intention to communicate, predict what is to be said, select the right words and organise them in the right order that conveys a meaning. Until last decade, the focus was to make machines intelligent enough so that they can read and acquire information from the input source. This aim has been accomplished successfully but now the problem has overturned. The computers now have access to large datasets and repositories though realising the value of data and extracting useful insights from the data is the new challenge. Language generation serves this purpose.

Language generation is a complex process as compared to language processing. In language processing, the source is fixed and obvious. The input for a language analyser is a series of written text or a sound wave. On the other hand, the source for language generation is not fixed. Rather, it is an abstract concept. According to Damerau and Indurkhya (2010), "the source [for language generation] is a 'state of mind' inside a speaker with 'intensions' acting in a situation" (p. 124). Lack of consistency in pointing out the precise source for language generation is the major problem in this area of research.

NLP based Intelligent Systems

Chatbots

Natural Language Processing offers myriad of potential applications. Out of many such possible applications, chatbot is one application that has been implemented successfully and is being used by the people extensively. Colossal brands like Tommy Hilfiger and food chains like Taco Bell use chatbots for customer support which otherwise is done by a representative. They decrease the customer's wait time and improve services offered. Models that apply the principles of NLP are used to design the brain of these chatbots. They not only handle customer requests but collect significant data while interacting for analysing the trends and extracting meaningful information.

Another in-use application of NLP is talking personal assistant that come integrated with devices such as mobile, tablet and laptop. Some common examples of this application are Amazon's Alexa, Microsoft's Cortana and Apple's Siri. Using natural language processing and machine learning, these applications interpret what the user is saying and delivers an appropriate or related output that the user expects. According to Gill (2017), "The true success of NLP lies in the fact that humans deceive into believing that they are talking to humans instead of computers" (p. 1). These bots are very realistic and give an impression of a real person talking and listening.

Information retrieval

Chatbots and personal assistants are two common existing applications of NLP. The scope is not restricted this only. NLP finds its applicability in variety of areas. Sentence embedding models and information retrieval are some other significant applications of NLP. Question answering technology is one of the major techniques to retrieve information. The survey data collected by the systems uses question answer model to collect data from various users so that the data can be studied to identify the trends. The amount of data produced every

year is increasing exponentially. Querying data from such large datasets stored on various local and distributed systems may prove costly in terms of time and money if the right technique is not employed. In contrast to traditional information retrieval method where all the information was considered relevant, the question answering technology focuses only on necessary words or phrases that serves the purpose of information retrieval (Damerau and Indurkhya, 2010). The precise identification of required piece of information and hence saving the retrieval cost is the main advantage of this technology.

NLP plays an important role in breaking down the information. Figure 2 clearly shows the steps of a Question Answering model. The phases of natural language processing are applied to retrieve precise output that serves as input for the retrieval models.

Question answering technology finds diverse applications that are currently contributing in making the overall system intelligent that retrieve information. It finds many domain specific applications such as ontology- based question answering for movies and cinemas and some other applications such as Ask.com (Kolomiyets & Moens, 2011). With extensive use of World Wide Web, precise answer to queries is necessary. Web-based question answering systems like Ask.com help to increase the efficiency in terms of information retrieval time required.

