

A Review Paper on “NLP”

Associate Prof. and Hod. Neeraj Prakash Shrivastava, Jiny Jain

Department of CSE,

Anand ICE, Jaipur-302012, Rajasthan, India;

Neeraj.shrivastava@anandice.ac.in, jinyjain21031999@gmail.com

Abstract – Natural language processing is a branch of computer science and artificial intelligence which is concerned with interaction between computers and human languages. Natural language processing is the study of mathematical and computational modelling of various aspects of language and the development of a wide range of systems. These includes the spoken language systems that integrate speech and natural language. Natural language processing has a role in computer science because many aspects of the field deal with linguistic features of computation. Natural language processing is an area of research and application that explores how computers can be used to understand and manipulates natural language text or speech to do useful things. The applications of Natural language processing include fields of study, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, and expert system.

Keywords–Deep learning, CLIR,Natural Language.

I. INTRODUCTION

Natural Language Processing (NLP) is that field of computer science which consists of interfacing computer representations of information with natural languages used by humans. It examines the use of computers in understanding and manipulating the natural language text and speech. Over the past years, a lot of research has been done in the field of NLP. Some of the recent works have been discussed here. Kumarana et al. (2011) have developed a multilingual content creation tool for Wikipedia. Optimal Search for Minimum Error Rate Training has been discussed by Michel and Chris (2011). Associating Web Queries with Strongly-Typed Entities [Patrick et al., 2011], Linguistic Style Accommodation in Social Media [Cristian et al., 2011], Predicting the Importance of Newsfeed Posts and Social Network Friends [Tim et al., 2010], Wiki BABEL: A System for Multilingual Wikipedia Content [Kumaran et al., 2010], The utility of article and preposition error correction systems for English language learners: Feedback and Assessment [Martin et al., 2010]. The work presented in this Section has been previously published [Khan, Dar and Quadri, 2012]

1.1 Theoretical developments in NLP

Theoretical developments in NLP can be grouped into following classes:

- (i) statistical and corpus-based methods in NLP
- (ii) use of WordNet for NLP research
- (iii) use of finite-state methods in NLP.

1.1.1 Statistical Methods

The models and methods used in solving NLP problems are broadly classified into two types: deterministic and

stochastic. A mathematical model is called deterministic if it does not involve the concept of probability; otherwise it is said to be stochastic. A stochastic model can be probabilistic or statistical, if its representation is from the theories of probability or statistics, respectively [Edmundson, 1968]. Statistical methods are used in NLP for a number of purposes, e.g., speech recognition, part-of- speech tagging, for generating grammars and parsing, word sense disambiguation, and so on. There has been a lot of research in these areas. Geoffrey Zweig and Patrick Nguyen (2009) have proposed a segmental conditional random field framework for large vocabulary continuous speech recognition [Geoffrey and Patrick 2009]. Gerasimos Potamianos, Chalapathy Neti, Ashutosh Garg, Guillaume Gravier and Andrew W. Senior (2003) have reviewed Advances in the Automatic Recognition of Audio- Visual Speech and have presented the algorithms demonstrating that the visual modality improves automatic speech recognition over all conditions and data considered [Gerasimos et al., 2003]. Raymond J. Mooney has developed a number of machine learning methods for introducing semantic parsers by training on a corpus of sentences paired with their meaning representations in a specified formal language [Raymond, 2007]. Marine CARPUAT and Dekai WU (2007) have shown that statistical machine translation can be improved by using word sense disambiguation. They have shown that if the predictions of the word sense disambiguation system are incorporated within a statistical machine translation model then the translation quality is consistently improved [Marine and Dekai, 2007].

1.1.2 Use of WordNet for NLP research

Mihalcea & Moldovan (1999) have proposed the use of WordNet to make the outcome of statistical analysis of natural language texts better. WordNet or the electronic

