© 2017 IJSRSET | Volume 3 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Research Problems in Natural Language Processing – A brief Overview

P. Selvaperumal
Assistant Professor, Department of Computer Applications, Don Bosco College, Yelagiri, Yelagiri, Tamil Nadu, India

ABSTRACT

Natural Language processing field concerns with computers processing human languages. Natural language processing (NLP) can be categorized into text and speech processing and also as Natural language understanding and Natural language generation. Various problems occur in language processing at every level of processing including morpheme/phoneme, syntactic, semantic, pragmatic, and Discourse level. This paper gives a brief overview of various research areas concerns with the text and speech processing. This assumes the significance keeping in mind the various applications like Dialogue system, Question and answering system, Information retrieval system, Information extraction system, Expert system etc both in text and speech processing areas. All the applications together enable the community to construct an intelligent system that process and generate human language just as a human being.

Keywords : Natural langauge Understanding; Natural langauge Generation, Text Mining, Speech processing.

I. INTRODUCTION

Natural language processing (NLP) is the field of computer science that deals with computers processing human language. The notion of computers understanding human language includes computer possessing language understanding skills and language generation skills. Thus Natural language processing is required for Natural Language Understanding (NLU) and Natural Language Generation (NLG) [1].

Language processing can be at syntax level or at semantic level. Processing text at syntax level includes Morphological analysis, syntactic analysis etc whereas semantic level processing includes higher level text processing techniques that may include syntactic level processing as preprocessing step or an inherent step in the process. Computers processing human languages can be viewed at two levels namely syntactic and semantic level. All the higher level language processing tasks involves basic language processing tasks like Morphological analysis, POS tagging, Named Entity Recognition (NER), shallow and deep sentence parsing etc. Higher level language processing includes semantic analysis, pragmatics, discourse processing etc.

Performance of NLP systems are depleted by a number of factors like ambiguity in natural languages which ranges from punctuation level ambiguity to discourse level ambiguities, complexity of sentence structure, difficulty in identifying named entities, slang, jargon, sarcasm, idioms, phrases, spelling variations(British/American), grammar variations, acronyms/abbreviations etc. While many of these are dealt in English language to reasonably well, there performance can be improved or resolving these problems systems in other languages is attractive field to work on. A study on performance enhancement of NLP system by resolving these issues is also interesting area to work. This paper explains ambiguity issue in detail below since it is much discussed area in NLP.

Natural languages are highly ambiguous. Computers trying to understand human language often end up with twisted meanings. Thus ambiguity resolution is an integral part important natural language processing tasks. Ambiguity in natural language can arise at every level from word to discourse (document or set of documents). Morphological ambiguity for example Unlockable can be broken into two ways like Un+lockable or Unlock+able given how serious the difference if the

context is not considered for ambiguity resolution. Lexical ambiguity arises because reasons including synonymy (many words having one meaning) and polysemy (one word having many meanings). This ambiguity stretches up to sematic level as well as in the example "Johan and Mary got married" can be sensed in two ways whether both got married to each other and got married separately. Similar to above mentioned ambiguities in text document, ambiguities in speech depreciate the NLP system performance. For example, Phonetic ambiguity "I got up late" and "I got a plate" seems same at speech level but gives two different senses [2]. Resolving ambiguity in English language may be a treaded path, but resolving ambiguities that are specific to other languages have lot of research scope.

Even though ambiguity resolution in NLP is much known area for several years still lot of research scope is there. Ambiguities that arise while processing one language may occur while another language is processed. Thus resolving ambiguities that arises while processing text or speech is attractive area to work. Conventional example given for sentence level ambiguity (PP attachment problem) "I saw a man with the telescope" that gives multiple senses (while seeing him, you had the telescope or he had the telescope). Ambiguity resolution is much worked area in language processing, but there are many topics in this area that has scope to be researched.

Issues at the syntactic level Processing language at the syntactic level involves strong grammar knowledge. Most of the problems at the syntactic level language processing for English language is researched but there are some scope for performance improvement. A similar study on other regional languages is interesting.

Text Processing

Referential ambiguity refers to which entity the known entity is referring to. For example, an entity "Super star" may refer to different persons (entities) in different documents. Such ambiguous entities adversely affects the performance of NLP systems like information extraction system. Another related problem is co-reference resolution. Co-reference resolution refers to extraction of all the entities that refers to a given entity. For example, if "henry ford" is referred by his surname "ford" or "Henry" or He etc in a document then extracting all such entities constitutes its co-reference chain. If the same process is done in multiple documents, then it is called as cross-document co-reference resolution. Extracting co-references in languages in non-English text corpus is interesting area to study.

POS Tagging refers to the process of assigning appropriate grammar tags for every word in a sentence. Training a POS tagger to tag new sentences forms the core part of Tagging process in NLP. POS tagging is also a preprocessing step for a number of higher level language processing tasks. For example it can be used by search engine to distinguish the senses between the queries "Who chaired the meeting" to "price of chair" where the former refers to who presided over the meeting where latter refers to object chair. Thus use of POS tagging to improve performance of a search engine or other NLP systems like dialogue system, question and answering system is attractive. But again problem may arises because of shorter length, unstructured or non-grammatical structure nature of queries.

