International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

A REVIEW ON NATURAL LANGUAGE PROCESSING

Neha Sharma[1], Sonika Jindal[2]

[1,2] *Computer science , Shaheed Bhagat Singh Technical Campus,Ferozepur, (India)*

ABSTRACT

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. In this paper NLP interface, various techniques for synonym are described.

Keywords: ***NLP Networks, Information Retrieval, Natural Language Interfaces***

I INTRODUCTION

Natural Language Processing (NLP) is a general term for a wide range of tasks and methods related to automated understanding of human languages. In recent years, the amount of available diverse textual information has been growing rapidly, and specialized computer systems can over ways of managing, sorting, filtering and processing this data more anciently. As a larger goal, research in NLP aims to create systems that can also `understand' the meaning behind the text, extract relevant knowledge, organize it into easily accessible formats, and even discover latent or previously unknown information using inference. For example, the _eld of biomedical research can benefit from various text mining and information extraction techniques, as the number of published papers is increasing exponentially every year, yet it is vital to stay up to date with all the latest advancements. Research in Machine Learning (ML) focuses on the development of algorithms for automatically learning patterns and making predictions based on empirical data, and it offers

useful approaches to many NLP problems. Machine learning techniques are commonly divided into three categories:

Supervised learning methods make use of labelled training data to build models that can generalise to unseen examples. These include many well known learning algorithms, such as support vector machines (Cortes &Vapnik, 1995; Joachims, 1998), conditional random fields (Lafferty et al., 2001), probabilistic neural networks (Specht, 1990) and random forests (Breiman, 2001).

Semi-supervised systems normally require smaller amounts of labeled data, and also make use of some unlabelled corpora. This can be achieved by bootstrapping the system with a small set of annotated examples and iteratively finding more from the unlabelled data (Agichtein&Gravano, 2000; Thelen&Riloff, 2002), or by propagating the labels of known examples to unseen instances using some similarity metric (Zhu &Ghahramani, 2002; Chen et al., 2006).

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 03, March 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

Unsupervised learning methods aim to find a hidden structure in the provided dataset, without using any explicit labelling information. Due to the restrictions of the problem, this usually reduces to some form of clustering, such as k-means (Lloyd, 1982), hierarchical clustering (Johnson, 1967), or self-organising maps (Kohonen, 1990).

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks. The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems, and so on.

II THE NETWORKS

All the NLP tasks above can be seen as tasks assigning labels to words. The traditional NLP approachis: extract from the sentence a rich set of hand-designed features which are then fed to astandard classification algorithm, for example, a Support Vector Machine (SVM), often with a linearkernel. The choice of features is a completely empirical process, mainly based first on linguisticintuition, and then trial and error, and the feature selection is task dependent, implying additionalresearch for each new NLP task.

