Journal of Theoretical and Applied Information Technology
31st March 2018. Vol.96. No 6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A STUDY ON NLP APPLICATIONS AND AMBIGUITY PROBLEMS

SHAIDAH JUSOH
King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan
E-mail: s.ibrahim@psut.edu.jo

ABSTRACT

Natural language processing (NLP) has been considered as one of the important area in Artificial Intelligence. However, the progress made in natural language processing is quite slow, compared to other areas. The aim of this study is to conduct a systematic literature review for identifying the most prominent applications, techniques and challenging issues in NLP applications. To conduct this review, I had screened 587 retrieved papers from major databases such as SCOPUS and IEEE Explore, and also from Google search engine. In searching relevant papers search keywords such as "*natural language processing*, *NLP applications, and complexity of NLP applications*" had been used. However, to focus to the scope of the study 503 papers were excluded. Only the most prominent NLP applications namely information extraction, question answering system and automated text summarization were chosen to be reviewed. It is obvious that the challenging issue in NLP is the complexity of the natural language itself, which is the ambiguity problems that occur in various level of the language. This paper also aims at addressing ambiguity problems which occur at lexical and structural levels and significance techniques or approaches for solving the problems. Finally, the paper briefly discuss the future of NLP.

Keywords: *Natural language processing, NLP applications, Ambiguity in NLP*

1. INTRODUCTION

Information is normally stored in text documents. These text documents can be found on personal desktop computers, intranets and on the Web. Valuable knowledge is normally embedded inside unstructured texts. The Web has been considered as the world's largest repository of knowledge, and it is being constantly augmented and maintained by millions of people around the world. However, it is not in the form of a database from which records and fields are easily manipulated and understood by computers, but in natural language texts which are intended for human reading. In spite of the promise of the semantic web, the use of English language and other natural language texts will continue to be a major medium for communication, knowledge accumulation, and information distribution [1]. This requires the study of NLP; a field of computer science and linguistics, concerning with the interaction between computers and a natural language. The term natural language is normally used to distinguish human languages such as Arabic or English from formal languages such as C++, C, Java or XML. The ultimate goal of NLP researchers is to create a software program that enables computers to understand or generate language used by humans. The foundation of NLP is basically lies on many disciplines, such as linguistics, mathematics, computer sciences, psychology, and so on.

Research work in NLP has been ongoing since more than six decades. Alan Turing's article "Computing Machinery and Intelligence" published in 1950s has been considered as the earliest work in NLP. Later in 1954, the work of Georgetown University and IBM on translating 60 Russian sentences into English language was considered among the earliest successful in NLP. In the year of 1978-1970, Terry Winograd, an American professor of computer science at Stanford University showed a successful computer program that can understand a natural language. The program was named as SHRDLU, developed using Lisp programming language. During the 70's NLP are based on conceptual ontologies and real world information. This field is moving rapidly and much work has been conducted in the last 10 years. Despite the goal of NLP's work far from being completely successful, significant positive outcomes has been shown in some research work. [2], [3], [4],[5] and [6]. The aim of this paper is to survey the most prominent applications and techniques in NLP. Despite a number of survey papers on NLP have

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

been published so far, to the best of my knowledge, none of them have discussed about the most prominent NLP applications and approaches in solving ambiguity problems comprehensively. This paper is organized as the following. Section 2 presents research questions and methods used, section 3 presents results and analysis, section 4 address the future of NLP, and the paper is concluded in section 5.

2. RESEARCH AND METHODS

This review is guided by the following 4 research questions:

RQ1: What are the different between natural language understanding and natural language generation?

RQ2: What are the most prominent NLP applications and their techniques?

RQ3: What are the challenging issues in NLP?

RQ4: What are approaches and techniques used to resolve ambiguity problems?

The first step of this systematic literature review is screening all 587 retrieved documents. The process of screening includes download the papers from the major databases such as SCOPUS, IEEE Explore, Google Search, and read their abstracts. Three main keywords for search were used: natural language processing, NLP applications, and complexity of NLP applications. During this screening process 300 papers were excluded because of poor presentation. Then 287 papers were reviewed and evaluated. This process excluded 203 papers. Only 84 papers were selected to be included in this paper. The selection was made based on three criteria: relevant to the research questions, well written, and had been cited by other researchers. Classifications of the selected papers are shown in Table 1.

Table 1: Classifications of extracted papers

Topic	Sub-Topics	References
NLP Application	Information Extraction	[11] to [25]
	Question Answering System	[26] to [32]
	Automated Text Summarization	[33] to [43]
Challenging Issues in NLP	Structural Ambiguity	[45] to [64]
	Lexical /Word Sense Ambiguity	[65] to [84]

3. RESULTS AND ANALYSIS

NLP encompass both text and speech, however, work on speech processing has evolved into separate fields. When NLP term is used, it normally refers to natural language generation (NLG) and natural language understanding (NLU). NLG involves with some form of computerized data into natural language, rather than the other way around. However NLG is different than techniques for 'report generation', 'document generation', 'mail merging' and so on. These techniques are simply plugging a fixed data structure such as a table of numbers or a list of names into a template in order to produce complete documents. On the other hand, NLG uses some level of underlying linguistic representations of texts, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic reliazer to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. A machine translation system is the most known application of NLG.. The system analyzes texts from a source language into a grammatical or conceptual representation, and then generates corresponding texts in the target language [7]. On the other hand NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic [8]. NLU is independent from speech recognition [9]. However, the combination of the two may produce a powerful human-computer interaction system. When combined with NLU, speech recognition transcribes an acoustic signal into a text. Then the text is interpreted by an understanding component to extract the meaning. The connection between speech recognition analysis and natural language understanding is illustrated in Figure 1.