Wordnet is a lexical database where words are grouped into clusters that are synonyms called synsets [3]. For example search for the word "can" yields 6 nouns and 3 verbs. i.e "can" can be used in 6 senses as noun and 3 senses as verb. Thus even the usage of POS tagger to resolve the ambiguity of the word "can" will not improve the results. Study of such ambiguous words and finding ingenious ways to resolve highly ambiguous words in attractive area. Similarly a study on new words added to languages [4] and their impact in language processing is also intriguing.

Speech Processing

Much of the speech processing involves study of processing the speech signals in digital representation. Similar to text, speech processing includes speech understanding and speech synthesis sub fields. Speech coding is related area where the job is to represent the speech in digital signal to transfer it or store it in digital media.

Identification of names of person, place etc in the speech content is a problem similar to Named Entity Recognition (NER) problem in text processing.

Like ambiguities that arises in text, there ambiguities that arises in speech both a syntactic and semantic levels. Syntactic level ambiguities includes phone ambiguity because of homophones or homonyms etc.

Issues at the Semantic Level

There are many language processing areas where the processing system need to know the meaning of what is

expressed in the text or speech to process the text or speech. Many area in semantic language processing still needs improvement.

Text Processing

Text synthesis is an interesting are in text mining where the task is to synthesize text for a given topic or theme. This is far more complex than topic capturing where the task is to detect the topic for a given passage or text. Text summarization is another related problem where the algorithm has to summarize the input text passage. Text paraphrasing is another enticing are to work on. The task in paraphrasing is to reproduce the same theme of a text passage in other easily understandable words. Text categorization or is a conventional problem in text mining where the text documents are to be classified into appropriate categories like classifying email into Personal/official or non-spam/spam. While text categorization is conventional the idea of web page categorization is slightly deviant from this as web pages have additional features like web links, anchor texts, Meta tags etc. which can't be found in plain text documents. Categorizing other types of texts like tweets, sms, whats app messages are slightly different as these are generally short messages and classifying them in conventional ways will be difficult because of dearth in words.

Sentimental analysis is widely regarded problem in nlp today. It involves analyzing the input sentiments in the form of text into positive, negative or neutral. Emotional analysis is another related work where the task is to find the emotion or the mood of the passage or review. A wide range of works coming out in the area leverages machine learning for analyzing the sentiments or opinions that are expressed in the review/feedback.

Most of these language processing tasks are done by employing the following categories of machine learning algorithms like supervised, unsupervised, semi-supervised or reinforcement learning algorithms. The choice of selecting particular category depends on a number of factors like availability of data, the nature of task, availability of algorithms for the specific task etc.

Automatic Machine translation is a separate area in language processing where the task is to translate input sentence or passage into sentence or passage of desired language. Issues relating to automatic translation can be at syntax level or at semantic level. Syntax level issues like structure of the language (English follows left to right pattern whereas Arabic, Hebrew, Persian etc follows right to left pattern), splitting of words (Sandhi splitting) etc. Issues at the sematic level includes translating idioms, phrases, sarcasm etc.

Approaches to machine translation includes neural machine translation, statistical machine translation, and hybrid approach that combines both.

Information extraction (IE) refers to the process of extracting interesting information from text documents. Since web is considered as largest repository of knowledge, extracting interested information from the web is an attractive field that people are already working on.

Question and answering system is another enticing area to work on. The process of composing answers with the help of information retrieval and extraction from knowledge base like Web or Wikipedia is attractive area to work. Question and answering system by MIT [5] START shows how fascinating the area is. While developing a complete question and answering system is far from trivial, domain specific question and answering system is less complex and quite easy than a general question and answering system which can answer any query.

Information retrieval (IR) system is widely dealt problem but still there are many areas in IR which needs to be addressed. Since natural language is highly ambiguous removal of intrinsic ambiguity in the query form inherent part of any information retrieval system. Ambiguity may be in names (synonymy, polysemy etc) or in any other parts of the sentences. Cross lingual information retrieval system [6] is another promising area where the task is to retrieve documents that are in other languages to that of the query language. To be precise, search engines has to retrieve documents of any language provided it is relevant to the query. Such search engines are generally regarded as semantic search engines which retrieves documents that are semantically related to the query. An extension of traditional Information retrieval system is web information retrieval system that involves retrieving relevant web pages for an input query. Research areas in Information retrieval includes query expansion, index creation and maintenance, information retrieval models etc.

Expert system is a broad field that uses natural language processing (both natural language understanding and

Natural language generation) and knowledge base to provide expert assistance to the users. Its uses ranges from Medical advice to agricultural expertise and education.

