International Journal of Emerging Technology and Innovative Engineering
Volume 4, Issue 2, February 2018 (ISSN: 2394 – 6598)

Paper Received on: 08.02.2018
Paper Published on: 24.02.2018

APPROACHES BASED ON DATA MINING IN NATURAL LANGUAGE PROCESSING

V.Deepa,M.Tech(CSE)[1#] A.Jenifa[2#] J.Pamina[3#]

[#]*Studentschool of computing, KARE, vkdeepa94@gmail.com*
[#]*student, school of computing,KARE, jenifarulraj@gmail.com*
[#]*Professorschool of computing, KARE, jpamina8@gmail.com*

ABSTRACT

With the increasing usage of the Web and other text application areas, the demands in both text mining and natural language processing (NLP) have been increasing. Researchers in text mining have hoped that NLP—the attempt to extract a fuller meaning representation from free text—can provide useful improvements to text mining applications of all kinds[1]. The primary goal of Natural language processing (NLP) is to implement within computers the skill to understand a normal human language or natural language. It is related to the field of computer- human interaction. One of the motivations of NLP is for the society whose access to web information is obstructed simply by their inability to use the key-board and operating system. Natural language comes under the domain of artificial intelligence with the goal of understanding and creating meaningful expressions in human language. Artificial intelligence is the capability of a machine to imitate intelligent human behavior[2] .Therefore NLP uses Artificial Intelligence and is used to recover information in data mining. This paper presents a survey on Natural language processing (NLP) and its approaches.

Keywords: Natural Language Processing, Artificial Intelligence, Data Mining, Text Mining, Human Language.

I. INTRODUCTION

The concept of natural language processing is to develop and computer systems that can analyse, understand and synthesise natural human languages. Natural language falls within the domain of artificial intelligence with the goal of understanding and creating meaningful expressions in the human language.

There are some terminologies used in NLP:

- Morphology – It is a study of the construction of words from primitive meaningful units
- Syntax − It is used to arrange the words to make a sentence.
- Semantics − It is concerned with the meaning of words and how to combine words into meaningful phrases and sentences.
- Discourse − It deals with how the immediately preceding sentence can affect the interpretation of the next sentence
- .Pragmatics − It is used to understand the sentences in different situations and how the interpretation of the sentence is affected.

There are general 5 steps in NLP:

A. Lexical analysis Individual words are analyzed into their component and non-words such as punctuation are separated from the words. It involves identifying and analyzing the structure of words. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words.

B. Syntax analysis The purpose of syntax analysis is to check that a string of words is well-arranged and to break it up into a structure that shows the relationships between the different words. A syntactic analyzer performs this using a dictionary of word definitions (the lexicon) and a set of syntax rules (the grammar)[3].

C. Semantic Analysis It draws the exact meaning from the text. The text is checked for meaningfulness. The semantic analyzer disregards sentence such as “illegal law”. The structure created by syntax analyzer is assigned meaning.

D. Pragmatics

It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected. The structure representing what was said is reinterpreted to determine what was actually meant. Eg “close the door?” should have been interpreted as a request rather than an order.

E. Discourse

The meaning of an individual sentence may depend on the sentences preceding it and may influence the meanings of the sentences that follow it. .Eg the word “it” in the sentence “she wanted it” depends upon the prior discourse context.

2. TEXT MINING

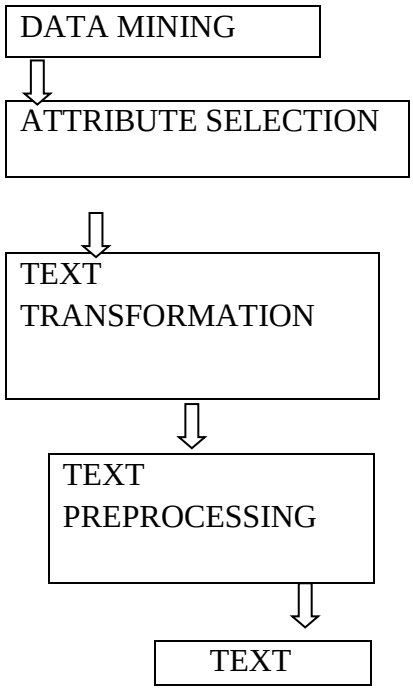
The goal of text mining is to discover relevant information in a text by transforming the text into data that can be used for further analysis. Text mining accomplishes this through the use of a variety of analysis methodologies; natural language processing (NLP) is one of them[4]. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information . It is the process to uncover the hidden information.