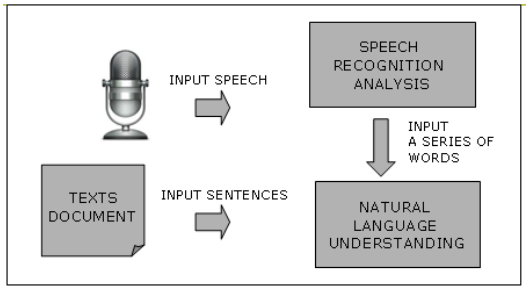

Figure 1. The diagram shows the connection between speech recognition analysis and natural language understanding.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1. NLP Applications

Applications of NLP have been included in many study fields, such as natural language user interface, automated text summarization, information extraction, machine translation, questions answering system, speech recognition, text mining, and document retrieval and so on. However, in this review the focus is given on techniques used in information extraction, automated text summarization, and questions and answering system. These applications are normally used as fundamental approaches for other types of NLP applications. For example, information extraction is a fundamental task in text mining data analysis and text mining applications.

3.1.1. Information Extraction

Information Extraction (IE) is defined as one of NLP's tasks for recognizing and extracting instances of a particular pre-specified class of entities, relationships, and events in natural language texts.

The IE field has been initiated by DARPA's MUC program (Message Understanding Conference) in 1987. MUC has originally defined IE as the task of (1) extracting specific, well-defined types of information from the text of homogeneous sets of documents in restricted domains and (2) fill pre-defined form slots or templates with the extracted information. The process of information extraction is also called text analysis. It turns the unstructured information embedded in texts into structured data. Information extraction is an effective way to populate the contents of a relational database. Figure 2 illustrates how entity extraction can be applied for text/data mining and visualization approaches. A typical information extraction system has three to four major components [10]. The first component is tokenization or zoning module. This module is used to split an input document into its basic building blocks. The typical building blocks are words sentences and paragraphs. The second component is a module for performing morphological and lexical analysis. This module handles activities such as assigning post tags to the document's various words. The third component is a module for syntactic analysis. This part of an IE system establishes the connection between the different parts of each sentence.

This is achieved either by doing full parsing or shallow parsing, however in many IE works, a shallow parsing approach is used. The forth component is a module for domain analysis. The information collected in the previous components is analyzed to determine the relationships between entities.

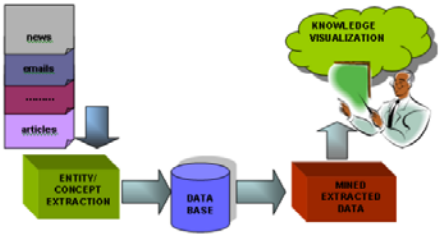

Figure 1. The diagram shows a database can be populated through information extraction technique. The populated data is then mined to extract the hidden knowledge. The knowledge can be visualized through visualization approach.

The major task in IE is known as named entity recognition (NER). A process of named entity recognition refers to the combined task of finding spans of text that constitute proper names and then classifying the entities referred to according to their types. It is an important task in many NLP applications such as information retrieval, machine translation, and question answering systems. The most studied entity types are categorized into "proper names", "names of persons", "locations", and "organizations". In general, early research works of NER focus on recognizing "proper names". The "location" type can, in turn, be divided into multiple subtypes of "fine-grained locations" [11]. Similarly, "fine-grained person" subcategories, like "politician" and "entertainer," appear in the work of [12]. The "person" type is quite common and used at least once in an original way by [13], who combine it with other cues for extracting medication and disease names (e.g., "Parkinson disease").

There are two known methods in NER. The first method is based on knowledge-based approach and the second one is based on statistical or machine learning approach. Knowledge-based approach is also known as unsupervised learning approach, while statistical and machine learning approach as a supervised learning approach. In a supervised learning approach, NER process is learned automatically on large text corpora and then supervised by human [14]. While in the unsupervised learning approach, an existing lexical database such as WordNet is used [15]. The combination of supervised and unsupervised

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

learning approaches produces a semi-supervised learning approach [16] [17].

Although a significant number of NER research work has been devoted to English language, some of other languages also have received attention by NER researchers. These include Turkish [18]), Danish [19], Hindi [20], Polish [21], and Arabic language [22]. However, this task is quite challenging in Arabic language because in this language, a proper name is not indicated by a capital letter (as in English language). Nevertheless, some efforts have been shown for Arabic language. For example, reference [23] has attempted to solve the problem using a hybrid system built based on both statistical methods and predefined rules.

Reference [24] presented an application that can evaluate the performance of a number of the Arabic root extraction methods. The implemented methods in this system are selected according to a previous classification, where these methods are classified into five categories: Light Stemmer, Arabic Stemming without a root dictionary, MT-based Arabic Stemmer, N-gram based on similarity coefficient and N-gram based on dissimilarity coefficient. The evaluation was conducted on the same terms in a corpus of two thousand words and their roots.

Reference [25] presented a working Arabic information extraction (IE) system that is used to analyze large volumes of news texts every day to extract the named entity (NE) types person, organization, location, date, and number, as well as quotations (direct reported speech) by and about people. In their work, they deployed multilingual NER to cover Arabic where they presented what Arabic language-specific resources had to be developed and what changes needed to be made to the rule sets in order to be applicable to the Arabic language.

3.1.2. Question and Answering Systems

Question and answering (QA) is a system that is able to automatically understand and build answers to human user questions which are posed in a natural language. The QA system is expected to allow users to ask questions in a "everyday language". Answers may be long or short, they may be lists or narrative. They may vary with intended use and intended user [26]. For example, if a user wants a justification, this requires a longer answer, while comprehension tests may require short answers phrases. To answer the question, the system must be able to analyze the question first. Questions can be distinguished by its type; factual, opinion or summary. If the question is in the context of ongoing interaction, the answer must be consulted with the online resource. Normally, the answer is presented in some kinds of form [26].

QA can be categorized into two types; according to its used methods. Firstly, shallow QA systems which use techniques like pattern matching in returning a final answer. Since such a method ignores the issue of semantic, thus, many relevant answers may be missed out or irrelevant answers may be retrieved. Secondly, deep QA systems; the deep parsing technique of NLP is used [27].

