A Comparative Study based on Natural Language Processing Techniques

Koyel Datta Gupta
Department of Computer Science & Engineering
Maharaja Surajmal Institute of technology

Abstract- The concept of natural language processing (NLP) is based on artificial intelligence and linguistics. The aim of NLP lies is to relate natural language of human beings and machine learning. There are diverse challenges in the elucidation of natural language through artificial intelligence. The study of natural language processing has been carried on in the areas like generation and interpretation of natural language, morphological partitioning and machine translation, tagging of speech, sentimental analysis of speech, and optical character identification. Recently unsupervised learning algorithms are being applied profusely in NLP. These unsupervised analyse information which is not physically annotated as well as fusion of non-annotated & annotated data. With the availability of large amount of non annotated information, it is essential to explore unsupervised algorithms which otherwise produces less precise results. The objective of this paper is to present a comparative study of different natural language processing techniques.

Keywords – natural language processing; morpheme; corpus.

I. INTRODUCTION

The science of natural language processing has the prime objective of communicating between natural language used by humans and computers [1]. It deals with the interpretation of human language by machines. However, there are diversified issues in this field including perceiving human natural languages by machines. To address all the issues the tasks of natural language processing has been categorised as machine translation, anaphora, analysis of discourse, generation and understanding of natural language, lemmatisation, , morphological separation, part of speech tagging, relationship extraction, recognition of speech and analysis of sentiments, terminology extraction and many more.

In the initial phase of research on NLP the main focus was on Machine Translation [2]-[4]. From the year 1954, the Journal of Mechanical Translation started publishing about the initial research work in this domain. The first phase of NLP research was very challenging and difficult because of the non-availability of high speed computers and high level programming languages. The main course of work was dictionary based word processing for resolving semantic and syntactic ambiguity. In the second era of NLP based study was focussed on Artificial Intelligence based semantic interpretation [5]-[7]. The third stage showed development of grammatical theory by linguists and use of logic for knowledge representation [8]-[9]. In the late 90's, statistical approach [10] became the most significant area of research.

Most techniques proposed for NLP are supervised in nature and requires annotated data for processing. However, in the real world the availability of annotated data is sparse. Hence, in recent times the research work in the field of natural language processing is oriented mainly around unsupervised algorithms which can process plain data (not annotated) or a mixture of annotated and plain data. The performance of these algorithms, nonetheless show less precise results.

II. SOME IMPORTANT DEFINITIONS

It is essential to define some important terms related to natural language processing

i. **Ambiguity:** The term indicates that a word, phrase or sentence in a particular context may convey multiple meaning.
ii. **Corpus:** The word means body of text.
iii. **Discourse:** The term defines a sequence of sentences generated by one or more person for communicating.
iv. **Morpheme:** The word indicates to a meaningful morphological piece (which cannot be partitioned any further) of any given language.
v. **Lexicalization:** In automatic text generation, lexicalization refers to the production of a suitable lexical article for specified semantic content.
vi. **Synset:** The term refers to a set of words that are considered to be synonyms in particular context.

vii. **Word-Net:** The word points to a list of words with all associated meanings for every context.

When process is loaded into the node then it is sent to the cluster task executor. All details of that process sent to the cluster manager. It stores the information about the nodes, IP address of node, number of executing tasks, number of completed tasks and interrupted tasks. Cluster manager server checks the availability of space and then it perform the load balancing and if there are interrupt occur then reallocation resource, which are going to store into the database. In cluster manager server there are four blocks present i.e. node load record which load the records of node then task analyzer if interrupt is occur that time it will be activate and third one is process scheduler it schedules the process and last one is resource allocation it allocates the resource to the process. In this system resend block is present there. It resends the processes by using TTL and leave count policy. Result analyzer verifies the output of process. Here we are going to use process scheduling algorithm for clustering. Here we are going to calculate the cluster efficiency by showing the charts.

III. STUDY OF SOME RECENT ALGORITHMS ADAPTED FOR NATURAL LANGUAGE PROCESSING

A. Morphological Splitting Technique

This technique is meant for partitioning words into distinct morphemes and identifying its corresponding meaning in a corpus. The authors [11] propose an unsupervised technique to study the compound pieces and morphological procedures essential to split compounds into corresponding compound pieces. The scheme applies a bilingual corpus to train the morphological techniques needed to partition a compound into pieces. In addition, "monolingual corpora" are applied to train and filter the compound candidate sets.

B. Word Sense Disambiguation Technique

The concept of word sense disambiguation states that a single word may have multiple meanings depending upon its context in the corpus. It is difficult to identify correct sense of a word used in a particular context in a given sentence. For these problems Word-Net can be referred. In this paper [12], presents a method to approximate sense distributions for small questions. The authors club the concept of predicting word senses in a corpus with a new approach of including word senses in language modeling and further study the incorporation of Synsets.

C. Machine Translation Technique

Machine Translation involves the process of conversion of text automatically present in one form of natural language to any other language. The notion of lexicalization is important here. In the work [13] authors have used semantic role labels to enhance a "string-to-tree translation" method. The work is based on "predicate-argument structure". For each verb, the authors generate rules which represent the complete predicate argument structure making the syntax to semantics flexible. In [14] a "two-level alignment model" is designed to discriminate between morphemes and words. To achieve this model, within an HMM based word alignment model, an IBM Model 1 is embedded.

