Volume 6, Issue 10, October 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Study of the Natural Language Processing Tasks to Address Semantics Ambiguities

Partha Sarkar
Research Scholar,
Department of Computer Science,
Assam University, Silchar, Assam, India

Bipul Syam Purkayastha
Professor,
Department of Computer Science,
Assam University, Silchar, Assam, India

Abstract: Natural Language Processing (NLP) is a kind of human-computer interaction where the elements of human language, be it spoken or written, are formulated so that a computer can perform tasks based on that interaction. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyze, understand, and generate languages that humans use naturally, so that eventually it will be able to address human communication. This goal is not easy to reach. Understanding language means knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. It is ironic that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master. Although machines are capable of inverting large number of data with speed and grace, they still fail to master the basics of our spoken and written languages. The obvious problems arise from the semantic ambiguities which in most of the cases becomes difficult to present through a software program. As an English speaker we can effortlessly understand a sentence like "My mind is flying in joy". However, this sentence presents difficulties to a software program that lacks both our knowledge of the world and our experience with linguistic structures. This paper is attempted to analyze the problems which can be well met by using a mix of knowledge-engineered and statistical/machine-learning techniques to disambiguate and respond to natural language input.

Keywords: Interaction, Linguistic structures, NLP input, Software, Semantic ambiguities, Value-adding tasks.

I. INTRODUCTION

Natural Language Processing (NLP) is a modern computational technology and a method of investigation and evaluation. Some prefer the term ‘computational linguistics’ in order to emphasize this latter function, but NLP is a term that links the history of Artificial Intelligence (AI) and the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations. Knowledge representation here means the need for representations of our knowledge of the world in order to understand human language with computers. In other words, Natural Language Processing (NLP) is the use of computers to process written and spoken language for some practical, useful, purpose: to translate languages, to get information from the web on text data banks so as to answer questions, to carry on conversations with machines, so as to get various information. These are only examples of major types of NLP, and there is also a huge range of lesser but interesting applications, e.g. getting a computer to decide if one newspaper story has been rewritten from another or not. NLP is not simply applications but the core technical methods and theories such as Machine Learning techniques, which is automating the construction and adaptation of machine dictionaries, modeling human agents' beliefs and desires etc. The goal of Natural Language Processing (NLP) is to design and build a computer system that will analyze, understand, and generate natural human-languages. Applications of NLP include machine translation of one human-language text to another; generation of human-language text such as fiction, manuals, and general descriptions; interfacing to other systems such as databases and robotic systems thus enabling the use of human-language type commands and queries; and understanding human-language text to provide a summary or to draw conclusions. It is comparatively easy for an NLP system is to parse a sentence to determine its syntax but a more difficult task is determining the semantic meaning of a sentence.

II. SEMANTICS AND NATURAL LANGUAGE PROCESSING TASKS

The entire purpose of a natural language is to facilitate the exchange of ideas among people about the world in which they live. These ideas converge to form the "meaning" of an utterance or text in the form of a series of sentences. The meaning of a text is called its *semantics*. Semantics and its understanding as a study of meaning covers most complex tasks like: finding synonyms, word sense disambiguation, constructing question-answering systems, translating from one NL to another, populating base of knowledge. Basically one needs to complete morphological and syntactical analysis before trying to solve any semantic problem. It is quite obvious that in order to solve complex NLP tasks, especially related to semantic analysis, we need formal representation of language i.e. semantic language. From the syntactic structure of a sentence the NLP system will attempt to produce the logical form of the sentence. Logical form is context-free in that it does not require that the sentence be interpreted within its overall context in the discourse or conversation in which it occurs. And logical form attempts to state the meaning of the sentence without reference to the particular natural language. Thus the intent seems to be to make it closer to the notion of a proposition than to the original sentence. From the computational point of view, no general solutions that would be adequate have been proposed for this area. A wide range of methods/tasks can be implemented to accomplish the disambiguation, some which require information as to the frequency with which each sense occurs in a particular corpus of interest, or in general usage, some which require consideration of the local context, and others which utilize pragmatic knowledge of the domain of the document. Following below is a brief study on that.

III. AUTOMATIC SUMMARIZATION

Automatic summarization is a method that produces a readable summary of a chunk of text. It is often used to provide summaries of text of a known type, such as articles in the editorial section of a newspaper. There are two types of summarization: extractive and abstractive. Extractive summarization techniques assume that the summary can be constructed by simply extracting the relevant sentences from the document or documents themselves. This assumption may not always hold but it seems to hold for news articles. So, most techniques in this area try to rank the sentences in the documents and figure out how to choose the most relevant ones in a non-redundant manner. Abstractive summarization, on the other hand, makes the assumption that a good summary needs to be constructed without using the sentences from the documents but instead by generating new sentences that contain the meaning and knowledge contained in the documents. As expected, this is particularly difficult. One of the ways people claim to achieve this is by doing what's called template-based natural language generation wherein a template for the summary is pre-defined (or generated based on the topic) and words and phrases are extracted from the documents to fill the blanks in the templates. I personally don't consider this abstractive summarization but just another form of extractive summarization.