BASEBALL is the best known early question answering program [28]. The program answers questions about baseball games played in the American league over one season. Given a question such as `Who did the Red Sox lose to on July 5?' or `How many games did the Yankees play in July?' or even on how many days in July did eight teams play?, BASEBALL analysed the question, using linguistic knowledge, into a canonical form which was then used to generate a query against the structured database containing the baseball data.

One example of current working QA system is START, the world's first Web-based question answering system (http://start.csail.mit.edu/), has been online and continuously operating since December, 1993. It has been developed by Boris Katz and his associates of the InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory. START aims to supply users with "just the right information," instead of merely providing a list of hits. Currently, the system can answer millions of English questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors) and people (e.g., birth dates, biographies). A question such as "Who is the ruler of Jordan?" can be given to the system. The system will return an answer as the following" Ruler: King Abdullah II (1999)" and other relevant answers such as the current prime minister of Jordan. If a question such as `Why the sky is blue?' is given, the system will give a long narrative answer. Reference [29] stated that START answers natural language questions by presenting components of text and multi-media information drawn from a set of information resources that are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

hosted locally or accessed remotely through the Internet. These resources contain structured, semi-structured and unstructured information.

Although most of the published work reported on QA systems in English, some efforts in designing and developing QA systems in Arabic language have been shown. AQAS has been considered as the earliest work for QA systems in Arabic language [30]. The system was developed based on knowledge based approach. It accepts Arabic queries which follow pre-defined rules and matches these against frames in its knowledge base. The architecture of the system is similar to early English based QA systems. The system was applied in a closed domain environment; however, no experimental results have been reported.

Reference [31] reported on the design and implementation of a question answering (QA) system called QARAB. QARAB takes natural language questions expressed in the Arabic language and attempts to provide short answers. The systems primary source of knowledge is a collection of Arabic newspaper text extracted from Al-Raya, a newspaper published in Qatar. To identify the answer, a keyword matching strategy was adopted. The extracted keywords in the questions are matched to the candidate documents which are selected using information retrieval approach. QA system for Arabic language based on keyword matching is also reported in [32]. Using a keyword, simple structures extracted from both a question and candidate documents selected by the IR system were used in the process of identifying the answer. In order to perform this process, an existing tagger is used to identify proper names and other crucial lexical items and build lexical entries.

3.1.3 Automated Text Summarization

Automated text summarization is a process of producing a readable, short, and meaningful summary from a long text document. Automatic text summarization has drawn a considerable interest since it provides a possible solution to a critical information overload problem that people are facing nowadays. An automated summarization tool can be utilized by busy managers to scan relevance information, researchers to have a quick glance on relevance research articles, students to have a quick understanding on subject matters, and so on. Technically, summarization is a process of deriving a shorter version of text from an original text, by selecting important contents. Text summarization was defined in [33] as "the process of distilling the most important information from a source to produce a shorter version for a particular user or task". Summarization is also considered as a distilling process to obtain the most important information from a source text to produce an abridged version for a particular user/users and task/tasks. Therefore, the process of summarization of any targeted text ranges from the interpretation of the source text, understanding and analyzing the meaning being imparted to it; after that representing the meaning of the source text which is based on certain areas of information, and at the end creating a summary of the source text that has been understood and finalize all that needs to be represented.

The basic idea of summarization is to understand the eventual meaning of the text which has been presented in a short time and in the form of a relatively shorter text. This shorter text, which is a subset of the original text, may not really convey all the details of the actual text, but it certainly aims at conveying the basic and actual idea which is trying to be conveyed in the text [34]. Research community in this area have been putting effort to apply automated text summarization in various applications such as document classification, information retrieval, document extraction, knowledge sharing, and so on. Text summarization is really a complex task itself, since a wide variety of techniques can be applied in order to condense content information, from pure statistical approaches to those using closer analysis of text structure involving linguistic and heuristic methods (anafora resolution, named entity recognition, lexical chains, etc,). In fact, many algorithms for feature reduction, feature transformation, feature weighting, etc. are directly related to this task, since they already try to select a proper and limited set of items that can be used as storing the core content of a given text.

Most of the working summarization systems are based on the extraction of a certain number of sentences found in the text which are considered to express most of the concepts present in the document [35]. Sentence extraction techniques are usually statistical, linguistics, and heuristic methods, or a combination of all those techniques in order to generate a final summary. The result is not syntactically or content wise altered. In a sentence extraction technique, a score is computed for each sentence based on features such as position of sentence in the document and the frequency of the word. Using these features, the most important

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sentences in a document are extracted. A generated summary is basically a collection of extracted sentences from the original text. Jusoh and her research associates [36] have shown an effort to improve the sentence extraction technique by introducing the sentence refinement technique. The summarization tool was developed and tested on texts which were written in English and Malay language. Their experimental results indicated shorter versions of summaries are obtained without losing its text context.

Ultimately, the aim of summarization techniques is to move one step forward; understanding and rearranging information of source texts to produce a readable and meaningful texts (sentence abstraction technique) [37]. The challenging task is to understand the whole text and generate a summary as a human does. Although this technique can produce a better summary, however, this technique is very difficult to be implemented. Furthermore, in a manual abstraction process, an editor would have to acknowledge six editing operations: reducing the sentences; combine them; transform them syntactically; paraphrasing its lexical; generalize and specify; and re-ordering the sentences [38]. Thus, automated abstraction requires a computer system to have the knowledge of a human editor. Sentence abstraction requires lots of computing time. On the other hand, sentence extraction is easier to be implemented as it does not require full understanding of the texts context.

Another possible technique is text categorization (a.k.a. text classification) [39],[40] and [41]. Basically, text categorization is the task of assigning pre-defined categories to free-text documents. Using this technique, a result of the summarization algorithm is a list of key-paragraphs, key-phrases or key-words that have been considered to be the most relevant ones. Although some methods are able to generate new sentences from the content, usually it consists a pure selection of textual fragments. It can provide conceptual views of document collections and has important applications in the real world. For example, news are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on. Another widespread application of text categorization is spam filtering, where email messages are classified into the two categories of spam and non-spam, respectively. While text classification in the beginning was based mainly on heuristic methods such as applying a set of rules based on expert knowledge; nowadays the focus has turned to fully automatic learning and even clustering methods.