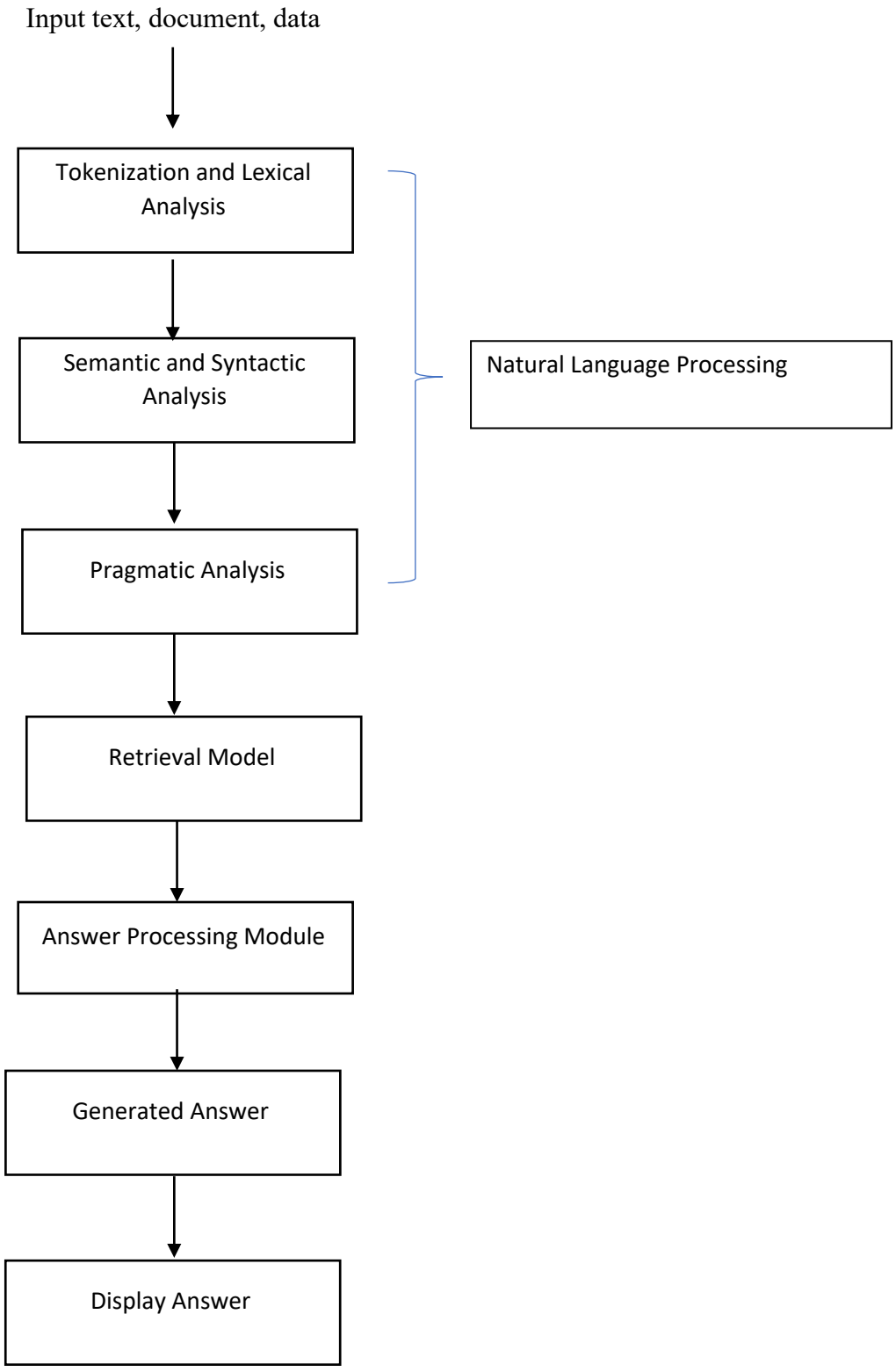
Input text, document, data
Tokenization and Lexical Analysis
Semantic and Syntactic Analysis
Pragmatic Analysis
Natural Language Processing
Retrieval Model
Answer Processing Module
Generated Answer
Display Answer

Figure 2: Flow of Question Answer Technology

Future Applications of Natural Language Processing

1. Biomedicine

Healthcare is one of the areas which has observed enormous impact of technology. NLP plays an important role in the field of biomedicine as it helps to convert the raw data into structured data that can be used by computer. According to Friedman et al (2013), NLP can be employed to recognise entities in a text adequately. The entities may comprise of persons, patients, drugs, body substances or diseases. Taking one step ahead, other than recognising the entities, concepts of NLP can be employed to derive relation among these entities.

Another possible application of NLP is information extraction and hence the decision making (Friedman et al., 2013). Along with cost effective data reading techniques such as storing annotated information and parse tress, NLP can be used in combination to achieve decision making by building reasoning chains. For example, the above method can be used to gain knowledge about protein-protein interaction and deduce hypotheses for drug-drug interactions.

The scientists and research groups are constantly trying to discover what more NLP has to offer for healthcare. Presently, systems employing NLP are based on supervise learning, but researchers stress on the use of automated learning to regulate the costs of previous systems.