IV. COREFERENCE RESOLUTION

Coreference resolution, on the other hand, is a method in which if given a sentence or larger chunk of text, determine which words refer to the same objects. Anaphora resolution is a specific example of this task, and is specifically concerned with matching up pronouns with the nouns or names that they refer to. For example, in a sentence such as “He entered John's house through the front door", "the front door" is a referring expression and the bridging relationship to be identified is the fact that the door being referred to is the front door of John's house. In simple words, coreference resolution is the process in which we identify the noun phrases that are referring to a same real-world entity. In this context, such noun phrases are called mentions, or just anaphoric noun phrases. Mentions can be named, nominal or pronominal.

V. DISCOURSE ANALYSIS

Discourse analysis is a interdisciplinary term which is concerned with "the use of language in a running discourse, continued over a number of sentences, and involving the interaction of speaker (or writer) and auditor (or reader) in a specific situational context, and within a framework of social and cultural conventions The meaning of any sentence depends upon the meaning of the sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence. During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of language which require real world knowledge. Discourse includes a number of related tasks. One task is identifying the discourse structure of connected text, i.e. the nature of the discourse relationships between sentences. Another possible task is recognizing and classifying the speech acts in a chunk of text.

VI. MORPHOLOGICAL SEGMENTATION

Morphological segmentation separate words into individual morphemes and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the morphology of the language being considered. English has simple morphology, especially inflectional morphology, and thus it is often possible to ignore this task entirely and simply modeling all possible forms of a word as separate words. In languages such as Turkish, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.

VII. NAMED ENTITY RECOGNITION (NER)

Named Entity Recognition (NER) determine from the text which items in the text match to proper names, such as people or places, and its type. Here it should be noted that, although capitalization can aid in recognizing named entities in languages such as English, this information cannot aid in determining the type of named entity, and in any case is often inaccurate or insufficient. For example, the first word of a sentence is also capitalized, and named entities often span several words, only some of which are capitalized. Furthermore, many other languages in non-Western scripts (e.g. Chinese or Arabic) do not have any capitalization at all, and even languages with capitalization may not consistently use it to distinguish names. For example, German capitalizes all nouns, regardless of whether they refer to names, and French and Spanish do not capitalize names that serve as adjectives.

VIII. CHUNKS

Natural language understanding convert chunks of text into more formal representations such as first-order logic structures that are easier for computer programs to manipulate. Natural language understanding involves the identification of the intended semantic from the multiple possible semantics which can be derived from a natural language expression which usually takes the form of organized notations of natural languages concepts. Introduction and creation of language metamodel and ontology are efficient however empirical solutions. An explicit formalization of natural languages semantics without confusions with implicit assumptions such as closed world assumption (CWA) vs. open world assumption(OWA), or subjective Yes/No vs. objective True/False is expected for the construction of a basis of semantics formalization.

IX. PART OF SPEECH TAGGING

Part-of-speech tagging , as the name gives clues, determines the part of speech for each word in a given sentence. Many words, especially common ones, can serve as multiple parts of speech. For example, "table" can be a noun (The table is made of wood) or verb (Table the files). However, languages with little inflectional morphology, such as English are particularly prone to such ambiguity. Chinese is prone to such ambiguity because it is a tonal language during verbalization. Such inflection is not readily conveyed via the entities employed within the orthography to convey intended meaning.

It is relevant here to take a bird's eye view on the task of part of speech tagging, i.e. determining the correct part of speech of each word in a given sentence. A typical machine-learning-based implementation of a part of speech tagger proceeds in two steps-a training step and an evaluation step. The training step makes use of a corpus of training data, which consists of a large number of sentences, each of which has the correct part of speech attached to each word. An example of such a corpus, common in use, is the Penn Treebank. This includes, among other things, a set of 500 texts from the 'Brown Corpus', containing examples of various genres of text, and 2500 articles from the 'Wall Street Journal'. This corpus is analyzed and a learning model is generated from it, consisting of automatically created rules for determining the part of speech for a word in a sentence, typically based on the nature of the word in question, the nature of surrounding words, and the most likely part of speech for those surrounding words. It not only works on the training data but also tries to be as simple as possible in analyzing the data. In the second step i.e. in the evaluation step, the model that has been learned is used to process new sentences. An important part of the development of any learning algorithm is testing the model that has been learned on new, previously unseen data. It is critical that the data used for testing is not the same as the data used for training; otherwise, the testing accuracy will be unrealistically high.