Some work on automated text summarization for Arabic texts document can be found in the published literatures [42] and [43]. For example, the Arabic query-based text summarization system (AQBTSS) was reported in [42]. The system takes an Arabic document and a query (in Arabic) and attempts to provide a reasonable summary. In AQBTSS, each sentence is compared against the user query to find relevant sentences. In this case, the query has been used to select the documents. In 2009, El-Haj and associates presented a concept-based summarizer system (ACBTSS) [43]. This system takes a bag-of-words representing a certain concept as the input the system instead of a user's query. Each sentence in a document is matched against a set of keywords that represent a given concept.

3.2. Challenges Issues in NLP

Since decades, ambiguity has been a challenging issue for NLP researchers. In spite of some results on resolving ambiguity problems have been obtained, a number of important research problems have not been solved yet [44]. Ambiguity is still a great challenge for computational linguists and computer scientists. The concept of ambiguity is closely connected to semantic gap between the user's intentions and how she/he is able to convey, since it can lead to more than one interpretation of the user's input. Intentions are always a matter of interpretations. Ambiguity has been a critical issue for human computer interaction because of it pervasiveness in everyday life, yet its emergent nature challenges the role of design. Failure in giving a correct user's interpretations may cause the user to mistrust the system and discontinue use. Several types of ambiguity have been identified. These include: structural, syntactical, form class, word sense and local ambiguity.

3.2.1 Structural Ambiguity

Structural ambiguity occurs when a sentence can be analyzed as having more than one syntactic structure or parse tree. For example, the utterance "You can have *peas and beans or carrots*" can be analyzed in one of two ways, as indicated by the bracketing [*peas and beans*] or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[*peas and carrots*]. Syntactical ambiguity is a grammatical ambiguity of a whole sentence that occurs in sub-part-of a sentence. It is a grammatical construct, and results from the difficulty of applying universal grammatical laws to a sentence structure. For example, a sentence "*Salman hits the boy with the stick*". This phrase is ambiguous, as to whether a boy was hit with a stick, or whether a boy with a stick was struck by Salman.

Form class ambiguity arises when a given word can be analyzed as more than one part-of-speech. For example, *book*, may be either a *noun* or a *verb*, *plastic* can be either an *adjective* or a *noun*, and so on. Form class ambiguity necessarily gives rise to structural ambiguity as well, as in the famous example "*He saw her duck*". The words *her* and *duck* are both form class ambiguous. Taking *her* as possessive pronoun and *duck* as a noun, we get a structure of [*noun phrase*, *verb*, *noun*, *phrase*]. However, taking *her* as a personal pronoun and *duck* as a verb, we get a structure of [*noun phrase*, *verb*, [*noun phrase*, *verb*].

3.2.1.1 Structural disambiguation approach

The work on structural ambiguity reports on preposition phrase (**PP**) attachment. The number of published papers on lexical resolution is more than the number of published papers for structural ambiguity. Prepositions are often among the most frequent words in a language. For example, based on the British National Corpus (BNC) [45], four out of the top-ten most-frequent words in English are prepositions (of, to, in, and for). Despite their frequency, however, they are notoriously difficult to master, even for humans [46]. For instance, less than 10% of upper-level English as a Second Language (ESL) students can use and understand prepositions correctly [47].

Naturally, the number of PP contexts with attachment ambiguity is theoretically unbounded. The bulk of PP attachment research, however, has focused exclusively on the case of a single PP occurring immediately after an NP, which in turn is immediately preceded by a verb. PP attachment research has undergone a number of significant paradigm shifts over the course of the last three decades, and been the target of interest of theoretical syntax, AI, psycholinguistics, statistical NLP, and statistical parsing [48]. Two large areas of research on the syntactic aspects of prepositions are (a) PP attachment and (b) prepositions in multiword expressions. A sentence "*Malik eats rice with a spoon*" as an example of the PP attachment. PP attachment is the task of finding the governor for a given PP. In the given example above, the PP *with a spoon* is governed by either the noun *rice* or the verb *eats*. Determining the correct attachment site for PP is one of the major sources of ambiguity in natural language parsing and analysis.

Early research on PP attachment focused on the development of heuristics intended to model human processing strategies, based on analysis of competing parse trees independent of lexical or discourse context. As the research communities grow up, many researchers have attempted to resolve PP attachment ambiguity in many different angles. A significant shift in NLP research on PP attachment was brought by the authors of [49] who were the harbingers of statistical NLP and large-scale empirical evaluation. Researchers have been trying to tackle the problem by a variety of smoothing methods and machine learning algorithms including backed-off estimation [50], instance-based learning 51], maximum entropy learning [52], decision trees ([53], [54], neural networks ([55], [56], boosting [57]) as well as corpus [58].

Reference [59] also described a neural network based approach for resolving ambiguity in PP attachment. To disambiguate the PP attachment, the constituent namely verbs, noun, PP are associated with semantic classes from WordNet. The neural network method for PP attachment involves three phases, training, validation and testing. The approach is classified into supervised learning approaches. In their experiment, only one structure is used. A sentence does not go through a deep parsing process. The work [58] proposed to resolve the PP attachment ambiguity based on a *four-tuple* composed of the head verb of the verb phrase, the head noun of the noun phrase, and the preposition and head noun in the prepositional phrase. A corpus with known results, the Penn Treebank, is used for training and testing purposes.

