International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:12, 2014

Role of Natural Language Processing in Information Retrieval; Challenges and Opportunities

Khaled M. Alhawiti

***Abstract*—**This paper aims to analyze the role of natural language processing (NLP). The paper will discuss the role in the context of automated data retrieval, automated question answer, and text structuring. NLP techniques are gaining wider acceptance in real life applications and industrial concerns. There are various complexities involved in processing the text of natural language that could satisfy the need of decision makers. This paper begins with the description of the qualities of NLP practices. The paper then focuses on the challenges in natural language processing. The paper also discusses major techniques of NLP. The last section describes opportunities and challenges for future research.

***Keywords*—**Data Retrieval, Information retrieval, Natural Language Processing, Text Structuring.

I.Introduction

NATURAL Language Processing (NLP) aims to develop a computer program that could parse natural human language. The goal of the processing is to enable communication between humans and computers and vice versa. NLP occurs in multiple stages that can be implemented in an application. The stages may include parsing, part of speech tagging, and recognition of entities. The development of the program also requires expertise in computational linguistics. In particular, the skills needed in NLP include sentence understanding, probabilistic parsing and tagging, machine translation, grammar induction, automatic question answering, word sense disambiguation, text generation, speech generation, text clustering, and information retrieval [1].

There are seven principles that must be taken into consideration for the evaluation of NLP techniques. The first principle is to define a function that is objective and adopt a threshold that is optimal. The second principle is to cross validate the performance of various techniques of NLP. The third principle is to test the differences in performance for NLP techniques statistically. The fourth principle is to test equivalence among various NLP techniques statistically. The fifth principle is to analyze the impact of difficulty. The sixth principle is to adjust scores for the analysis of boundaries. The seventh principle is the validation of the random score [2].

II.Qualities of NLP Practices

Decision making in government and business organizations is dependent directly on information quality. The web has become the richest source for most of the information of business intelligence. The enterprise systems of searching the knowledge base may be developed based on ontologies or computing that is meaning-based. The textual information in these technologies is indexed and the knowledge base is tagged. However, tagging in the current web is not in a proper way semantically. Hence enterprise search methods do not result in meaningful information retrieval. There is a need of effective search methods for extracting the best and relevant information that could improve the decision making process [3].

One of the approaches of NLP is evidence-based NLP. It consists of three integrative iterative processes. The first step filters the search results to obtain a set of information that is relevant. It overcomes the limitations of the mechanisms of keyword and ranking. In the next step, the set of relevant information is applied to the concepts of grounded theory. In the third and final step, information quality is tested using evidential analysis [3].

In the web search that is conducted with traditional approach, several techniques are applied. These include building an index of the web content, building a database of the indices, and search for those keywords matching the database contents. The problem with this approach is that it does not support the acquisition of intelligence information. For example, the search engine of Google can find thousands of web pages and show top 1000 results quickly. However, the pages may not be related semantically. Duplicate content filtering of Google is regarded as the best technology in the world. However, the technology cannot analyze the text meanings. Due to this limitation, Google cannot omit results that are similar in semantics. Hence, the keyword matching technique may miss out important information. Also, ranking algorithm may prioritize search results that are irrelevant. It is also important to note that the keyword used in the search is provided by the author. It may not be a true representative of the text. Possible relevant information may be missed out in this approach [3].

The problems of the web search, which is conducted with traditional approach, can be overcome by a new approach called concept search. Concept search analyzes plain natural language text to extract information that resembles in conceptual terms to the information supplied in a search query. The approach works on the basis of ideas. Ideas expressed in the information are matched with the ideas presented in the query of the search [3].

The most important resource that mankind has is knowledge, i.e. information. In today's age of information, the efficient management of this knowledge depends on the use of

Khaled M. Alhawiti is with the Tabuk University, Tabuk, Saudi Arabia (phone: +9660555903348; e-mail: alhowity@hotmail.com).

all other natural, industrial and human resources.