X. PARSING

The method of Parsing aims at grammatical analysis of a given sentence. The grammar for natural languages is ambiguous and typical sentences have multiple possible analyses. In fact, there is the possibility that for a typical sentence there may be thousands of potential parses. It is the process of assigning structural descriptions to sequences of words in a natural language. It assigns the structure and grammatical category of the sentence. In other words, a parser accepts as input a sequence of words (or their surrogates) in some language and an abstract description of possible structural relations that may hold between words or sequences of words in the language, and produces as output zero or more structural descriptions of the input as permitted by the structural rule set. There will be zero descriptions if either the input sequence cannot be analyzed by the grammar, i.e. is ungrammatical, or if the parser is incomplete, i.e. fails to find all of the structure the grammar permits. There will be more than one description if the input is ambiguous with respect to the grammar, i.e. if the grammar permits more than one analysis of the input. The input symbol sequence to a parser may or may not consist solely of words in a natural language. Parsing in NLP may be of word sequences, part-of-speech tag sequences, or of sequences of complex symbols. The idea of parsing is grounded on the belief that grammatical structure contributes to meaning and that discovering the grammatical structure of an NL word sequence is a necessary step in determining the meaning of the sequence. In some parsers the construction of a meaning representation is carried out in parallel with the derivation of a structural analysis according to the grammar.

XI. RELATIONSHIP EXTRACTION

Relationship extraction identifies the relationships among named entities when given a chunk of text, (e.g. who is the wife of whom). Many applications in information extraction, natural language understanding, and information retrieval require an understanding of the semantic relations between entities. Relations are the subject, action and object relations within sentences. It aims at parsing sentences into subject, action and object form and then adding additional semantic information such as entity extraction, keyword extraction, sentiment analysis and location identification. The computer needs to know how to recognize a piece of text having a semantic property of interest in order to make a correct annotation. Thus, extracting semantic relations between entities in natural language text is a crucial step towards natural language understanding applications.

XII. SENTIMENT ANALYSIS

In the domain of NLP, sentiment analysis, which is also known as opinion mining, refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. This algorithm takes an input string and assigns a sentiment rating in a range from very negative, negative, neutral, positive, and very positive. Sentiment extracts subjective information usually from a set of documents, often using online reviews to determine "polarity" about specific objects. It is especially useful for identifying trends of public opinion in the social media, for the purpose of marketing.

XIII. INFORMATION RETRIEVAL (IR)

Information retrieval (IR) is concerned with storage, searching and retrieving information. It is a separate field within computer science. The field of IR relies on some NLP methods like stemming. Some current research and applications seek to bridge the gap between IR and NLP. Information extraction (IE) is, on the other hand, concerned in general with the extraction of semantic information from text. This covers tasks such as named entity recognition, conference resolution, relationship extraction, etc. A textual information retrieval system carries out various tasks in response to a user's query. In the first phase of indexing the collection of documents, NLP techniques are applied to generate an index containing document descriptions. Normally each document is described through a set of terms that, in theory, best represents its content. In the second phase, when a user formulates a query, the system analyses it, and if necessary, transforms it with the hope of representing the user's information needs in the same way as the document content is represented. Then, the system compares the description of each document with that of the query, and presents the user with those documents whose descriptions are closest to the query description. The results are usually listed in order of relevancy, that is, by the level of similarity between the document and query descriptions.

XIV. MACHINE LEARNING ALGORITHMS

Machine learning has been studied from a variety of perspectives, sometimes under different names. Modern approaches to natural language processing (NLP) are grounded in machine learning. The area of machine learning is slightly different from that of most prior attempts at language processing. Where the prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules, the machine-learning paradigm, now a days, uses general learning algorithms, often supported by statistical inference, to automatically learn such rules through the analysis of large corpora. There are a number of different classes of machine learning algorithms which have been applied to NLP tasks. In common to all of these algorithms is that they take as input a large set of "features" that are generated from the input data. As an example, for a part-of-speech tagger, typical features might be the identity of the word being processed, the identity of the words immediately to the left and right, the part-of-speech tag of the word to the left, and whether the word being considered or its immediate neighbors are content words or function words. The algorithms differ, however, in the nature of the rules generated. Some of the earliest-used algorithms, such as decision trees, produced systems of hard if-then rules similar to the systems of hand-written rules that were then common. In recent years, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage that they can express the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system. In addition, models that make soft decisions are generally more robust when given unfamiliar input, especially input that contains errors.