However, statistical approaches are not appropriate or adequate in accounting for inferring prepositional phrase attachments in cognitive modeling systems, as human cognition is generally not a completely statistical process. Pure statistical models for disambiguation tasks also suffer from sparse-data problem. The hybrid method was introduced in [60], [61], where in [60] a corpus-based approach was integrated with knowledge-based techniques. In their work four head words were used; main verb (**v**), head noun (**n1**), the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

preposition (**p**), and the head noun (**n2**), where it was referred as quadruple (v, n1, p, n2). The clues include, syntactic cue, co-occurrance, syntactic features and conceptual relationships between v and n2 or between n1 and n2. They reported that the results of their experiments are considered as good. Reference [62] proposed a theoretical approach for the detecting ambiguities connected with the meaning of the user's input using a formal structure for the multimodal input. The proposed approach is also a hybrid approach which combines constraints multiset grammar with linear logic. They claimed that the hybrid approach provides an adaptive treatment of the ambiguities.

The importance of PP semantics has been discussed by many researchers. For example, [63] used preposition semantics in a cooperative question answering system in the context of cross-language question answering (CLQA), and further later [64] successfully applied their preposition word sense disambiguation (WSD) method in a paraphrase recognition task, namely, predicting that "*Kim covered the baby in blankets*" and "*Kim covered the baby with blankets*" have essentially the same semantics. They proposed seven general senses of prepositions (e.g. PARTICIPANT, INSTRUMENT, and QUALITY), and annotated prepositions occurring in 120 sentences for each of 10 prepositions. IE is one application where prepositions are crucial to a system accuracy. As a matter of fact, PP attachment plays an important role in named entities and in IE patterns and in linking the elements in a text.

3.2.2 Word Sense Ambiguity

Lexical ambiguity or word sense disambiguation (WSD) has been recognized as an AI-hard problem. A break-through in this field would have a significant impact on many relevant Web-based applications, such as Web information retrieval, improved access to Web services, IE, and so on [65]. WSD has obvious relationships to other fields such as lexical semantics, whose main endeavor is to define the relationships between "word" and "meaning" and "context" [66]). WSD can be viewed also as a classification task; word senses are the classes, and an automatic classification method is used to assign each occurrence of a word to one or more classes based on the evidence from the context and from external knowledge sources. WSD can be described as "given a set of words (e.g., a sentence or a bag of words), a technique is applied which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context" [67].

In [67] WSD was divided into two groups; lexical sample and all words WSD. In a lexical sample, a system is required to disambiguate a restricted set of target words usually occurring one per sentence. In this type of systems, a number of instances are labeled manually (training set) and then applied to unlabeled instances (test set). This is also known as a supervised system. In all words WSD, a system is required to disambiguate all open-class words in a text. These include nouns, verbs, adjectives and adverbs. This task requires a wide coverage of systems. Thus supervised systems can potentially suffer from the problem of data sparseness, as it is unlikely that a training set of adequate size is available for a wide coverage. This is a point where the use of external knowledge is considered for WSD. This type of systems is classified into unsupervised systems. Unsupervised systems based their disambiguation decisions on knowledge sources. The sources may belong to one of broad classes: syntactic, semantic and pragmatic [68]. Syntactic knowledge sources have to do with the role of a word within the grammatical structures of sentences. Semantic knowledge relates the word to its properties. This was demonstrated by the work of [69] where they have combined knowledge gathered from WordNet with results of an anaphora resolution algorithm. Knowledge sources include corpora (a collection of text), machine readable dictionaries and semantic network.

3.2.2.1. Knowledge-based approach

The use of knowledge-based approach has been demonstrated in the early WSD work. For example, [70] and [71] used manually encoded semantic knowledge for WSD. Unfortunately, the manual creation of knowledge resources is an expensive and time consuming effort, which must be repeated every time the disambiguation scenario changes. In recent years, existing lexical resources such as machine-readable dictionaries (MRDs) like WordNet [72], [73], [67] and Oxford Dictionary of English have been applied as an external source of knowledge in WSD work. According to [65] word senses clearly fall under the category of objects that are better described through a set of structured features. Thus they have applied structural pattern recognition approach to disambiguate word senses. In their work, graph representations of word senses are automatically generated from WordNet 2.7. Others who used WordNet include researchers of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[74] [75] and [76]. Early approaches to WSD based on knowledge representation techniques, have been replaced in the past few years by more robust machine learning and statistical techniques. However, according to [65] the results of recent comparative evaluations of WSD systems show that machine learning and statistical techniques have inherent limitations. On the other hand, the increasing availability of large-scale, rich lexical knowledge resources seems to provide new challenges to knowledge-based approaches [67].

3.2.2.2. Machine learning approach

The work presented in [77] has been considered as an early application of machine learning to the WSD problem. Several disambiguation cues, such as first noun to the left/right and second word to the left/right were extracted from parallel text. The senses are defined by determining the differences between them. This technique was also applied for machine translation. On the other hand, [78] used the *flip-flop algorithm* to decide which of the important cues for each word by using mutual scores between words. Syntactic relations between *subject-verb*, *verb-objec*t and *adjective-noun* have been used by [79] to determine the cues.

According to [80], most of the previous corpus-based approaches to the resolution of word-sense ambiguity are based on lexical information from the context of the word to be disambiguated suffer from the problem of data sparseness. To address this problem, they proposed a disambiguation method using co-occurring concept codes (CCCs). The use of concept-code features and concept-code generalization effectively alleviate the data sparseness problem and also reduce the number of features to a practical size without any loss in system performance. They claimed that the effectiveness of the CCC features and the concept-code generalization by experimental evaluations. The proposed disambiguation method was applied to a Korean-to-Japanese MT system that experimented with various machine-learning techniques. In a lexical sample evaluation, their CCC-based method achieved a precision of 82.00%, with an 11.83% improvement over the baseline. Also, it achieved a precision of 83.51% in an experiment on real text, which shows that their proposed method is very useful for practical MT systems.

The work presented in [81] demonstrated an effort to resolve ambiguous terms using sense-tagged corpora and UMLS with the motivation that the UMLS has been used in natural language processing applications such as information retrieval and information extraction systems. In their work, machine-learning techniques have been applied to sense-tagged corpora, in which senses (or concepts) of ambiguous terms have been most manually annotated. Sense disambiguation classifiers are then derived to determine senses (or concepts) of those ambiguous terms automatically. However, they conclude that manual annotation of a corpus is an expensive task.