dictionary is developed at Princeton University. It is a large database that serves as an important NLP tool consisting of nouns, verbs, adjectives and adverbs. These are arranged in the form of synonym sets (synsets). Each set represents one underlying lexical concept. These sets are linked with each other by means of conceptual-semantic and lexical relations. There are different wordnets for about 50 different languages, but they are not complete like the original English WordNet [Gerard and Gerhard, 2009]. WordNet is now used in a number of NLP research and applications. One of the most important applications of WordNet in NLP is EuroWordNet developed in Europe. EuroWordNet is a multilingual database which consists of WordNets for the European languages. It has been structured in the same way as the WordNet for English. A methodology for the automatic construction of a large-scale multilingual lexical database has been proposed where words of many languages are hierarchically organized in terms of their meanings and their semantic relations to other words. This database is capable of organizing over 800,000 words from over 200 languages, providing over 1.5 million links from words to word meanings. This universal wordnet has been derived from the Princeton WordNet. Lars Borin and Markus Forsberg have given a comparison between WordNet and SALDO. SALDO is a Swedish lexical resource which has been developed for language technology applications [Lars and Markus, 2009]. Japanese WordNet currently has 51,000 synsets with Japanese entries. Methods for enhancing or extending the Japanese Wordnet have been discussed. These include: increasing the cover, linking it to examples in corpora and linking it to other resources. In addition various plans have been outlined to make it more useful by adding Japanese definition sentences to each synset [Francis et al., 2009]. The use of WordNet in multimedia information retrieval has also been discussed and the use of external knowledge in a corpus with minimal textual information has been investigated. The original collection has been expanded with WordNet terms in order to enrich the information included in the corpus and the experiments have been carried out with original as well as expanded topics [Manuel et al., 2011]. A Standardized Format for Wordnet Interoperability [Claudia et al., 2009] has been given i.e., WordNet-LMF. The main aim of this format is to provide the WordNet with a format representation that will allow easier integration among resources sharing the same structure (i.e. other wordnets) and, more importantly, across resources with different theoretical and implementation approaches.

1.1.3. Use of finite state methods in NLP

The finite-state automation is the mathematical tool used to implement regular expressions – the standard notation for characterizing text sequences. Different applications of the Finite State methods in NLP have been discussed [Jurafsky and Martin, 2000; Kornai, 1999; Roche and Shabes, 1997]. From past many years the finite state

methods have been used in presenting various research studies on NLP. The FSMNLP workshops are the main forum of the Association for Computational Linguistics (ACL) Special Interest Group on Finite-State Methods (SIGFSM) [Anssiet et al., 2011].

II. NLP SOFTWARE

A number of NLP software packages and tools have been developed, some of which are available for free, while others are available commercially. These tools have been broadly classified into different types some of which are mentioned here. General Information Tools (e.g. Sourcebank – a search engine for programming resources., The Natural Language Software Registry), Taggers and Morphological Analyzers (e.g. A Perl/Tk text tagger, AUTASYS – which is a completely automatic English Wordclass analysis system, TreeTagger – a language independent part-of-speech tagger, Morphy – which is a tool for German morphology and statistical part-of-speech tagging), Information Retrieval & Filtering Tools (e.g. Rubryx: Text Classification Program, seft – a Search Engine For Text, Isearch – software for indexing and searching text documents, ifile – A general mail filtering system, Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering), Machine Learning Tools (e.g. Machine Learning Toolbox (MLT), The Machine Learning Programs Repository), FSA Tools (e.g. FSA Utilities: A Toolbox to Manipulate Finite-state Automata), HMM Tools (e.g. Hidden Markov Model (HMM) Toolbox, Discrete HMM Toolkit, A HMM mini-toolkit), Language Modeling Tools (e.g. Maximum Entropy Modeling Toolkit, Trigger Toolkit, Language modeling tools), Corpus Tools (e.g. WebCorp, Multext: i.e. Multilingual Text Tools and Corpora, TACT- i.e. Text Analysis Computing Tools, Textual Corpora and Tools for their Exploration). Some more tools include DR-LINK (Document Retrieval using LINGuistic Knowledge) system demonstrating the capabilities of NLP for Information Retrieval [Liddy et al, 2000], NLPWin: an NLP system from Microsoft that accepts sentences and delivers detailed syntactic analysis, together with a logical form representing an abstraction of the meaning [Elworthy, 2000]. Waldrop (2001) has described the features of three NLP software packages, viz. Jupiter: a product of the MIT research Lab that works in the field of weather forecast, Movieline: a product of Carnegie Mellon that talks about local movie schedules, and MindNet from Microsoft Research, a system for automatically extracting a massively hyperlinked web of concepts.

III. APPLICATIONS

There are a number of applications of NLP e.g. machine translation, natural language text processing and summarization, user interfaces, multilingual and cross

language information retrieval (CLIR), speech recognition, and expert systems, and so on. In this paper we discuss automatic abstracting and information retrieval.