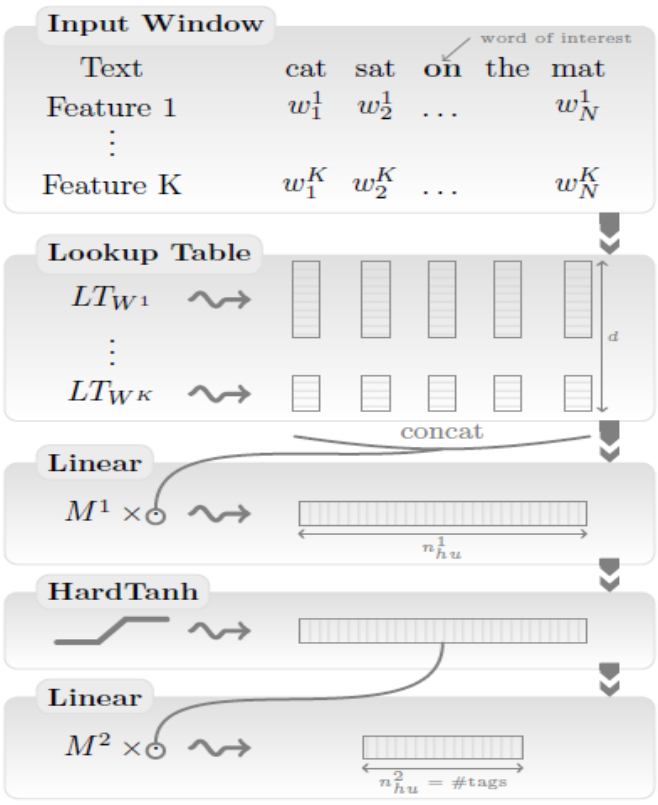

Figure 1: Window approach network.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

complex features (e.g., extracted from a parse tree) which can impact the computational cost whichmight be important for large-scale applications or applications requiring real-time response.Instead, we advocate a radically different approach: as input we will try to pre-process ourfeatures as little as possible and then use a multilayer neural network (NN) architecture, trained inan end-to-end fashion. The architecture takes the input sentence and learns several layers of featureextraction that process the inputs. The features computed by the deep layers of the network areautomatically trained by backpropagation to be relevant to the task. We describe in this section ageneral multilayer architecture suitable for all our NLP tasks, which is generalizable to other NLPtasks as well.

Natural Language Understanding

At the core of any NLP task there is the important issue of natural language understanding. The process of building computer programs that understand natural language involves three major problems: the first one relates to the thought process, the second one to the representation and meaning of the linguistic input, and the third one to the world knowledge. Thus, an NLP system may begin at the word level – to determine the morphological structure, nature (such as part-of-speech, meaning) etc. of the word – and then may move on to the sentence level – to determine the word order, grammar, meaning of the entire sentence, etc.— and then to the context and the overall environment or domain. A given word or a sentence may have a specific meaning or connotation in a given context or domain, and may be related to many other words and/or sentences in the given context.

Liddy and Feldman suggest that in order to understand natural languages, it is important to be able to distinguish among the following seven interdependent levels, that people use to extract meaning from text or spoken languages:

• phonetic or phonological level that deals with pronunciation
• morphological level that deals with the smallest parts of words, that carry a meaning, and suffixes and prefixes
• lexical level that deals with lexical meaning of words and parts of speech analyses
• syntactic level that deals with grammar and structure of sentences
• semantic level that deals with the meaning of words and sentences
• discourse level that deals with the structure of different kinds of text using document structures and
• pragmatic level that deals with the knowledge that comes from the outside world, i.e., from outside the contents of the document.

A natural language processing system may involve all or some of these levels of analysis.

Natural Language Text Processing Systems:

Manipulation of texts for knowledge extraction, for automatic indexing and abstracting, or for producing text in a desired format, has been recognized as an important area of research in NLP. This is broadly classified as the area of natural language text processing that allows structuring of large bodies of textual information with a view to retrieving particular information or to deriving knowledge structures that may be used for a specific purpose. Automatic text processing systems generally take some form of text input and transform it into an output of some different form. The central task for natural language text processing systems is the translation of

ISSN (O) 2454-9665
ISSN (P) 2454-0663

potentially ambiguous natural language queries and texts into unambiguous internal representations on which matching and retrieval can take place (Liddy, 1998). A natural language text processing system may begin with morphological analyses. Stemming of terms, in both the queries and documents, is done in order to get the morphological variants of the words involved. The lexical and syntactic processing involve the utilization of lexicons for determining the characteristics of the words, recognition of their parts-of-speech, determining the words and phrases, and for parsing of the sentences.

Information Retrieval

Information retrieval has been a major area of application of NLP, and consequently a number of research projects, dealing with the various applications on NLP in IR, have taken place throughout the world resulting in a large volume of publications. Lewis and Sparck Jones comment that the generic challenge for NLP in the field of IR is whether the necessary NLP of texts and queries is doable, and the specific challenges are whether non-statistical and statistical data can be combined and whether data about individual documents and whole files can be combined. They further comment that there are major challenges in making the NLP technology operate effectively and efficiently and also in conducting appropriate evaluation tests to assess whether and how far the approach works in an environment of interactive searching of large text files. Feldman suggests that in order to achieve success in IR, NLP techniques should be applied in conjunction with other technologies, such as visualization, intelligent agents and speech recognition.