Research of [82] proposed a method for lexical ambiguity resolution using corpus and concept information. Since the extracted knowledge is stored in words themselves, these methods require a large amount of space with a low recall rate. On the contrary, they resolve word sense ambiguity by using concept co-occurrence information extracted from an automatically sense-tagged corpus. The tested accuracy of their method exceeds 82.4% for nominal words, and 83% for verbal words.

Although WSD have not been applied to real task applications widely, a few researchers have taken an effort to do so. For example, reference [83] have extended the previous work by mining biological named entity tagging (BNET) that identifies names mentioned in text and normalizes them with entries in biological databases. They concluded that that names for genes/proteins are highly ambiguous and there are usually multiple names for the same gene or protein. Reference [84] investigated a particular technique for resolving ambiguity that is motivated by task-level ambiguity. In their study, they explored a technique to find commonly occurring patterns of part-of-speech in a query and allow the patterns to be transformed into clarification questions. The patterns are centered on a single query word and incorporate a small number of words on either side.

4. THE FUTURE OF NLP

Despite the lack of high-performing methods had preventing an extensive use of NLP techniques in many areas of information technology, such as information retrieval, natural language interfaces, query processing, advanced Web search, and many more real applications, NLP based applications are emerging technologies for business.

It is no doubt that NLP is an enabler for deploying natural, intelligent, and intuitive applications for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

everyday use. It is transforming the way how human interact with computers. Thus, resolving the complexity issues in a human language is indeed critical, vital, and urgent.

Applications such as chatbot, smart search, recommender, customer service, personal assistant, multi lingual automated translation machine, question answering, caption generation are expected to be able to capitalize NLP techniques for human-like understanding of speech and texts. Deeper applications such as extracting insights and analysis from a vast amount unindexed and unstructured data, mining texts, images, audios and videos or reading, filtering, analyzing, extracting, and visualizing pieces of knowledge from text documents such as emails, short messages, reviews, and so on, are seen as critical technologies of NLP in the future.

When machines are intelligent enough to understand and communicate in a human language, human users are able to be more effective and efficient in accessing, analyzing, and leveraging huge amount of data. NLP market is growing. According to a 2017 Tractica report [86], NLP market is estimated to be around 22.3 billion USD by 2025. This estimation has included the total NLP software, hardware and services. Furthermore, NLP solutions that leveraging AI will see a market growth from 136 million USD in 2016 to 5.4 billion USD by 2025.

5. CONCLUSION

This paper has successfully present the most prominent applications of NLP. These include information extraction, question answering systems, and automated text summarizations. Because of the mechanism (the citation numbers) used in selecting papers to be reviewed, a number of current literatures which discussed the selected topic might be left out unintentionally. In overall, this paper has given a depth overview of main applications of NLP. NLP has been also considered as one of AI hard problems. The complexity of natural language processing is caused by the ambiguity problems which always occur in a human language. Although the ambiguity problem may occur in all levels of a natural language, the most common problems always occur at lexical and structural levels. The paper also addresses, discusses, and distinguishes between approaches in resolving the ambiguity problems. The future technologies which are based on NLP are briefly highlighted.

REFRENCES:

[1] McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. Queue, 3 , 48–57.

[2] Sekimizu, T., Park, H., & Tsuji, J. (1998). Identifying the interactions between genes and gene products based on frequently seen verbs in medline abstract. Tokyo Japan: Universal Academy Press.

[3] Chu, C.-T., Sung, Y.-H., Yuan, Z., & Jurafsky, D. (2006). Detection of word fragments in mandarin telephone conversation. In International Conference on Spoken Language Processing. URL pubs/fragment-icslp-06.pdf

[4] Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. A. (2009). Topic modeling for the social sciences. In Workshop on Applications for Topic Models: Text and Beyond (NIPS 2009). Whistler, Canada.

[5] Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: identifying interactional style in spoken conversation. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '09), (pp. 638–646). Morristown, NJ, USA: Association for Computational Linguistics.

[6] Grenager, T., Klein, D., & Manning, C. D. (2005). Unsupervised learning of field segmentation models for information extraction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (pp. 371–378).

[7] Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing: An introduction to Natural language Processing, Computational Linguistics and Speech Recognition. United States of America: Prentice Hall.

[8]Allen, J. (1988). Natural Language Understanding, United States of America: The Ben-jamin/Cummings Publishing Company.

[9] Karat, C., Vergo, J., & Nahamoo, D. (2003). Conversational interface technologies. In J. A. Jacko, & A. Sears (Eds.). The Human-Computer Interaction Handbook, (pp. 169–186). Lawrence Erlbaum Associates.

[10] Feldman, R., & Sanger, J. (2007). The text mining Handbook: Advanced Approaches in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Analyzing Unstructured Data. United State of America: Cambridge University Press.

[11] Lee, S., & Lee, G. (2005). Heuristic methods for reducing errors of geographic named entities learned by bootstrapping. In Proceeding of the International Joint Conference on Natural Language Processing.

[12] Fleischman, M., & Hovy, E. (2002). Fine grained classification of named entities. In Proceeding of the 19th International Conference on Computational Linguistics (COLING).

[13] Bodenreider, O., & Zweigenbaum, P. (2000). Identifying proper names in parallel medical terminologies. Stud Health Technol Inform, 77 , 443–447.

[14] McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. In Proceedings of the Conference on Computational Natural Language Learning .

[15] Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. In Proceedings of the 1st International Conference on General WordNet, (pp. 466–471).

[16] Chang, C. H., & Kup, S.-C. (2004). A semi-supervised approach of web data extraction with visual support. Intelligent System, 19 (6), 56–64

[17] Nadeau, D. (2007). Semi-Supervised Named Entity Recognition:Learning to Recognize 100 Entity Types with Little Supervision. Ph.D. thesis, University of Ottawa.