Throughout the history of humanity knowledge, mostly communicates, is stored and managed in the form of natural language-Greek, Latin, English, Spanish, etc. The present age is no exception: the knowledge still exists and creating in the form of documents, books, articles, even if they are stored in electronic form, or digital. The breakthrough is that in this way, and computers can be a huge help in the processing of this knowledge.

To combat this, much effort is spent, especially in developed countries, the development of science that enables computers to understand the text. This science, depending on the practical versus theoretical approach, the degree to which it is expected to achieve understanding and other aspects has several names: natural language processing, text processing, language technology, computational linguistics. In any case, it is render text by their sense and not as a binary file.

The general scheme of most of the systems and methods involving processing language is as follows:

- First, the text is not processed directly but is transformed into a formal representation preserving their relevant characteristics for the job or the specific method (e.g., a set of strings of letters a database table, a set of logical predicates, etc.).
- Then the main program manipulates this representation, transforming according to the task, looking at it the necessary substructures, etc.
- Finally, if necessary, changes made to the formal representation (or response generated in this way) are transformed into natural language.

As the reader may have guessed, the complexity associated with natural language is particularly important when we need to retrieve textual information that satisfies the information need of a user. It is for this reason that in the area of Textual Information Retrieval NLP techniques are widely used, both to facilitate the description of the content of the documents, as to represent the query formulated by the user, and therefore with the aim of comparing both descriptions and present the user with documents that satisfy a greater extent their information needs.

In other words, a system of textual information retrieval performs the following tasks to answer the queries of a user:

- Indexing the document collection: in this phase, by applying NLP techniques, an index that contains descriptions of documents is generated. Normally, each document is described by the set of terms, hypothetically, best represents its content.
- When a user asks a query, the system analyzes it, and if necessary transforms, in order to represent the information need of the user in the same way that the contents of the documents.
- The system compares the description of each document with the description of the query, and presents the user those documents whose descriptions most closely resemble the description of your inquiry.
- The results are often shown in terms of their relevance, i.e., sorted according to the degree of similarity between the descriptions of documents and query.

III.Statistical Natural Language Processing

The statistical natural language processing represents the classic model of information retrieval systems, and is characterized in that each document is described by a set of keywords called index terms.[5]

This approach is very simple, and is based on what has been termed as "bag of words", the bag-of-words model is a simplifying demonstration used in natural language processing as well as information retrieval (IR). [10]

In this approach, all words in a document are treated as index terms for that document. In addition, a weight to each term according to their importance is attached, usually determined by their frequency of occurrence in the document. Thus, no consideration is taken in order, the structure, and meaning, etc. words. [11]

IV.Linguistic Processing of Natural Language

This approach is based on the application of different techniques and rules explicitly encoding linguistic knowledge. The documents are analyzed from different linguistic levels, as mentioned above, language tools that incorporate the text specific to each level annotations. The different steps to take are to perform a linguistic analysis of documents although this does not imply that apply to all systems. [12]

Morphological analysis is performed by taggers (taggers) that assign to each word its grammatical category from the identified morphological features. After identifying and analyzing the words that form a text, the next step is to see how they relate and combine to form larger units, phrases and sentences. Therefore, it is to perform text parsing. At this point grammars (parsers) that are descriptive language formalisms that aim to establish the syntactic structure of text apply. [13] The techniques used to implement and build grammars are varied and depend on the purpose for which the parsing is performed. In the case of information retrieval surface analysis, which identifies only the most significant structures used to apply: noun phrases, verb and prepositional phrases, entities, etc. This level of analysis is often used to optimize resources and does not slow the response time of the systems [14].

From the syntactic structure of the text, the next goal is to get the meaning of the sentences that compose it. It is about getting the semantic representation of sentences, from the composite elements. [15]

One of the tools used in semantic processing is based on WordNet lexical data. It is a semantic lexicon annotated in different languages, consisting of groups of synonyms called synsets of short definitions that are provided and the various semantic relations between these synonym groups are stored.

Since in an article no space to describe all the interesting applications of the techniques of natural language processing, we can only mention here that caught more attention or received further development in recent years.

