Towards Understanding Theoretical Developments in Natural Language Processing

Mehnaz Khan
Research Scholar
Department of Computer Science
University of Kashmir

Dr. Mehraj-ud-Din Dar
Director IT & SS
University of Kashmir

Dr. S.M.K. Quadri
Director
Department of Computer Science
University of Kashmir

ABSTRACT

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. The main aim of the researchers in this field is to collect the necessary details about how natural languages are being used and understood by humans. They use these details to develop the tools for making the computers understand and manipulate the natural languages to perform the desired tasks. In this paper we describe some of the theoretical developments that have influenced research in NLP. We also discuss automatic abstracting and information retrieval in natural language processing applications. We conclude with a discussion on Natural Language Interfaces, NLP software and the future research in NLP.

General Terms

Natural language processing, Artificial Intelligence.

Keywords

Automatic abstracting, information retrieval, interfaces.

1. INTRODUCTION

Artificial intelligence (AI)[1] is a computer science discipline which deals with the study and creation of intelligent computer systems i.e. the systems that show some form of intelligence. These systems are able to learn new concepts and can deduce useful conclusions. Natural Language Processing is a field of AI that consists of analyzing how computers can be used to understand and manipulate natural language text or speech to do useful things. Emphases are being laid on communicating with the systems in a natural form. Therefore it is essential to develop programs to understand natural language. These programs should be able to interface with the humans in a natural way, and the most essential medium for that purpose is the natural language.

2. RECENT WORK IN NLP

Over the past years, a lot of research has been done in the field of NLP. Some of the recent issues include: Multilingual Content Creation Tool for Wikipedia[40], Optimal Search for Minimum Error Rate Training[41], Associating Web Queries with Strongly-Typed Entities[42], Linguistic Style Accommodation in Social Media[43], Predicting the Importance of Newsfeed Posts and Social Network Friends[44], Wiki BABEL: A System for Multilingual Wikipedia Content[45], The utility of article and preposition error correction systems for English language learners: Feedback and Assessment[46].

3. THEORITICAL DEVELOPMENTS

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.

3.1 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[2]. 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 [3]. 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[4]. 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 [5]. 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 [6].

3.2 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 is a large lexical database that has been developed at Princeton University. It’s also called an electronic dictionary. It is an important NLP tool which consists of English nouns, verbs, adjectives and adverbs organized into sets of cognitive synonym sets (synsets), each representing one underlying lexical concept. Synsets are interlinked 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[12]. 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[12]. 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[13]. 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[14]. 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[15]. A Standardized Format for Wordnet Interoperability [16] 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.

3.3 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 [17, 18, 19]. 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)[20].

4. NLP 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.

4.1 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[7]. The main aim of summarization is to differentiate between the more informative or important parts of the document and the less ones[11]. Radev et al.[10] have defined a summary as “a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually 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[8]. 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[7]. When the summaries are produced from a single document, it is called single document summarization. Multidocument summarization is the process of producing a single summary of a set of related source documents[7]. A lot of research has been done on automatic abstracting and text summarization. Zajic etal[21] 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[9] 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[8] 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[22]. 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[23]. 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 [24].

4.2 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 [26]. 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) [25] 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. ChengXiang Zhai (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[27]. Nicholas J. Belkin has identified and discussed few challenges for information retrieval research which fall within the scope of association with users[28]. 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[26]. Michael et al have discussed the various problems, directions and future challenges of content-based music information retrieval [29]. A unified framework has been proposed that combines the modeling of social annotations with the language modeling-based methods for information retrieval[30].

.

5. NLP INTERFACES

A natural language interface accepts commands in natural language and sends data to the system which then provides the appropriate responses to the commands. A natural language interface translates the natural language statements into appropriate actions for the system. A large number of natural language interfaces have been developed[31]. 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. These systems use a variety of IE and IR operations to get the correct answer from the source texts. In information retrieval and NLP, question answering (QA) is the task of automatically answering a question posed in natural language. To find the answer to a question, a QA computer program may use either a pre-structured database or a collection of natural language documents. Unlike information retrieval systems(Internet search engines), QA systems do not retrieve documents, but instead provide short, relevant answers located in small fragments of text. That is why QA systems are significantly slower and require more hardware resources than information retrieval systems[32]. QA track of TREC (Text Retrieval Conference) have shown some interesting results. Several steps were included in the technology used by the participants in the QA track. First, words like ‘who’, ‘when’ 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. In TREC-8, the first QA track of TREC, the most accurate QA systems could answer more than 2/3 of the questions correctly[34]. 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[33]. In the first two QA tracks the questions were simple. In TREC 2001 QA track, which was the third running of a QA track in TREC, several new conditions were added to increase the realism and difficulty of the task[35]. The TREC 2002 track repeated the main and list tasks from 2001, but with the major difference of requiring systems to return exact answers. The change to exact answers was motivated by the belief that a system’s ability to recognize the precise extent of the answer is crucial to improving question answering technology[36]. These runnings of QA track have been carried out every year till date by adding different conditions to make the QA tracks more realistic.