1. Automatic Abstracting

Automatic abstracting or text summarization is a technique used to generate abstracts or summaries of texts. Due to the increase in the amount of online information, it becomes very important to develop the systems that can automatically summarize one or more documents [Dragomir et al., 2002]. The main aim of summarization is to differentiate between the more informative or important parts of the document and the less ones [Dipanjan and Andre, 2007]. According to Radev et al. (2002) a summary can be defined as piece of text that can be produced from one or more texts in a way such that it conveys important information in the original text(s), and whose size is not more than half of the original text(s) and mostly significantly less than that". The summary can be of two types i.e. abstraction or extraction. Abstract summary is one in which the original documents' contents are paraphrased or generated, whereas in an extract summary, the content is preserved in its original form, i.e., sentences [Krystaet al, 2007]. Extracts are formed by using the same words, sentences of the input text, while abstracts are formed by regenerating the extracted content. Extraction is the process of identifying the important contents in the text while in abstraction the contents are regenerated in new terms. When the summaries are produced from a single document, it is called single document summarization. Multidocument summarization has been defined as a process of producing a single summary from a number of related documents. A lot of research has been done on automatic abstracting and text summarization. Zajicetal [David et al., 2008] have presented single-document and multi- document summarization techniques for email threads using sentence compression. They have shown two approaches to email thread summarization i.e. Collective Message Summarization (CMS) and Individual Message Summarization (IMS). NeATS [Chin and Eduard, 2002] is a multidocument summarization system in which relevant or interesting portions about some topic are extracted from a set of documents and presented in coherent order. NetSum [Krystaet al, 2007] is an approach to automatic summarization based on neural networks. Its aim is to obtain those features from each sentence which helps to identify its importance in the document. A text summarization model has been developed which is based on maximum coverage problem and its variant [Hiroya and Manabu, 2009]. In this some decoding algorithms have been explored such as a greedy algorithm with performance guarantee, a randomized algorithm, and a branch-and-bound method. A number of studies have been carried out on text summarization. An efficient linear time algorithm for calculating lexical chains has been developed for preparing automatic summarization of documents [Silber and McCoy, 2000]. A method of

automatic abstracting has been proposed that integrates the advantages of both linguistic and statistical analysis. Jin and Dong-Yan (2000) have proposed a methodology for generating automatic abstracts that provides an integration of the advantages of methods based on linguistic analysis and those based on statistics [Songand Zhao, 2000].

2. Information Retrieval

Information retrieval (IR) is concerned with searching and retrieving documents, information within documents, and metadata about documents. It is also called document retrieval or text retrieval. IR concerns with retrieving documents that are necessary for the users' information. This process is carried out in two stages [Jun and Jianhan, 2009]. The first stage involves the calculation of the relevance between given user information need and the documents in the collection. In this stage probabilistic retrieval models that have been proposed and tested over decades are used for calculating the relevance to produce a "best guess" at a document's relevance. In the second stage the documents are ranked and presented to the user. In this stage the probability ranking principle (PRP) [Cooper, 1971] is used. According to this principle the system should rank documents in order of decreasing probability of relevance. By using this principle the overall effectiveness of an IR system maximizes.

There has been a lot of research in the field of information retrieval. Some of the recent developments are included here. ChengXiangZhai (2008) has given a critical review of statistical language models for information retrieval. He has systematically and critically reviewed the work in applying statistical language models to information retrieval, summarized their contributions, and pointed out outstanding challenges [ChengXiang, 2008]. Nicholas J. Belkin has identified and discussed few challenges for information retrieval research which come under the range of association with users [Nicholas, 2008]. An efficient document ranking algorithm has been proposed that generalizes the well-known probability ranking principle (PRP) by considering both the uncertainty of relevance predictions and correlations between retrieved documents [Jun and Jianhan, 2009]. Michael et al have discussed the various problems, directions and future challenges of content-based music information retrieval [Michael et al., 2008]. A unified framework has been proposed that combines the modeling of social annotations with the language modeling-based methods for information retrieval [Ding et al., 2008].

IV. CURRENT AND FUTURE PROGRESS OF NLP

Some of the active researches on NLP phenomena include the Syntactic phenomena: those that pertain to the

structure of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning (e.g. discriminative models for scoring parses, coarse to fine efficient approximate parsing, dependency grammar); Machine translation (e.g. models and algorithms, low- resource and morphological complex language); Semantic

phenomena : those that pertain to the meaning of a sentence relatively independent of the context in which the language occurs (e.g. sentiment analysis, summarization, information extraction ,slot-filling, discourse analysis, textual entailment); Pragmatic phenomena such as Speech: those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue) or, non-linguistic (such as knowledge about person who produced the language, about goals of the communication, about the objects in the current visual field, etc. (e.g. language modelling-syntax and semantics, models of acoustics, pronunciation). Speech recognition and information retrieval have finally gone commercial and there is a ton of text and speech on the Internet, cell phones, etc. It is now clear that studies regarding anything about a language are possible, e.g. formalizing some insights e.g. discrete knowledge (what is possible) and continuous knowledge (what is likely); studying the formalism mathematically; developing and implementing algorithms and testing on real data. The current and on-going future changes or improvements which need to be done to NLP are: to add features to existing interfaces, back end processing should be fully implemented (e.g. information extraction and normalization to build databases. Another anticipated improvement is of having hand held devices with translators and personal conversation recorder with topical searches.