Text mining

one of the techniques of data

mining, is an analysis of data contained in natural language text

Text mining is the process of producing high-quality information from unstructured data. Text mining also is known as intelligent text analysis, text data mining or knowledge discovery in the text (KDT). Text mining is used to convert the unstructured data into structured data or meaningful information. The data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets, whereas the data that refers to the information that doesn't reside in a traditional row-column database is called unstructured data. Text Mining uses Natural Language Processing to increase the efficiency of mining.



A. Text mining process in text mining process, raw text is collected in the form of documents. Further mining process consists of following sub-process:

Text Pre-processing: - In this process, the unwanted text or letters or phrase such as punctuation mark, were get cleaned up and tokenization is formed.

Text Transposition: - In this process text representation and the required text selection is performed.

Attribute Selection: -This process comprises of removal of the unwanted attributes and reduces the dimensionality.

Data Mining: - Different data mining algorithms will perform to categorize the text.

Interpretation & Evaluation: - The derived text are interpreted and analysed and then the whole process will either terminate or iterate.

3.APPROACHES USED IN NLP

Liang's four categories of approaches to semantic analysis in NLP

A. Distributional It employs large-scale statistical tactics of Machine Learning and Deep Learning. Distributional approaches include the large-scale statistical tactics of machine learning and deep learning. These methods typically turn content into word vectors for mathematical analysis and perform quite well at tasks such as part-of-speech tagging (is this a noun or a verb?), dependency parsing (does this part of a sentence modify another part?), and semantic relatedness (are these different words used in similar ways?). These NLP tasks don't rely on understanding the meaning of words, but rather on the relationship between words themselves.

Such systems are broad, flexible, and scalable[5]. They can be applied widely to different types of text without the need for hand-engineered features or expert-encoded domain knowledge. The downside is that they lack a true understanding of real-world semantics and pragmatics. Comparing words to other words or words to sentences, or sentences to sentences can all result in different outcomes.

B. Frame—Based "A frame is a data-structure for representing a stereotyped situation," explains Marvin Minsky in his seminal 1974 paper called "A Framework for Representing Knowledge." Think of frames as a canonical representation for which specifics can be interchanged. Liang provides the example of a commercial transaction as a frame. In such situations, you typically have a seller, a buyers, goods being exchanged, and an exchange price. Sentences that are syntactically different but semantically identical – such as "Cynthia sold Bob the bike for $200" and "Bob bought the bike for $200 from Cynthia" – can be fit into the same frame. Parsing then entails first identifying the frame being used, then populating the specific frame parameters – i.e. Cynthia, $200.The obvious downside of frames is that they require supervision. In some domains, an expert must create them, which limits the scope of frame-based approaches. Frames are also necessarily incomplete. Sentences such as "Cynthia visited the bike shop yesterday" and "Cynthia bought the cheapest bike" cannot be adequately analysed with the frame we defined above.

C. Model-Theoretical Approach

The third category of semantic analysis falls under the model-theoretical approach. To understand this approach, we'll introduce two important linguistic concepts: "model theory" and "compositionality". Model theory refers to the idea that sentences refer to the world, as in the case with grounded language (i.e. the block is blue). In compositionality, meanings of the parts of a sentence can be combined to deduce the whole meaning. Liang compares this approach to turning language into computer programs. To determine the answer to the query "what is the largest city in Europe by population", you first have to identify the concepts of "city" and "Europe" and funnel down your search space to cities contained in Europe. Then you would need to sort the population numbers for each city you've shortlisted so far and return the maximum of this value.

D. Interactive Learning

Paul Grice, a British philosopher of language, described language as a cooperative game between speaker and listener. Liang is inclined to agree. He believes that a viable approach to tackling both breadth and depth in language learning is to employ interactive, interactive environments where humans teach computers gradually. In such approaches, the pragmatic needs of language inform the development.

4.CONCLUSION

Natural language processing is a branch of artificial intelligence & computer science and it uses text mining to make the interaction between human and computer, though its purpose is to have interaction among natural language of human beings and computers. Current research in NLP is shown more interest in learning different algorithms which are based on unsupervised learning.

REFERENCES

[1].Achlioptas, P., Sch¨olkopf, B., and Borgwardt, K. (2011). Two-locus association mapping in subquadratic time. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pages 726–734.

[2].Azencott, C., Grimm, D., Sugiyama, M., Kawahara, Y., and Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics, 29(13):171–179.

[3]Becker, C., Hagmann, J., M¨uller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the arabidopsis thaliana methylome. Nature, 480(7376):245–249.

[4].Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[5].Borgwardt, K. M. (2013). Machine Learning in Computational Biology. Machine Learning Summer School 2013, T¨ubingen, Germany.

[6].Borgwardt, K. M. and Kriegel, H. (2005). Shortest-path kernels on graphs. In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA, pages 74–81.