Arguing that syntactic phrases are more meaningful than statistically obtained word pairs, and thus are more powerful for discriminating among documents, Narita and Ogawa use a shallow syntactic processing instead of statistical processing to automatically identify candidate phrasal terms from query texts. Comparing the performance of Boolean and natural language searches, Paris and Tibbo found that in their experiment, Boolean searches had better results than freestyle (natural language) searches. However, they concluded that neither could be considered as the best for every query. In other words, their conclusion was that different queries demand different techniques.

Pirkola shows that languages vary significantly in their morphological properties. However, for each language there are two variables that describe the morphological complexity, viz., index of synthesis (IS) that describes the amount of affixation in an individual language, i.e., the average number of morphemes per word in the language; and index of fusion (IF) that describes the ease with which two morphemes can be separated in a language. Pirkola shows that calculation of the ISs and IFs in a language is a relatively simple task, and once they have been established, they could be utilized fruitfully in empirical IR research and system development.

Chandrasekar&Srinivas propose that coherent text contains significant latent information, such as syntactic structure and patterns of language use, and this information could be used to improve the performance of information retrieval systems. They describe a system, called Glean, that uses syntactic information for effectively filtering irrelevant documents, and thereby improving the precision of information retrieval systems.

A number of tracks (research groups or themes) in the TREC series of experiments deal directly or indirectly with NLP and information retrieval, such as the cross-language track, filtering track, interactive track, question-answering track, and the web track. Reports of progress of the NLIR (Natural Language Information Retrieval) project are available in the TREC reports (Perez-Carballo&Strzalkowski,;Strzalkowski. et al..,). The major goal

ISSN (O) 2454-9665
ISSN (P) 2454-0663

of this project has been to demonstrate that robust NLP techniques used for indexing and searching of text documents perform better compared to the simple keyword and string-based methods used in statistical full-text retrieval (Strzalkowski, T. et al..,). However, results indicate that simple linguistically motivated indexing (LMI) did not prove to be more effective than well-executed statistical approaches in English language texts. Nevertheless, it was noted that more detailed search topic statements responded well to LMI compared to terse one-sentence search queries. Thus, it was concluded that query expansion, using NLP techniques, leads to a sustainable advances in IR effectiveness (Strzalkowski et al..,).

Natural Language Interfaces

A natural language interface is one that accepts query statements or commands in natural language and sends data to some system, typically a retrieval system, which then results in appropriate responses to the commands or query statements. A natural language interface should be able to translate the natural language statements into appropriate actions for the system.

A large number of natural language interfaces that work reasonably well in narrow domains have been reported in the literature (for review of such systems see Chowdhury, 1999b, Chapter 19;Haas, ; Stock,).

Much of the efforts in natural language interface design to date have focused on handling rather simple natural language queries. A number of question answering systems are now being developed that aim to provide answers to natural language questions, as opposed to documents containing information related to the question. Such systems often use a variety of IE and IR operations using NLP tools and techniques to get the correct answer from the source texts. Breck et al. report a question answering system that uses techniques from knowledge representation, information retrieval, and NLP. The authors claim that this combination enables domain independence and robustness in the face of text variability, both in the question and in the raw text documents used as knowledge sources. Research reported in the Question Answering (QA) track of TREC (Text Retrieval Conferences) show some interesting results. The basic technology used by the participants in the QA track included several steps. First, cue words/phrase like ‘who’ (as in ‘who is the prime minister of Japan’), ‘when’ (as in ‘When did the Jurassic period end’) were identified to guess what was needed; and then a small portion of the document collection was retrieved using standard text retrieval technology. This was followed by a shallow parsing of the returned documents for identifying the entities required for an answer. If no appropriate answer type was found then best matching passage was retrieved. This approach works well as long as the query types recognized by the system have broad coverage, and the system can classify questions reasonably accurately. In TREC-8, the first QA track of TREC, the most accurate QA systems could answer more than 2/3 of the questions correctly. In the second QA track (TREC-9), the best performing QA system, the Falcon system from Southern Methodist University, was able to answer 65% of the questions. These results are quite impressive in a domain-independent question answering environment. However, the questions were still simple in the first two QA tracks. In the future more complex questions requiring answers to be obtained from more than one documents will be handled by QA track researchers.