[18] Cucerzan, S., & Yarowsky, D. (1999). Language independent named entity recognition combining morphological and contextual evidence. In Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

[19] Bick, E. (2004). A named entity recognizer for Danish. In Proceedings of the Conference on Language Resources and Evaluation.

[20] May, J., Brunstein, A., Natarajan, P., & Weischedel, R. (2003). Surprise! what‟s in a cebuano or Hindi name? ACM Transactions on Asian Language Information Processing (TALIP), 2 (3), 169–180

[21] Piskorski, J. (2004). Named-entity recognition for Polish with SProUT. In L. Bolc, Z. Michalewicz, & T. Nishida (Eds.) Lecture Notes in Computer Science, vol. 3490, (pp. 122–133).

[22] Huang, F. (2005). Multilingual Named Entity Extraction and Translation from Text and Speech. Ph.D. thesis, Carnegie Mellon University.

[23] Abuleil, S. (2006). Hybrid system for extracting and classifying arabic proper names. In Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, (pp. 205–210). Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS).

[24] AlHajjar, A., Hajjar, M., & Khaldoun, Z. (2010). A system for evaluation of arabic root extraction methods. In Proceedings of the 2010 Fifth International Conference on Internet and Web Applications and Services, ICIW ‟10, (pp. 506–512). Washington, DC, USA: IEEE Computer Society.

[25] Zaghouani, W. (2012). Renar: A rule-based arabic named entity recognition system. 11 (1), 2:1–2:13.

[26] Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering:the view from here. Natural Language Engineering, 7, 275–300.

[27] Al-Harbi, O., Jusoh, S., & Norwawi, N. M. (2011). Lexical disambiguation in natural language questions-nlqs). Journal of Computer Science Issues, 8, 143 International –150.

[28] Green, B., Wolf, A., Chomsky, C., & Laughery, K. (1961). Baseball: An automatic question answerer. In Proceedings Western Joint Computer Conference, vol. 19, (pp. 219–224).

[29] Katz, B., Borchardt, G., & Felshin, S. (2006). Natural language annotations for question answering. In Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006).

[30] Mohammed F. A, Khaled Nasser, & Harb H.M. (1993). A knowledge based Arabic question answering system (AQAS). SIGART Bull. 4, 4 (October 1993), 21-30.

[31] Hammo, B., Abu-Salem, H., & Lytinen, S. (2002). Qarab: a question answering system to support the arabic language. In Proceedings of the ACL-02 workshop on Computational approaches to semitic languages, SEMITIC ‟02, (pp. 1–11). Stroudsburg, PA, USA: Association for Computational Linguistics.

[32] Kanaan, G., Hammouri, A., Al-Shalabi, R., & Swalha, M. (2009). A new question answering system for the arabic language. American Journal of Applied Sciences, 6, 797–805.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[33] Mani, I., & Benjamin, J. (2002). Review of automatic summarization. Journal of Computational Linguistics, 28, 221–223.

[34] Mani, I. (1999). Advances in Automatic Text Summarization. Cambridge, MA: MIT Press.

[35] Loo, P., & Tan, C. (2002). Word and sentence extraction using irregular pyramid. In Proceedings of the 5th International Workshop on Document Analysis Systems V (DAS ‟02), (pp. 307–318). Heidelberg: Springer-Verlag, London, UK.

[36] Jusoh, S., Masoud, A. M., & Alfawareh, H. M. (2011). Automated text summarization: Sentence refinement approach. In V. Snasel, J. Platos, & E. El-Qawasmeh (Eds.) Digital Information Processing and Communications, vol. 189 of Communications in Computer and Information Science, (pp. 207–218). Springer Berlin Heidelberg.

[37] Chan, S. (2006). Beyond keyword and cue-phrase matching: a sentence-based abstraction technique for information extraction. Decision Support System, 42, 759–77. Chang, C. H., & Kup, S.-C. (2004). A semi-supervised approach of web data extraction with visual support. Intelligent System, 19 (6), 56–64.

[38] Jeek, K., & Steinberger, J. (2008). Automatic text summarization: The state of the art and new challenges. In Proceedings of the Znalosti 2008, (pp. 1–12).

[39] Devasena, C. L., & Hemalatha, M. (2012, March). Automatic text categorization and summarization using rule reduction. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 594-598). IEEE.

[40] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

[41] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, *1*(1), 69-90.

[42] El-Haj, M., & Hammo, B. (2008). Evaluation of query-based arabic text summarization system. In Proceeding of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE08, (p. 17). IEEE Computer Society.

[43] El-Haj, M., Kruschwitz, U. and Fox, C., 2009, November. Experimenting with Automatic Text Summarisation for Arabic. In *LTC* (pp. 490-499).

[44] Alfawareh, H.M. & Jusoh, S. (2011). Resolving ambiguous entity through context knowledge and fuzzy approach. International Journal on Computer Science and Engineering (IJCSE), 3 (1), 410 – 422.

[45] Burnard, L. (2000). Reference Guide for the British National Corpus. Oxford, UK: Oxford University Computing Services.

[46] Chodorow, M., Tetreault, J., & N.Han (2007). Detection of grammatical errors involving prepositions. In Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, (pp. 25–30).

[47] Lindstromberg, S. (2001). Preposition entries in UK monolingual learners dictionaries: Problems and possible solutions. Applied Linguistics, 22 (1), 79–103.

[48] Baldwin, T., Kordoni, V., & Villavicencio, A. (2009). Prepositions in applications: A survey and introduction to the special issue. Computational Linguistic, 35 (2), 119–149.

[49] Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. Computational Linguistics., 19 (1), 103–120.

[50] Collins, M., & Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In Proceedings of the 3rd Annual Workshop on Very Large Corpora, (pp. 27–38).

[51] Zavrel, J., Daelemans, D., & Veenstra, J. (1997). Resolving PP attachment ambiguities with memory-based learning. In Proceedings of the Conference on Computational Natural Language Learning (CoNLL-97), (pp. 136–144).