6. 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 – A Fully Automatic English Wordclass Analysis System, TreeTagger – a language independent part-of-speech tagger, Morphy – An integrated 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: Multilingual Text Tools and Corpora, Text Analysis Computing Tools (TACT), 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 [37], 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 [38]. 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[39].

7. FUTURE RESEARCH ISSUES IN NLP

NLP is an emerging field of research. There are a number of open research issues in various areas of NLP. One of the most important among those is machine learning. Although a lot of research has been done in this field however there are some issues that need to be solved e.g. Interactive Learning for Domain Adaptation Scenarios, Active Learning for Aggregating Structures.

8. ACKNOWLEDGEMENT

The authors are thankful to the faculty, Department of Computer Science, University of Kashmir for their constant support.

9. REFERENCES

[1] Patterson, D. W. Introduction to Artificial Intelligence and expert systems. University of Texas

[2] Edmundson, H. P. 1968. Mathematical Models in Linguistics and Language Processing.

[3] Geoffrey, Z., and Patrick, N. 2009. A Segmental CRF Approach to Large Vocabulary Continuous Speech Recognition. Microsoft Research, Redmond, WA.

[4] Gerasimos, P., Chalapathy, N., Guillaume, G., Ashutosh, G., and Andrew, W. S. 2003. Recent Advances in the Automatic Recognition of Audio-Visual Speech. In Proceedings of the IEEE, 91(9), September 2003.

[5] Raymond, J. M. 2007. Learning for Semantic Parsing. Computational Linguistics and Intelligent Text Processing: Proceedings of the 8th International Conference, CICLing 2007, Springer, Berlin, Germany, 311-324.

[6] Marine, C., and Dekai, W. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. Department of Computer Science and Engineering, University of Science and Technology, Clear Water Bay, Hong Kong, 61-72.

[7] Dragomir, R. R., Kathleen, M., and Eduard, H. 2002. Introduction to special issue on summarization. Association for Computational Linguistics, 28(4), 399-408.

[8] Krysta, M. S., Lucy, V., and Christopher, J.C. 2007 Enhancing Single-document Summarization by Combining RankNet and Third-party Sources. Association for Computational Linguistics. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 448–457, Prague, June 2007.

[9] Chin, Y. L., and Eduard, H. 2002. From Single to Multi-document Summarization: A Prototype System and its Evaluation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, (July 2002), 457-464.

[10] Radev, D. R., Hovy, E., and McKeown, K. 2002. Introduction to the special issue on summarization. Computational Linguistics. 28(4):399-408. [1, 2]

[11] Dipanjan, D., and Andre, F. T. M. 2007. A Survey on Automatic Text Summarization. Language Technologies Institute, Carnegie Mellon University.

[12] Gerard, M., and Gerhard, W. 2009. Towards a Universal WordNet by Learning from Combined Evidence. CIKM'09, November 2–6, 2009, Hong Kong, China.

[13] Lars, B., and Markus, F. 2009. All in the Family: A Comparison of SALDO and WordNet. Sprakbanken, University of Gothenburg, Sweden.

[14] Francis, B., Hitoshi, I., Sanae, F., Kiyotaka, U., Takayuki, K., and Kyoko, K. 2009. Enhancing the Japanese WordNet. Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, 1-8.

[15] Manuel, C. D. G., Maria, T. M. V., Alfonso, U. L., and Jose, M. P. O. 2011. Using WordNet in Multimedia Information Retrieval. Workshop. Springer-Verlag Berlin Heidelberg. 185-188.

[16] Claudia, S., Monica, M., and Piek, V. 2009. WordNet-LMF: Fleshing out a Standardized Format for WordNet Interoperability. CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

[17] Jurafsky, D., and Martin, J.H. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. Upper Saddle River, NJ: Prentice Hall.

[18] Kornai, A. 1999. Extended Finite State Models of Language (Studies in Natural Language Processing), Cambridge University Press.

[19] Roche, E., and Shabes, Y. 1997. Finite-State Language Processing (Language, Speech and Communication), MIT Press.

[20] Anssi, Y. J., Andras, K., and Jacques, S. 2011. Finite-state methods and models in natural language processing. Natural Language Engineering 17 (2): 141–144.

[21] David, M. Z., Bonnie, J. D., and Jimmy, L. 2008. Single-Document and Multi-Document Summarization Techniques for Email Threads Using Sentence Compression. College of Information Studies, University of Maryland.