Owei argues that the drawbacks of most natural language interfaces to database systems stem primarily from their weak interpretative power which is caused by their inability to deal with the nuances in human use of natural language. The author further argues that the difficulty with NL database query languages (DBQLs) can

ISSN (O) 2454-9665
ISSN (P) 2454-0663

be overcome by combining concept based DBQL paradigms with NL approaches to enhance the overall ease-of-use of the query interface.

Zadrozny et al. suggest that in an ideal information retrieval environment, users should be able to express their interests or queries directly and naturally, by speaking, typing, and/or pointing; the computer system then should be able to provide intelligent answers or ask relevant questions. However, they comment that even though we build natural language systems, this goal cannot be fully achieved due to limitations of science, technology, business knowledge, and programming environments. The specific problems include (Zadrozny et al., 2000):

• Limitations of NL understanding;

• Managing the complexities of interaction (for example, when using NL on devices with differing bandwidth);

• Lack of precise user models (for example, knowing how demographics and personal characteristics of a person should be reflected in the type of language and dialogue the system is using with the user), and

• Lack of middleware and toolkits.

III CONCLUSION

This paper carries out various approaches for network construction in NLP. Paper also describes about various interfaces and techniques for NLP.

REFERENCES

[1] Privacy, Security, and Data Mining, pp.1-8, 2002.] Han Jiawei, M. Kamber, and Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp.1-40, 2006.

[2] V.S.Verykios, E.Bertino, I.N.Fovino, L.P.Provenza, Y.Saygin, Y.Theodoridis, “State-of-the-art in Privacy Preserving Data Mining”, New York, ACM SIGMOD Record, vol.33, no.2,Pp.50-57, 2004.

[3] N. Zhang, “Privacy-Preserving Data Mining”, Texas A&M University, pp.19-25, 2006.

[4] R. Agrawal, R. Srikant, “Privacy-Preserving Data Mining”, ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.

[5] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, “Privacy Preserving Mining of Association

[6] Rules”, Information System, vol.29, no.4, pp.343-364, 2004.

[7] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.

[8] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland,USA, pp.37-48, 2005.

[9] D. Agrawal, C.C. Aggarwal, “On the Design and Quantification of Privacy Preserving Data Mining Algorithms”, In Proceedings of the 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems, pp.247-255, 2001.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

[10] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp.217-228, 2002.

[11] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[12] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", J. Am. Stat. Assoc., vol.60, no.309, pp.63-69, 1965.

[13] S.J. Rizvi, J.R. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th VLDB conference, pp.1-12, 2002.

[14] W. Du, Z. Zhan, "Using Randomized Response Techniques for Privacy Preserving Data Mining", In Proceedings 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.505-510, 2003.

[15] Guo, S. Guo, X. Wu, "Privacy Preserving Market Basket Data Analysis", In Proceedings the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases,Pp.103-114, 2007.

[16] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.557-570, 2002.

[17] R. Bayardo, R. Agrawal, "Data Privacy Through Optimal k-Anonymization", In Proceedings the 21st International Conference on Data Engineering, pp.217-228, 2005.

[18] K. Lefevre, J. Dewittd, R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity", In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Pp.49-60, 2005.

[19] B. Fung, K. Wang, P. Yu, "Top-down Specialization for Information and Privacy Preservation", In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.

[20] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.571-588, 2002.