2. Online Essay Grading Module

The aspects of education have revolutionised to a great degree as compared to education practices that were followed couple of decades ago. Today, almost everything is done online. E-books, online documents, online lectures and interactive course platforms have successfully replaced books, notebooks, physical presence in a lecture and in-class assignments respectively. To some extent, assessment of a student's progress also takes place over the internet and is made available for students for analysis. Major advantages of online assessment include time efficiency, absence of human error and instant feedback and grading. Answer to question types such as multiple choice and single choice, can be assessed using algorithms and relevant software. However, answer to some question types, such as an essay, requires semantic and syntactic analysis which in turn requires teacher's supervision and efforts.

Moodle is one of the popular LMS (Learning Management System) which is widely used for academic purposes by schools and universities. It is an efficient integrated system that offers many features although, it lacks to provide assessment functionality for essay type answers. Ajitiono and Widyani (2016) propose a system that uses the concepts of NLP to analyse answers semantically and syntactically availing teachers to reduce the total grading time required.

The various steps involved in the overall grading system are as shown in the figure.

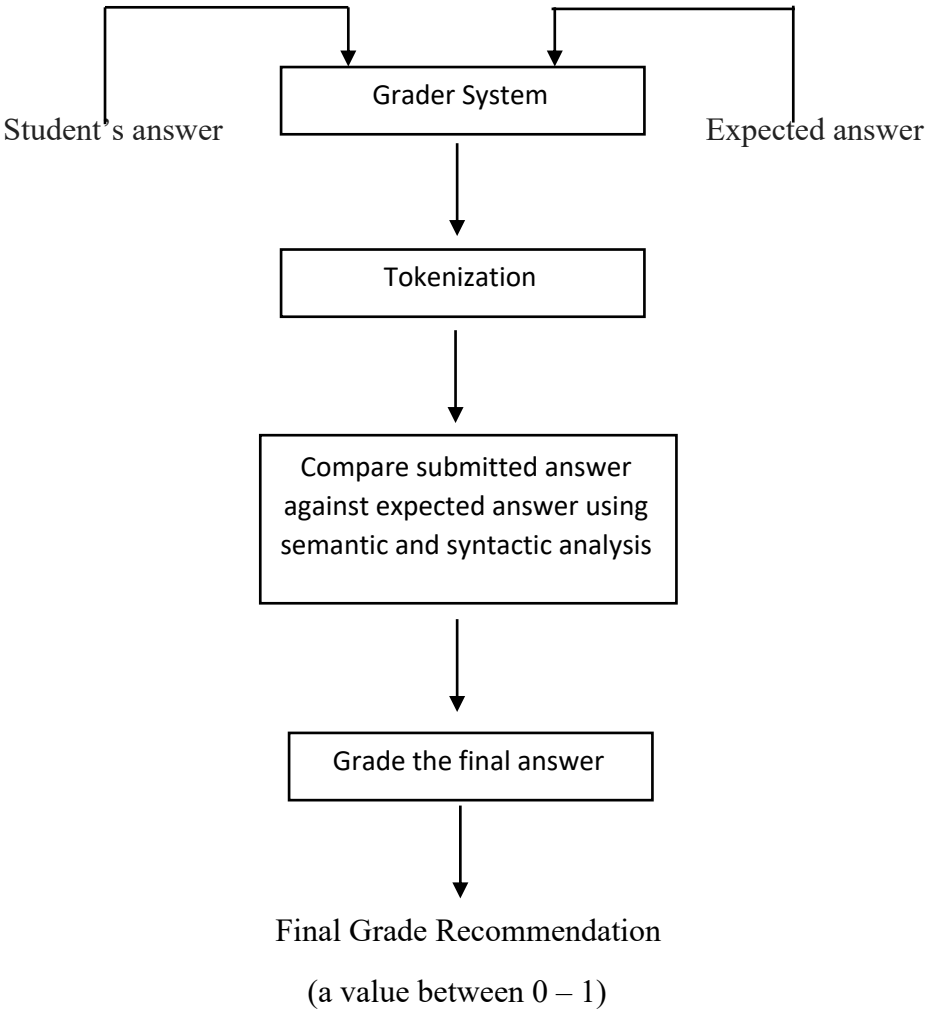

Figure 3: High-level model of Moodle Grading System

The proposed grader system accepts two values: student's answer and expected answer. The answers are then fed into the INANLP, a NLP tool used as plugin for Indonesian

language. Natural language processing starts at this point. In INANLP, the answer goes through the various stages of NLP i.e. tokenisation, part of speech tagging and comparison of sentences based on semantic and syntactic analysis. After analysis, the final grade recommendations are generated for teacher's reference. This final output is a decimal number between 0 to 1 where '0' refers to the worst fit of the sentence and '1' refers to the best fit of the sentence.

