International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

RESEARCH ARTICLE

Natural Language Processing Interface for Synonym

Navdeep Kaur[1], Vandana Pushe [2], Rupinderdeep Kaur[3]

Student of master Technology ,Department of CSE, Desh Bhagat University, Mandi Gobindgarh, Punjab, India
Assistant Professor Department of CSE, Desh Bhagat University, Mandi Gobindgarh, Punjab, India
Assistant Professor Department of CSE, Thapar University Patiala, Punjab, India
[1] navdeep.panag384@gmail.com, [2] vandanapushe@gmail.com, [3] rupinderdeep@thapar.edu

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— Machine learning, Multilayer neural network, Thought process, Linguistic input, Sublanguage analyses

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 cover ways of managing, sorting, filtering and processing this data more efficiently. 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 field 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 labeled training data to build models that can generalize 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).
Unsupervised learning methods aim to find a hidden structure in the provided dataset, without using any explicit labeling 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-organizing 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 approach is: extract from the sentence a rich set of hand-designed features which are then fed to a standard classification algorithm, for example, a Support Vector Machine (SVM), often with a linear kernel.[1] The choice of features is a completely empirical process, mainly based first on linguistic intuition, and then trial and error, and the feature selection is task dependent, implying additional research for each new NLP task. Complex tasks like SRL then require a large number of possibly
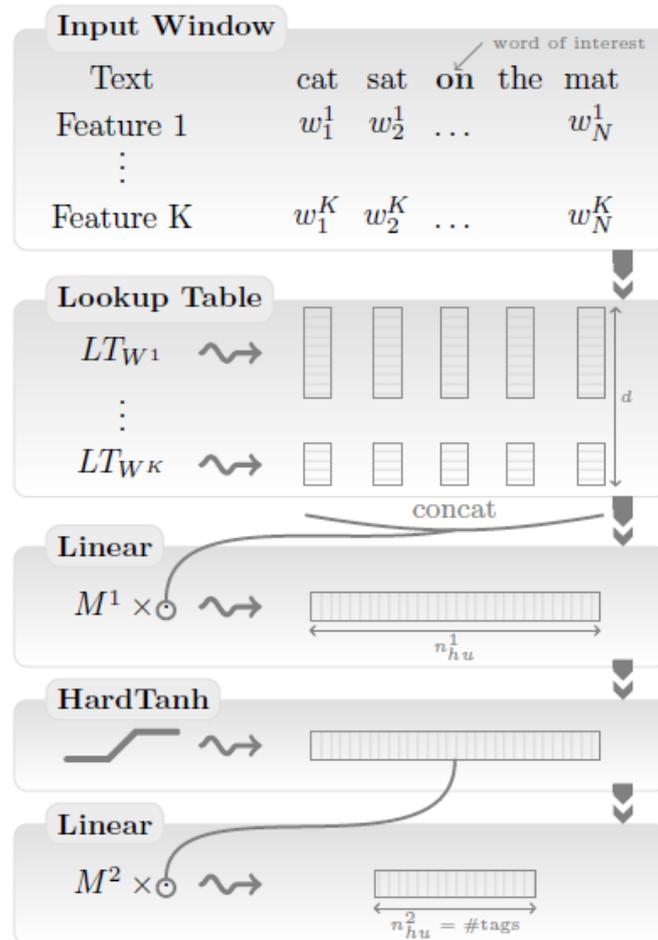

Figure 1: Window approach network.

complex features (e.g., extracted from a parse tree) which can impact the computational cost which might be important for large-scale applications or applications requiring real-time response.[2] Instead, we advocate a radically different approach: as input we will try to pre-process our features as little as possible and then use a multilayer neural network (NN) architecture, trained in an end-to-end fashion. The architecture takes the input sentence and learns several layers of feature extraction that process the inputs. The features computed by the deep layers of the network are automatically trained by back propagation to be relevant to the task. We describe in this section a general multilayer architecture suitable for all our NLP tasks, which is generalizable to other NLP tasks as well.[3]

A. 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. [4]

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.

B. 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.[5] 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 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.

Past research concentrating on natural language text processing systems has been reviewed by Haas, Mani & Maybury, Smeaton, and Warner. Some NLP systems have been built to process texts using particular small sublanguages to reduce the size of the operations and the nature of the complexities. These domain-specific studies are largely known as 'sublanguage analyses' (Grishman & Kittredge. Some of these studies are limited to a particular subject area such as medical science, whereas others deal with a specific type of document such as patent texts.

C. 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.[5]

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.[6]

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.

Khoo et al. report an experiment that investigates whether information obtained by matching cause-effect relations expressed in documents with the cause-effect relations expressed in user queries can be used to improve results in document retrieval compared with the use of only the keywords without considering the relations. Their experiment with the Wall Street Journal full text database revealed that causal relations matching where either the cause or the effect is a wildcard can be used to improve information retrieval effectiveness if the appropriate weight for each type of matching

can be determined for each query. However, the authors stress that the results of this study were not as strong as they had expected it to be.[7]

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 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..,).

D. 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.[8] 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.[8] 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 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. We describe the various techniques of machine learning and seven independent levels to extract meaning from text or spoken languages. Paper also describes about various interfaces and techniques for NLP.

REFERENCES

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research (JMLR)*, 6:1817–1953, 2005.

[2] R. M. Bell, Y. Koren, and C. Volinsky. The BellKor solution to the Netflix Prize. Technical report, AT&T Labs, 2007. http://www.research.att.com/˜volinsky/netflix.

[3] Y. Bengio and R. Ducharme. A neural probabilistic language model. In *Advances in Neural Information Processing Systems (NIPS 13)*, 2001.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks.In *Advances in Neural Information Processing Systems (NIPS 19)*, 2007.

[5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conferenceon Machine Learning (ICML)*, 2009.

[6] L. Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nˆımes*. EC2,1991.

[7] L. Bottou. Online algorithms and stochastic approximations. In David Saad, editor, *Online Learningand Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

[8] L. Bottou and P. Gallinari. A framework for the cooperation of learning algorithms. In *Advances in Neural Information Processing Systems (NIPS 3)*. 1991.

[9] L. Bottou, Y. LeCun, and Yoshua Bengio. Global training of document processing systems using graph transformer networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 489–493, 1997.

[10] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman Souli´e and J. H´erault, editors, *Neurocomputing: Algorithms, Architectures and Applications*, pages 227–236. NATO ASI Series, 1990.

[11] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992a.

[12] P. F. Brown, V. J. Della Pietra, R. L. Mercer, S. A. Della Pietra, and J. C. Lai. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–41, 1992b.

[13] C. J. C. Burges, R. Ragno, and Quoc Viet Le. Learning to rank with nonsmooth cost functions. In *Advances in Neural Information Processing Systems (NIPS 19)*, pages 193–200. 2007.

[14] Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

[15] O. Chapelle, B. Schlkopf, and A. Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., USA, September 2006.

[16] E. Charniak. A maximum-entropy-inspired parser. In *Conference of the North American Chapter of the Association for Computational Linguistics & Human Language Technologies (NAACL-HLT)*, pages 132–139, 2000.