XV. ADVANTAGE OF MACHINE-LEARNING ALGORITHMS

It has been found that the systems based on machine-learning algorithms have many advantages. Firstly, the learning procedures used during machine learning automatically focus on the most common cases, whereas when writing rules by hand it is often not obvious at all where the effort should be directed. Secondly, automatic learning procedures can make use of statistical inference algorithms to produce models that are robust to unfamiliar input and erroneous input. Generally, handling such input properly with hand-written rules is extremely difficult, error-prone and time-consuming. Thirdly, in systems, based on automatically learning, the rules can be made more accurate simply by supplying more input data. However, systems based on hand-written rules can only be made more accurate by increasing the complexity of the rules, which is a much more difficult task. In particular, there is a limit to the complexity of systems based on hand-crafted rules, beyond which the systems become more and more unmanageable. However, creating more data as input to machine-learning systems requires man-hours worked. The study of the computational processes underlying comprehension and generation of natural language is an important scientific and engineering task. Instilling machines with abilities that allow them interact to intelligently with humans depends on increasing levels of natural language comprehension or generation, although shallow levels of understanding can already successfully support challenging applications such as intelligent information extraction, automatic translation, summarization and others.

XVI. CONCLUSION

In recent years, many of the notable successes occurred in the field of machine translation. These systems were able to take advantage of existing multilingual textual corpora. Research in recent times has increasingly focused on unsupervised and semi-supervised learning algorithms. Generally, this task is much more difficult than supervised learning, and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available which can often leads to inferior results. It is generally accepted today that a statistical machine learning component must have a central role in supporting natural language related tasks. A significant amount of work, thus, has been devoted in the last few years to develop statistics based machine learning methods for these tasks, with considerable success.

ACKNOWLEDGMENT

This is to humbly acknowledge that this research paper was supported by Prof. Bipul Shyam Purkayastha, Professor, Department of Computer Science, Assam University, Silchar, India who provided insight and expertise that greatly assisted the research. We thank for the support and help by the Library of Assam University, a Central University for providing the books, articles as well as the web support whenever needed.

REFERENCES

[1] Andreas, Steve, and Faulkner, Charles, eds. *"NLP: The New Technology of Achievement".* William Morrow Paperbacks, February 19, 1999.Print.

[2] Baeza-Yate, Dr Ricardos, Ribeiro-Neto , Dr Berthier , eds. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd ed , Addison Wesley, 23 December 2010.Print.

[3] Bender, Emily, M., *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax (Synthesis Lectures on Human Language Technologies).* Morgan & Claypool Publishers , June 1, 2013.Print.

[4] Bishop, Christopher . *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 2nd ed, Springer, 15 February 2010. Print.

[5] Charniak, E. "*Statistical Language Learning"*. MIT Press, 1993. Print.

[6] Charniak, E., *"Tree-bank grammars".* Technical report, Department of Computer Science, Brown University, 1996. Print.

[7] Charniak, E."*Statistical Parsing with a Context-free Grammar and Word Statistics".* Proceedings of the Fourteenth National Conference on Artificial Intelligence, 1997. Print.

[8] Clark, Alexander, Fox, Chris , and Lappin, Shalom eds. *The Handbook of Computational Linguistics and Natural language Processing* , 1st ed, Wiley-Blackwell, October 4, 2012. Print.

[9] Collins, M. "*Head-driven statistical models for natural language parsing".* Computational Linguitics, 2003. Print.

[10] Daniel, Jurafsky. *Speech And Language Processing*, Dorling Kindersley India, 2008.Print.

[11] Forman, G. "An extensive empirical study of feature selection metrics for text classification". *Journal of Machine Learning Research.* 2003: 78-91. Print.

[12] Grune, Dick , and Jacobs, Ceriel J.H., eds. *Parsing Techniques: A Practical Guide (Monographs in Computer Science)*, 2nd ed, Springer, 13 December 2007. Print.

[13] Kubler, Sandra , McDonald, Ryan , and Nivre ,Joakim , eds. *Dependency Parsing (Synthesis Lectures on Human Language Technologies)*, Morgan and Claypool Publishers, January 23, 2009. Print.

[14] Manning, D., Christopher. , Raghavan, Prabhakar , and Schütze, Hinrich eds. *Introduction to Information Retrieval* , Cambridge University Press , 7 July 2008.Print.

[15] Manning, Christopher . *Foundations of Statistical Natural Language Processing*, MIT Press, 30 July 1999.Print.

[16] Palmer, Stone, Martha. *"Semantic Processing for Finite Domains (Studies in Natural Language Processing)"* 1st ed. Cambridge University Press, February 13, 2006. Print.

[17] Roark, Brian, and Sproat , Richard eds. *Computational Approaches to Morphology and Syntax (Oxford Surveys in Syntax & Morphology)*, Oxford University Press, September 27, 2007. Print.

[18] Vladimir. A., and Fomichov eds., *"Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms",* Springer, December 4, 2009. Print.

[19] Vaknin, Shlomo. *"NLP For Beginners: Only The Essentials,* Inner Patch Publishing, July 25, 2009. Print.