Discourse processing [7] involves processing collection of text documents. While much work in NLP focuses on individual words or sentences, discourse is considering the document as a whole or corpus as a whole. Some attractive areas includes anaphoric reference, modelling users' expertise level etc.

Speech Processing

Dialogue system is an interesting area to work on. The Dialogue system itself can be considered both as an application as well as a system area because it requires many language processing tasks like parsing, entity name recognition etc. Although the complexity of text dialogue system can't be compared with speech dialogue system but in general it is generally considered speech dialogue system is more complex than text system. Google's now, Apples Siri, and Microsoft's Cortana are virtual assistants that inherently recognizes human speech.

Opinion mining or Sentimental analysis can also be done on speech content. For a given set of feedbacks for a product in the speech form, the task is to detect whether the person is having positive, negative or neutral opinion about the product. Similar to text mining sentimental analysis, emotional analysis etc can also be done on speech documents. In emotional analysis, the task is to find the emotion of the speaker i.e whether they are in happy or sad. There are differences between text emotional analysis and speech emotional analysis as the text EA can be done only based on the words in the passage whereas many emotions are expressed in speech like boos, jeers or whining etc which are quite difficult to express in written form.

Like topic capturing in text, identifying the theme of the monologue or a dialogue is challenging. Resolving those issues for a smooth topic finding from a speech is attractive field to work with. Other simple but attractive semantic level speech processing tasks like identifying the gender of the speaker or ethnicity or character of the speaker are all attractive fields to work and requires a range of features to be extracted from the speech content. Processing speech files that contains audio from two or more different languages is challenging and thus an interesting area to work. There are many areas in text to speech synthesis that remains unresolved. Finding problems in the speech synthesis will help to develop programs that can speak with humans. Thus it helps in achieving one of the important goal of NLP namely language generation.

Like topic capturing in text form, for a piece of speech, an algorithm has to detect what the speaker is tasking about (topic). While this itself is difficult, if the speech is dialogue instead of monologue, identifying the theme or topic of the conversation is much more difficult. This problem is very much related to speech dialogue system where the task is to strike a dialogue with the user. This dialogue system may a simpleton like simple speaking with user or an intelligent that provides suggestion or solution or answer for the user questions as well along with the dialogue. Thus dialogue system is very much related to question and answering system either in text or in speech form. In speech form, question and answering system involves getting questions in the speech form and providing answers for those question in the same speech form.

Machine translation can also be in speech form where the sentence or passage is in the audio form and the output can be either in text or in audio form. Such applications are paramount importance for people who need to engage with foreigners who will be speaking a different language.

Computational humor [8] is another interesting research area where the main task is to identify humor in the text passage or to generate humor in a given context. The very notion of giving the ability for the computers to understand humor and react to it shows that this field is another milestone in artificial intelligence which aims to simulate human intelligence.

In general, many language processing tasks are being done by leveraging machine learning algorithms. The choice between supervised, semi-supervised or Un-supervised depends on factors like availability of training data, nature of the language processing task etc.

II. CONCLUSION AND FUTURE WORK

In this paper, we discussed a number of familiar research areas that have potential to work in language processing. With the advent of machine learning and deep learning, the landscape of language processing research has underwent a paradigm shift, Rule based

language processing tasks that require writing cumbersome and complex rules is replaced with using machine learning or deep learning algorithms that can learn from the training set by itself without need to explicitly programmed.

This paper delves deeper into various research areas in language processing. Although this paper touches upon some main areas in NLP, there are many other areas that are equally enticing as the above discussed areas that may have not touched upon or would have researched less. It is also important to know that there are research areas that are specific to one language that may not be suitable for another language. Identifying such problems and solving them is also promising area for budding researchers. The ultimate pursuit of research in NLP is aiming for developing systems that can understand human language and that can interact with humans like a fellow human.

III. ACKNOWLEDGMENT

We would like to profusely thank departmental colleagues and management of our college for carrying out this brief survey on various research areas in Natural Language Processing.

IV. REFERENCES

[1] James Allen, *Natural language understanding*. Pearson, 1995.

[2] Pushpak Bhattacharyya, "Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality." *CSI journal of computing* 1.2 ,2012, 1-13.

[3] Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 , 1995, 39-41.

[4] B Ela Kumar, *Natural language processing*. IK International Pvt Ltd, 2011.

[5] start.csail.mit.edu/

[6] El-Assady, Mennatallah, et al. "Interactive visual analysis of transcribed multi-party discourse." *Proceedings of ACL 2017, System Demonstrations* ,2017, 49-54.

[7] Sharma, V. K., & Mittal, N. (2018). Cross-Lingual Information Retrieval: A Dictionary-Based Query Translation Approach. In Advances in Computer and Computational Sciences (pp. 611-618). Springer, Singapore.

[8] Binsted, Kim, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G. Ritchie, R. Manurung, H. Pain, Annalu Waller, and D. O'Mara. "Computational humor." IEEE Intelligent Systems 21, no. 2 (2006): 59-69.