The developed module by the authors offers a method to improve the time efficiency of the teachers using the NLP for grading essay type answers. The deployment of this system will ensure automation of the existing system to increased degrees.

3. Other applications

Natural Language Processing can support myriad of applications in future. Current technological advancement demands intelligent systems and NLP can help to achieve the development of such automated systems. Other than applications discussed till now, there are many other ideas that can be converted into full-fledged working models of NLP.

With large amount of data available, the challenge is to make the data available among the users all over the world without any language barriers. Machine translation is considered an important application of NLP for this purpose (Gill, 2017). The goal of machine translation is not only translation of words but also to preserve the meaning of the words. NLP can be deployed in such a scenario since preserving meaning is the main challenge faced by the researchers studying language processing and automated learning systems.

Spam filter algorithms can be deployed to solve the problem of email flooding by unwanted emails. The concepts of NLP can be used to break down the incoming text and

define the intention of the content of the mail and classify them as spam or non-spam (Damerau and Indurkhya, 2010).

Summarization is another potential application of NLP (Gill, 2017). The amount of information and data available due to the successful techniques of data collection and data extraction is far beyond what a human mind can comprehend altogether. Natural Language Processing can be used to generate summaries of lengthy documents. This will help human workers to absorb substantial information from a large chunk of data in less time than usual.

Conclusion

This paper presents a summary of what natural language processing is and infers its future applications. NLP is gaining popularity due to its potential of addressing problems of semantics and retaining meaning of the processed text. Storing mechanisms along with intelligent systems that can make sense of the data are required to handle the current and anticipated explosion of data. And therefore, due to wide variety of identified challenging problems, NLP has been an active area for research. Techniques such as summarization, machine translation offer a great deal of advantages and must be explored for real world implementations. Finally, the paper concludes by highlighting the importance of NLP and throws light on the immense potential of this field by inferring some of the possible applications of NLP.

Prospects of NLP are burgeoning currently as it delivers the power to achieve that was once impossible. With advancement in NLP, a person who is not an expert in a particular area will be able to overcome the domain- related obstacles without spending much efforts. Recently, a robot named Sophia received the citizenship of Saudi Arabia. The robot is almost the replica of that of a real human being. Indeed, NLP has revolutionised the face of Artificial Intelligence and will oblige us to witness many such wonders in the coming future.

References

Adi, Y., Kermany, E., Belinkov, Y., Goldberg, Y., & Lavi, O. (2017). Analysis of Sentence Embedding Models Using Prediction Tasks in Natural Language Processing. *IBM Journal of Research and Development*, *61*(4)*,* 3-1.

Ajitiono, T., & Widyani, Y. (2016). Indonesian Essay Grading Module Using Natural Language Processing. *In Data and Software Engineering (ICoDSE), 2016 International Conference,*1-5. IEEE.

Buchanan, B. G. (2005). A (Very) Brief History of Artificial Intelligence. *Ai Magazine, 26*(4), 53.

Damerau, F. J., & Indurkhya, N. (Eds.). (2010). Handbook of Natural Language Processing (Vol. 2). *CRC Press, Taylor and Francis Group, Boca Raton, FL*. ISBN 978-1420085921.

Friedman, C., Rindflesch, T. C., & Corn, M. (2013). Natural Language Processing: State of the Art and Prospects for Significant Progress, A Workshop Sponsored by The National Library Of Medicine. *Journal of Biomedical Informatics, 46*(5)*,* 765-773.

Gill N. (2017). Overview of Artificial Intelligence and Role of Natural Processing in Big Data. *Data Science Central.* Retrieved from

http://www.datasciencecentral.com/profiles/blogs/overview-of-artificial-intelligence-and-role-of-natural-language

Kolomiyets, O., & Moens, M. F. (2011). A Survey on Question Answering Technology From An Information Retrieval Perspective. *Information Sciences, 181*(24), 5412-5434.