D. Speech Tagging and Recognition

Many languages show ambiguity where a word can act as a noun or a verb depending upon its context in a corpus. Part of Speech tagging is essential to identify this ambiguity.

On the contrary speech recognition involves identifying textual message of any speech by hearing to sound clip of a discourse. There is an existence of co-articulation in many natural languages, where the pronunciation of consecutive words mix with each other making recognition very hard.

In the paper [15], a deep neural network is presented that is trained to perform character-level representation of words and relate them with standard word depiction to execute part of speech tagging. The technique is applied on English and Portuguese.

The work [16] presents a speech recognition technique which utilizes "Mel Frequency Cepstrum Coefficients" (MFCC), Vector Quantization an HMM. The authors illustrate the method of speech recognition which is preceded by speaker recognition technique. MFCC is used for extraction of characteristics of every word pronounced by any speaker and then feature vectors are quantized before applying HMM to recognize the word.

E. Generation of Natural Language

Generation of natural language engages conversion of data stored knowledge databases into human readable natural language form. Identification of semantics using lexicalization is crucial. The paper [17] proposes a "statistical language generator based on a semantically controlled Long Short-term Memory (LSTM) structure". The

LSTM generator uses simple cross entropy training principle to learn from unaligned information by optimizing both surface understanding and sentence arrangement.

A brief description of the contribution of the individual work is given in Table I.

TABLE I. STUDY OF RECENT ALGORITHMS ADAPTED FOR NATURAL LANGUAGE PROCESSING

Researchers	*Year*	*Approach*	*Contribution*
Klaus Macherey et al.	2011	Morphological Splitting	A language-independent method for Splitting is proposed.
Z. Zhong et al.	2012	Word Sense Disambiguation	The proposed method annotates meaning to words in short queries and also attempts to incorporate senses into a language model for retrieving information.
M. Bazrafshan et al.	2013	Machine Translation	Two approaches are designed to integrate semantic role labels in a string-to-tree machine translation system.
E. Eyigŏz et al.	2013	Machine Translation	Two-level association model for morphologically enriched languages are presented.
Cicero Dos Santos et al.	2014	Speech Tagging	The novice idea of using convolutional neural networks for character-level feature extraction and using them with word-level features is presented.
Suma Swamy et al.	2013	Speech Recognition	MFCC is used for speaker recognition and HHM model is applied for speech recognition.
T. H. Wen et al.	2015	Generation of Natural Language	A neural network based generator proficient in producing natural linguistically diversified responses is presented.

IV.CONCLUSION

Researchers have been studying various aspects of Natural language processing over the years. In recent times it is observed that deep neural network is specifically applied for diverse NLP problems. The paper tries to capture the essence of few NLP based research works in current years. A brief description of individual's work is provided underlining their respective contributions. The problem of NLP especially generation of natural languages and sentimental analysis is extremely difficult and requires more attention.

REFERENCES

[1] Manaris, B. "Natural Language Processing: A Human–Computer Interaction Perspective," Appears in Advances in Computers , Academic Press, Vol. 47, 1998, pp. 1-66.

[2] Locke,W.N. and Booth. A.D. "Machine Translation of Languages,"John Wiley, 1995.

[3] Plath. W. "Multiple Path Analysis and Automatic Translation," in Booth 1967, pp 267-315.

[4] Minsky. M. "Semantic Information Processing," MIT Press,1968.

[5] Rustin. R. " Natural anguage Processing," Allgorithimics Press, 1973.

[6] Winograd.T. " A procedural model of language understanding," Readings in natural language processing, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1986

[7] Woods . W.A. " Semantics and quantification in natural language question answering," Readings in natural language processing, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1986

[8] Schank. R.C. "Language and memory, Readings in natural language processing," Morgan Kaufmann Publishers Inc., San Francisco, 1986.

[9] Small. S.L., Cottrell. G.W., Tanennaus. M.K., "Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence," Morgan Kaufmann Publishers Inc, San Francisco, CA, 1988.

[10] Manning. C.D. , Schuetze . H. "Foundations of Statistical Natural Language Processing," MIT Press, Cambridge, 1999.

[11] Macherey. K., Dai. A.M., Talbot. D., Popat A.C., Och. F., "Language-independent compound splitting with morphological operations," Proceedings *of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp.1395-1404.

[12] Zhong. Z. , Ng H. W. , "Word Sense Disambiguation Improves Information Retrieval," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 273–282, 8-14 July 2012.

[13] Bazrafshan, M. , Gildea. D."Semantic Roles for String to Tree Machine Translation". In: *ACL 2013*.

[14] Eyigŏz. E., Gildea, D., & Oazer, K. "Simultaneous word-morpheme alignment for statistical machine translation," Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 32 -40.

[15] Santos. C.D. , Zadrozny. B. , "Learning Character-level Representations for Part-of-Speech Tagging ," Proceedings of The 31st International Conference on Machine Learning, pp. 1818–1826, 2014

[16] Swamy S., Ramakrishnan. K.V., "An Efficient Speech Recognition System," Computer Science & Engineering: An International Journal, Vol. 3, No. 4, August 2013,pp 21-27.

[17] Wen. T.-H., Gasic. M., Mrksic, N., Su, P.-H., Vandyke, D., and Young, S. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. Empirical Methods on Natural Language Processing (EMNLP) (2015).