[22] Hiroya, T., and Manabu, O.2009. Text Summarization Model based on Maximum Coverage Problem and its Variant. Proceedings of the 12th Conference of the European Chapter of the ACL, 781–789, Athens, Greece, 30 March – 3 April 2009.

[23] Silber, H.G., and McCoy, K. F. (2000) Efficient text summarization using lexical chains In: H. Lieberman(Ed.). Proceedings of IUI 2000 International Conference on Intelligent User Interfaces, 9-12 Jan. 2000, New Orleans, LA. New York: ACM. 252-255.

[24] Song, J. and Zhao, D.Y. (2000). Study of automatic abstracting based on corpus and hierarchical dictionary. Journal of Software,11, 308-14.

[25] W. S. Cooper. The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. University of California, Berkeley, 1971.

[26] Jun, W., and Jianhan, Z. 2009. Portfolio Theory of Information Retrieval. SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

[27] ChengXiang, Z. 2008. Statistical Language Models for Information Retrieval A Critical Review. Foundations and Trends in Information Retrieval 2(3):137–213.

[28] Nicholas, J. B. 2008. Some(what) Grand Challenges for Information Retrieval. ACM SIGIR FORUM. 42(1):47-54 June 2008.

[29] Michael, A. C., Remco, V., Masataka, G., Marc, L., Christophe, R., and Malcolm, S. Content-Based Music Information Retrieval: Current Directions and Future Challenges. Proceedings of the IEEE 96(4):668-696, April 2008.

[30] Ding, Z., Jiang, B., Shuyi, Z., Hongyuan, Z., and Lee, G. Exploring Social Annotations for Information Retrieval. ACM 2008, Beijing, China.

[31] Stock, O. 2000. Natural language processing and intelligent interfaces. Annals of Mathematics and Artificial Intelligence, 28, 39-41.

[32] Mihai, S., Dan, I. M., and Sanda, M. H. Performance Analysis of a Distributed Question/Answering System. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 13(6):579-596.

[33] Voorhees, E. 2000. The TREC-9 question answering track report. [Online] Available: http://trec.nist.gov/pubs/trec9/papers/qa-report.pdf

[34] Voorhees, E. 1999. The TREC-8 question answering track report. [Online] Available: http://trec.nist.gov/pubs/trec8/papers/qa-report.pdf

[35] Ellen, M. V. Overview of the TREC 2001 Question Answering Track. National Institute of Standards and Technology, Gaithersburg, MD 20899.

[36] Ellen, M. V. Overview of the TREC 2002 Question Answering Track. National Institute of Standards and Technology, Gaithersburg, MD 20899.

[37] Liddy, E.., Diamond, T., and McKenna, M. (2000). DR-LINK in TIPSTER III. Information Retrieval, 3, 291-311.

[38] Elworthy, D. (2000). Question answering using a large NLP system. The Ninth Text Retrieval Conference (TREC 9).

[39] Waldrop, M.M (2001). Natural language processing, Technology Review, 104, 107-108.

[40] Kumarana, Narend, Ashwani, S., and Vikram, D. (2011) WikiBhasha: Our Experiences with Multilingual Content Creation Tool for Wikipedia, in Proceedings of Wikipedia Conference India, November 2011, Wikimedia Foundation.

[41] Michel, G., and Chris, Q. Optimal Search for Minimum Error Rate Training, in Proc. of Empirical Methods in Natural Language Processing, July 2011.

[42] Patrick, P., and Ariel, F. Jigs and Lures: Associating Web Queries with Strongly-Typed Entities, in Proceedings of Association for Computational Linguistics - Human Language Technology (ACL-HLT-11), June 2011.

[43] Cristian, D.N.M., Michael, G., and Susan, D. Mark My Words! Linguistic Style Accommodation in Social Media In Proceedings of WWW 2011, Hyderabad, India. ACM, 1 April 2011.

[44] Tim, P., Michael, G., Scott, C., David, M. C., and Aman, D. Predicting the Importance of Newsfeed Posts and Social Network Friends, American Association for Artificial Intelligence , July 2010.

[45] Kumaran, A., Naren, D., Ashok, B., Saravanan, K., Anil, A., Ashwani, S., Sridhar, V., Vidya, N., Vikram, D., and Sandor, M. WikiBABEL: A System for Multilingual Wikipedia Content, in in Proceedings of the 'Collaborative Translation: technology, crowd sourcing, and the translator perspective' Workshop (co-located with AMTA 2010 Conference), Denver, Colorado, Association for Machine Translation in the Americas, 31 October 2010.

[46] Martin, C., Michael, G., and Joel, T. The utility of article and preposition error correction systems for English language learners: Feedback and assessment , in Language Testing, Sage, July 2010.