As we have discussed in this article, the most valuable treasure of the human race-their knowledge and their culture is concentrated in large collections of texts (books, magazines, newspapers) written in natural language. Traditionally such collections are called libraries and played a unique role in dissemination and preservation of culture and knowledge. [16]

However, until now the technology keeping was rudimentary libraries: books stores were very basic support to find a book if the author and title and is known. The "return" of such diffusion of knowledge was very low: we can say that most of the information written in the books was not needed and by whom at the time when needed [17].

With the digital information processing utility libraries which in this case are called digital-libraries is increased until they become integrated services and complex cultural, scientific and technical information. Obviously intelligent search facilities provided by natural language technologies are only part of the overall solution, which also involves technical, administrative, legal, and cultural aspects. [6]

Another possibility that is opened with the appearance of large volumes of texts, which is also growing steadily creating specific databases through the information that, is communicated in the texts. For example, creating a database that keeps the tourist attractions for venues, dates and services, extracting this information automatically from the descriptions on the websites and advertising tourist companies, Or, a database of supply and demand for technological solutions that may be helpful for a consulting company. Obviously, this type of work requires some understanding of text by machine, although in a bounded domain [7]

Another way to filter the relevant information in the sea of irrelevant is the summary presentation, or (using an anglicized) summarization of information. It is analyzing a large text (or a large collection of texts), generate a short report of what these texts relevant to say, to give the reader an idea of its content without the need for it to read all the texts. [8]

A variant of summarization is the summarization of text theme: make a brief report on the issues (but not the ideas) to be discussed at a given text, for example, the text talks about war, politics and drug trafficking; the other speaks about science, technology and transportation. Despite lower richness of this presentation-compared with summaries completes- has some advantages: it is simpler to obtain and then gives more secure and stable results; also allows perform mathematical operations on sets (vectors) of subjects achieved.[9]

V.Text Mining

Text Mining predominantly uses techniques mainly focused in the fields of information retrieval, statistics, and machine learning. Its purpose usually is not to understand all or even a large part of what a given speaker/writer has said, but rather the extraction of patterns across a large number of documentation.

A.Intelligent Handling of Official Documents (E-Government)

Democratic societies tend to be too bureaucratic. This is because, first, the large number of documents circulating since each citizen gives effect to their rights to petition, appeal, review, etc., and second, the large number of people involved in the consideration of such documents so that the power of decision is not concentrated in the hands of one or a few people. On the other hand, this causes delays and disorder when the flow of documents exceeds the capabilities of the bureaucratic system [4].

An efficient solution to this problem, allowing democracy with efficiency is automatic document processing, at least in the areas of classification and distribution of document flow, searching for relevant documents, and similar, etc. For example, an automated system can rotate documents to relevant departments or officials. You can group documents describing similar cases for their joint account in a single meeting. You can provide the official search of similar cases in the past, in its opinion, to be left for consideration if a similar ruling could apply to the case in question. [18]

VI.Challenges in Natural Language Processing

The ultimate goal of NLP is the creation of a software program that could recognize human language and generate a language used by humans. The field is moving at an accelerating pace and much work has been done during the past 10 years. The two main components in NLP are Natural Language Generation (NLG) and Natural Language Understanding (NLU). For example, in an automated machine translation (MT) tool or an automated question answer tool, NLU and NLG play vital roles. NLU refers to a system that is used for the computation of meaning representation. The system is not dependent on the recognition of the speech. However, NLU can be used in speech recognition for transcribing an acoustic signal into a text. An understanding component interprets the text and extracts the meaning. In NLU, the two important components are syntactic analysis and semantic analysis. Syntactic analysis assigns a parse tree to a sentence of the source language. Semantic analysis translates the parse tree into a semantic representation that is unambiguous and precise representation of the source language sentence. For NLP community, semantic analysis is a challenging issue due to the involvement of very complex working [19].

NLG generates automated text of natural language by using techniques of linguistic representation of text. It also ensures the grammatical correctness of the generated text. It is accomplished through a syntactic analyzer. The analyzer ensures that grammatical rules are obeyed in the auto-generated text of natural language. There is also a text planner that is used for the coherent arrangement of sentences and paragraphs. Automated machine translation from one language to another language is an example of NLG. The application of the concept can be seen in Google Translate or the Translate feature in Microsoft Word. However, the present translation technology is unable to provide high quality translations. There is always a need of a human translator to edit the outputs from these systems. The translation technologies use two common methods; phrase alignment and word alignment.