[52] Ratnaparkhi, A., Reynar, J., & Roukos, S. (1994). A maximum entropy model for preposi¬tional phrase attachment. In Proceedings of the workshop on Human Language Technol¬ogy, (pp. 250–255).

[53] Merlo, P., Crocker, M. W., & Berthouzoz, C. (1997). Attaching multiple prepositional phrases: Generalized backed-off estimation. In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97), (pp. 149–155).

[54] Alam, Y. S. (2004). Decision trees for sense disambiguation of prepositions: Case of over. In Proceedings of the Workshop on Computational Lexical Semantics, (pp. 52–59).

[55] Sopena, J. M., Lloberas, A., & Moliner, J. L. (1998). A connectionist approach to prepo¬sitional phrase attachment for real world texts. In Proceedings of the 36th Annual

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Meeting of the ACL and 17th International Conference on Computational Linguistics (COLING/ACL-98), (pp. 1233–1237).

[56] Alegre, M. A., Sopena, J. M., & Lloberas, A. (1999). PP-attachment: A committee machine approach. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), (pp. 231–238).

[57] Abney, S., Schapire, R. E., & Singer, Y. (1999). Boosting applied to tagging and pp attachment. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), (pp. 38–45).

[58] Nadh, K., & Christian, H. (2009). Prepositional phrase attachment ambiguity resolution using semantic hierarchies. In Proceedings of the Ninth IASTED International Conference on Artificial Intelligence and Applications, (pp. 73–80).

[59] Srinivas, M., & Bhattacharyya, P. (2006). Prepositional phrase attachment through seman¬tic association using connectionist approach. In Proceedings of the Third International WordNet Conference (GWC2006), (pp. 273–277).

[60] Wu, H., & Furugori, T. (1996). Prepositional phrase attachment through a hybrid disambiguation model. In Proceedings of the 16th conference on Computational linguistics, (pp 1070-1073). Morristown, NJ, USA: Association for Computational Linguistics.

[61] Hartrumpf, S. (1999). Hybrid disambiguation of prepositional phrase attachment and in-terpretation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99), (pp. 111–120).

[62] Chiara, C. M., Fernando, F., & Patrizia, G. (2008). Ambiguity detection in multimodal systems. In Proceedings of the Working Conference on Advanced visual interfaces, (pp. 331–334). New York, NY, USA: ACM.

[63] Benamara, F. (2005). Reasoning with prepositions within a cooperative question-answering framework. In Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications, (pp. 145–152).

[64] Boonthum, C., Toida, S., & Levinstein, I. (2006). Preposition senses: Generalized disambiguation model. In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2006), (pp. 196–207).

[65] Navigli, R., & Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis an

[66] Agirre, E., & Edmonds, P. (2007). Introduction. In E. Agirre, & P. Edmonds (Eds.) Word Sense Disambiguation: Algorithms and Applications, (pp. 1–28). New York: Springer Verlag.

[67] Navigli, R. (2009). Word sense disambiguation: a survey. ACM Computing Surveys, 41 (2), 1–69.

[68] Agirre, E., & Stevenson, M. (2007). Knowledge sources for WSD, (pp. 217–251). New York: Springer Verlag

[69] McCarthy, D., Carroll, J., & Preiss, J. (2001). Disambiguating noun and verb senses using automatically acquired selectional preferences. In Proceedings of the SENSEVAL-2 Workshop at the European Chapter ACL, (pp. 119–122). Toulouse, France.

[70] Schank, R., & Abelson, R. (1977). Scripts, Plans, Goals, and Understanding. Hillsdale, N.J: Lawrence Erlbaum.

[71] Wilks, Y. (1978). A preferential pattern-seeking semantics for natural language inference. Artificial Intelligence, 6 , 53–74.

[72]Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the International Joint Conference Artificial Intelligence (IJCAI), (pp. 448–453).

[73] Mihalcea, R., & Moldovan, D. (2001). A highly accurate bootstrapping algorithm for word sense disambiguation. International Journal of Artificial Intelligence Tools, 10 (1-2), 5– 21.

[74] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics.

[75] Agirre, E., & Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING Luxembourg 2000 , (pp. 11–19).

[76] Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relat-edness. In Proceedings of the 18th international

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

joint conference on Artificial intelligence, (pp.805–810).

[77] Brown, P., Stephen, E., Pietra, D., Vincent, J., Pietra, D., & Mercer, R. L. (1991). Word sense disambiguation using statistical methods. In Proceedings of the 29th Annual Meeting for Computational Linguistics, (pp. 264–270).

[78] Nadas, A., Nahamoo, D., Picheny, M., & Powell, J. (1991). An iterative approximation of the most informative split in the construction of decision trees. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (pp. 565– 568). Toronto.

[79] Yarowsky, D. (1996). Homograph disambiguation in text-to-speech synthesi. In J. Hirschberg, R. Sproat, & J. van Santen (Eds.) Progress in Speech Synthesis, (pp. 159–175). New York: Springer Verlag. [77]

[80] Youjin, C., & Jong-Hyeok, L. (2005). Practical word-sense disambiguation using c-occurring concept codes. Machine Translation, 19 (1), 59–82.

[81]Liu, H., Hu, Z., Torii, M., Wu, C., Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. Journal of the American Medical Informatics Associations (JAMIA), 13, 497–507.

[82] Liu, H., Johnson, S. B., & Friedman, C. (2002). Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. Journal of the American Medical Informatics Associations (JAMIA), 9, 621–636.

[83] Liu, H., Hu, Z., Torii, M., Wu, C., & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. Journal of the American Medical Informatics Associations (JAMIA), 13, 497–507.

[84] James, A., & Hema, R. (2002). Using part-of-speech patterns to reduce query ambiguity. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, (pp. 307–314). New York, NY, USA: ACM.

[85]Tractica, Natural Language Processing Market to Reach $22.3 Billion by 2025, August 21, 2017, Retrieved from: https://www.tractica.com/newsroom/press-releases/natural-language-processing-market-to-reach-22-3-billion-by-2025/