V. CHALLENGES AND FAILURES

Church and Rau points out that even though we should know better, it is so appealing to fantasize about intelligent computers that understand human communication, that hyperbole is practically unavoidable. Sometimes these practices work out for the best. Symantec, for example, a highly successful vendor of software tools for the PC, started with a product called Q&A, an NLP program for querying a database. The Q&A was successful because of its unique packaging of AI/NLP with a good simple database facility. Neither would have been successful in isolation. The AI/NLP generated initial sales, but the real value was in the database. People bought the product because they were intrigued with the AI/NLP technology, but most users ended up turning off the AI/NLP features . But all too often excessive optimism results in a manic-like cycle of euphoric activity followed by severe depression. In 1954, Georgetown University demonstrated what would now be called a “toy” system. It was designed to translate

a small corpus of approximately 50 Russian sentences into English. Little if any attempt was made to generalize to sentences beyond the tiny test corpus.

The limitations of today’s practical language processing technology have been summarized by Bobrow and Weischedel as follows:

1.Current systems have limited discourse capabilities that are almost ex-clusively handcrafted. Thus current systems are limited to viewing interaction, translation, and writing text as processing a sequence of either isolated sentences or loosely related paragraphs. Consequently, the user must adapt to such limited discourse.

2.Domains must be narrow enough so that the constraints on the relevant semantic concepts and relations can be expressed using current knowledge presentation techniques, i.e., primarily in terms of types and sorts. Processing may be viewed abstractly as the application of recursive tree re-writing rules, including filtering out trees not matching a certain pattern.

3.Handcrafting is necessary, particularly in the grammatical components of systems (the component technology that exhibits least dependence on the application domain). Lexicons and axiomatizations of critical facts must be developed for each domain, and these remain time-consuming tasks.

VI. CONCLUSION

As a computerized approach of analyzing text, NLP is continually striving forward. Researchers are continually trying to gather knowledge on how human beings understand and use various languages. This aid in the development of appropriate tools and techniques which make computer systems understand and manipulate natural languages to perform the various tasks. Technologies, such as string matching, keyword search, glossary look up are now on the past as, to more forward looking technologies such as grammar checkers, conceptual search, event extraction, interlingual on going and striving forward.

REFERENCES

- [1]. E.D. Liddy, Natural Language Processing, 2001.
- [2]. N. Kaur¹, V. Pushe and R Kaur, “Natural Language Processing Interface for Synonym”, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pp. 638-642 ,ISSN 2320–088X.
- [3]. S. Vijayarani¹, J. Ilamathi and Nithya, “Preprocessing Techniques for Text Mining - An Overview”, International Journal of Computer Science & Communication Networks, Vol.5, issue.1, pp. 7-16 7 ISSN: 2249-57891] E.D. Liddy, Natural Language Processing, 2001. IJREAS VOLUME 6, ISSUE 3 (March, 2016) (ISSN 2249-3905)

- [4]. International Journal of Research in Engineering and Applied Sciences (IMPACT FACTOR – 6.573)
International Journal of Research in Engineering & Applied Sciences
- [5]. Email:- editorijrim@gmail.com,
<http://www.euroasiapub.org> 209
- [6]. L.Liddy, E. Hovy, J.Lin, J.Prager, D. Radev, L.Vanderwende, R.Weischedel, “Natural Language Processing”, This report is one of five reports that were based on the MINDS workshops.
- [7]. [5]G.Chowdhury, “Natural language processing”, Annual Review of Information Science and Technology, 2003, 37. pp. 51-89, ISSN 0066-4200.
- [8]. S. Jusoh and H.M. Alfawareh, “Natural language interface for online sales”, in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007),Malaysia:IEEE, November 2007, pp. 224-228
- [9]. E.K. Ringger, R.C. Moore, E. Charniak, L. Vanderwende, and H Suzuki, “Using the Penn Treebank to Evaluate Non-Treebank Parsers”, In Proceedings of the 2004 Language Resources and Evaluation Conference (LREC), 2004, Lisbon, Portugal.
- [10].T. Winograd, Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, 1971, MIT-AI-TR-235
- [11].W. A. Woods, “Transition Network Grammars for Natural Language Analysis”, Communications of the ACM 13:10, 1970.
- [12].W.C. Mann & S. Thompson, “Rhetorical Structure Theory: Toward a Functional Theory of Text Organization”, 1988. Text 8 (3). Pp. 243-281.
- [13].E. Charniak, K. Knight, and K.Yamada, “Syntax-based Language Models for Statistical Machine Translation”. In Proceedings of MT Summit IX, 2003.