The tools are not able of producing accurate meaning for ambiguous words. Also, the tools fail at many places to produce the correct structure of the sentence. Similarly, there are also rooms for improvements in the question answer automated tools [20].

VII. Major Techniques of NLP

The available NLP techniques can be classified into four dimensions. The categorization of each dimension is based on a variation point. The first classification is algebraic models. The second category is known as term extraction. The third classification is called weighting schema. The fourth category is known as similarity metric. Algebraic models are used for the evaluation of the semantic similarity among words. In the use of synonymies, the main issue is the computation of the level of similarity. In term extraction, the text fragments are preprocessed before the application of any comparisons. The preprocessing of text is performed by stop-word removal and tokenization. In tokenization, a series of tokens are generated for a given text. The terms not contributing to the semantics are removed. For example, articles and pronouns are removed. Stop words are defined before the preprocessing of the text. In term weighting, text terms are assigned weights. The assignment of weight is dependent on the occurrences of the terms in the analyzed text fragments. Weights may be assigned as raw frequency, binary, or term frequency. In similarity metric, a formula is used for computing the fraction of common words between two fragments of text [2].

VIII. Future Research Opportunities and Challenges

Natural language statements are gathered from various sources such as user feedbacks, documents, notes, and interview excerpts. The properties on natural language are difficult to prove due to the ambiguities of natural language. Hence, future research should focus on the expressing of informal language requirements as formal representations. Different levels of linguistic analysis are used in NLP systems. These include phonetic level, lexical level, morphological level, syntactic level, and semantic level. Domain ontology is also gaining acceptance in the context of concept identification. The goal of domain ontology is the creation of domain knowledge. The knowledge is built on structured concepts. The concepts are such that they have semantic relevance to each other [5].

The promise of NLP is that it supports users in the analysis, transformation, and creation of meaningful information from large amounts of text. However, there is one potential area that requires consideration in future research. It is the question that how NLP techniques can be used in conjunction with existing infrastructures of information systems. For example, how these techniques can be combined with web portals to bring measureable improvements to their users. NLP tools can be used to assist portal users in interpreting, transforming, and retrieving content. The history of NLP dates back to the decade of 50s. NLP was used in a variety of desktop applications. For example, the techniques were used for providing summaries of texts in word processor applications. The techniques were also used in the applications of scientific nature. These include protein annotation, extraction of diagnoses and clinical conditions from medical records, extraction of relationships between genes and cancer related drugs, and so forth. However, there is a little focus in the literature on the integration of NLP techniques with web-based systems aimed at facilitating a common web user [6].

There is a significant conceptual distance between the text of a natural language and its low-level logic formalism. Hence, the key idea is to bridge the gap between two scenarios. It can be achieved by incorporating the specification of temporal requirements into a high level language. The whole process can be split into two key tasks. The first step may involve representing natural text to a high level formal system. The step so far has only been made partially automated and requires human involvement. From the high level formalism, the process can be transferred to low level logic formalism. Reasoning about the requirements occurs in this process of conversion. [21] The approach needs to be researched in future studies, as it can help in the identification of inconsistencies in temporal requirements that are inherent in natural language text. As an example, inconsistencies in defining software requirements may be attributed to temporal constraints. The constraints affect the operations and controls of the software system. Hence, inconsistencies in requirements should be detected before the design phase to reduce the cost of faulty requirements [22]. The applications of natural language processing can also be applied to data mining and systems based on fuzzy logic. These algorithms parse the huge amount of data and provide meaningful information to the management for decision making.

References

[1] Nadial M., Panggabean E., and S. Meryana, "A Study of Parsing Process on Natural Language Processing in Bahasa Indonesia," *Computational Science and Engineering*, pp. 309-316, Dec. 2013.

[2] F. Davide., and C. Gerardo, "Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques," *IEEE Transactions on Software Engineering*, vol. 39, no. 1, pp. 18-44, Jan. 2013.

[3] D. Natalia, and S. David, "Application of Natural Language Processing and Evidential Analysis to Web-Based Intelligence Information Acquisition," *European Intelligence and Security Informatics Conference*, pp. 268-273, Aug. 2012

[4] J. Shaidah, and A. M. Hejab, "Automated Translation Machines: Challenges and a Proposed Solution," *International Conference on Intelligent Systems, Modelling and Simulation*, pp. 77-82, Jan. 2011.

[5] I. Mohd, and A. Rodina, "Class diagram extraction from textual requirements using Natural language processing (NLP) techniques," *International Conference on Computer Research and Development*, pp. 200-204, May 2010.

[6] B. Fedor, S. Bahar, W. Rene, M. Marie-Jean, and K. Birgitta, "Natural Language Processing for Semantic Assistance in Web Portals," IEEE Sixth International Conference on Semantic Computing, pp. 67-74, Sep. 2012.

[7] L. Wenbin. "Toward consistency checking of natural language temporal requirements," IEEE/ACM International Conference on Automated Software Engineering, pp. 651-655, Nov. 2011.

[8] Chris Buckley, Amit Singhal, Mandar Mitra, Gerard Salton. 1995. ``New Retrieval Approches Using SMART: TREC 4''. In Proceedings of TREC4.

[9] Strzalkowski, Tomek and Jose Perez-Carballo. 1994. "Recent Developments in Natural Language Text Retrieval." Proceedings of the

Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 123-136.

[10] Strzalkowski, Tomek, Jose Perez-Carballo and Mihnea Marinescu. 1996. “Natural Language Information Retirieval: TREC-4 Report.” Proceedings of the Third Text REtrieval Conference (TREC-4), NIST Special Publication 500-2xx.

[11] Brenner, Everett. "Beyond Boolean--New Approaches to Information Retrieval." National Federation of Abstracting and Information Services, Philadelphia, 1996

[12] Hayes, Philip J. and Gail Koerner. "Intelligent Text Technologies and Their Successful Use by the Information Industry". Proceedings of the Fourteenth National Online Meeting, 1993 : pp. 189-196.

[13] Jacobs, Paul S., ed. Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

[14] Liddy, Elizabeth. "Enhanced Text Retrieval Using Natural Language Processing." ASIS Bulletin (April/May 1998). On the Web at http://www.asis.org/Bulletin/Apr-98/liddy.html.

[15] Spink, Amanda and Howard Greisdorf. "Partial Relevance Judgments and Changes in Users' Information Problems During Online Searching." National Online Meeting, 18th Proceedings (1997): pp. 323-334.

[16] Callan, J. P. and Croft, W. B. An evaluation of query processing strategies using the TIPSTER collection. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (June 27 - July 1, Pittsbrugh, PA). ACM/SIGIR, New York, 1993, pp. 347–355.

[17] Chinchor, N., Hirschman, L., and Lewis, D. D. Evaluating message understanding systems: An analysis of the Third Message Understanding Conference (MUC-3). Computational Linguistics19, 3 (1993), 409–449.

[18] Guthrie, Louise and Leistensnider, James, `A Simple Probabilistic Approach to Classification and Routing’, Proceedings of the TIPSTER Text Program Phase II Workshop, Sponsored by Defense Advanced Research Projects Agency, May 6-8, 1996.

[19] Strzalkowski, Tomek. 1995. “Natural Language Information Retrieval” Information Processing and Management, Vol. 31, No. 3, pp. 397-417. Pergamon/Elsevier

[20] Feldman, Susan E. "Searching Natural Language Search Systems" Searcher (October, 1994): pp. 34-39.

[21] Feldman, Susan E. "Testing Natural Language: Comparing DIALOG, TARGET, and DR-LINK." ONLINE 0, No. 6 (Nov. 1996) pp. 71-79.

[22] Harman, Donna K., ed. The First Text Retrieval Conference (TREC-1). Bethesda, MD: National Institute of Standards and Technology, March 1993. (Available from NTIS: PB93-191641).