A Study on Deep Linguistic Processing with Special Reference to Semantic and Syntactic Levels

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

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

ABSTRACT

Natural Language Processing (NLP) sets a relation between human and computer where the elements of human language, be it spoken or written, are organized so that a computer can perform tasks accordingly based on their interaction. The goal of the Natural Language Processing (NLP) is to design and make software that will help to analyze, understand, and generate languages that humans use naturally, so that in the long run we will be able to address our computer according to our convenience. This goal is not easy to reach because the natural language, the symbol system, that is easiest for humans to learn and use, is hardest for a computer to master and interpret in a meaningful way. Though machines today are capable of inverting large matrix with speed and grace, they still fail to master the basics of our spoken and written languages. The obvious problems arise from the semantic and syntactic ambiguities which in most of the cases becomes difficult to present through a software programme. As an English speaker we can effortlessly understand a sentence like "My mind is flying in joy". But this sentence presents difficulties to a software program that lacks both our knowledge of the world and our experience with linguistic structures. Deep Linguistic Processing, in this connection, is an important area of study to achieve this goal.

Keywords

Ambiguities, Deep Linguistic Processing, Interaction, Symbol-system, Semantic, Syntactic

1. INTRODUCTION

Natural language processing today has become a broad area to study and research. As found in the many definitions of NLP [1] it is said that Natural Language Processing (NLP) is both a modern computational technology [14] and a method of investigating and evaluating claims about human language itself. NLP is a term that links the areas of Artificial Intelligence (AI), the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations with the aim of representing 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 analyze and process written and spoken language for some practical, useful, purpose such as to translate languages, to get information from the web on various text data , to carry on conversations with machines, so as to get advice etc. NLP is not simply theories but also the technical approaches, methods and applications aim at representing natural languages in proper way. The goal of Natural Language Processing (NLP) is to design and build a computer system that will analyze, understand, and generate natural human languages. Deep Linguistic Processing, in this connection, is an important area of study to achieve this goal.

2. DEEP LINGUISTIC PROCESSING

Deep Linguistic Processing is a natural language processing framework which is closely related with syntactic and semantic theories like CCG, HPSG, LFG, TAG and the Prague School etc. However, to know about Deep Linguistic Processing, it is very important to know the difference between Deep Linguistic Processing and Swallow Natural language processing methods. The Deep Linguistic Processing approaches differ from shallower methods in that they give richer, more expressive, structural representation. The considerable knowledge of computational power forms the basis of the deep linguistic processing approach. Deep linguistic processing has traditionally been related with computational grammar development used in case of parsing and corpora analysis. However, the application of these grammars was syntactically and semantically complex to present and expensive to run. But the development of machine learning approaches and techniques have given a pace to the field of natural language processing. The research and invention of forceful and technically innovative NLP tools have substantially decreased the amount of manual labor. However, the fact cannot be denied that in order for computers to understand natural language or inference, detailed syntactic and semantic representation is necessary. Moreover, shallow methods may lack human language 'understanding'. While humans can easily understand a sentence and its meaning, shallow linguistic processing might lack human language 'understanding' and thus it always gets difficult for machine translation to give a proper language output. For example, “My mind is flying out of joy”. In this sentence, a shallow information extraction system might infer wrongly that the mind is actually flying. While as humans, only we can understand that a mind cannot fly because mind does not have wings. The word ‘fly’ is used only to refer the condition of excess joy in human mind. In short, we can say that while deep linguistic processing provides a deep knowledge based analysis of language through manually developed grammars and language resources, shallow linguistic processing provides only the surface analysis of the language through statistical or machine learning usage of texts or annotated linguistic resource. However, to provide this deep knowledge based analysis of language deep linguistic approach largely tends towards the semantic and syntactic levels of language.

3. THE SYNTACTIC LEVEL OF LANGUAGE

The syntactic level of language learning focuses on analyzing the words in a sentence to give a logical explanation of the grammatical structure of the sentence. This requires both a grammar and a parser. This level of language learning aims at revealing the relationship of structural dependency between the words. There are various grammars that can be utilized, and which will, in turn, impact the choice of a parser. All NLP applications do not require the full parsing of sentences. However, the challenges in parsing of prepositional phrase attachment and conjunction scoping no longer confuse those applications for which phrasal and clausal dependencies are sufficient. Syntax conveys meaning in most languages because order and dependency contribute to meaning. For example, the two sentences: 'The cat chased the mouse' and 'The mouse chased the cat' differ only in terms of syntax but convey quite different meanings.

4. THE SEMANTIC LEVEL OF LANGUAGE

The semantic level contributes to the meaning of the sentence. Semantic processing [11] determines the possible meanings of a sentence by focusing on the interactions among word meanings in the sentence. This level can include the semantic disambiguation of words [8] with multiple senses and in the same way it shows how syntactic disambiguation of words which works as multiple parts-of-speech is accomplished at the syntactic level. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation [13] of the sentence. For example, amongst other meanings, 'file' as a noun can mean either a folder for storing papers, or a tool to shape one's fingernails, or a line of individuals in a queue. If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would do the disambiguation. A wide range of methods 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.

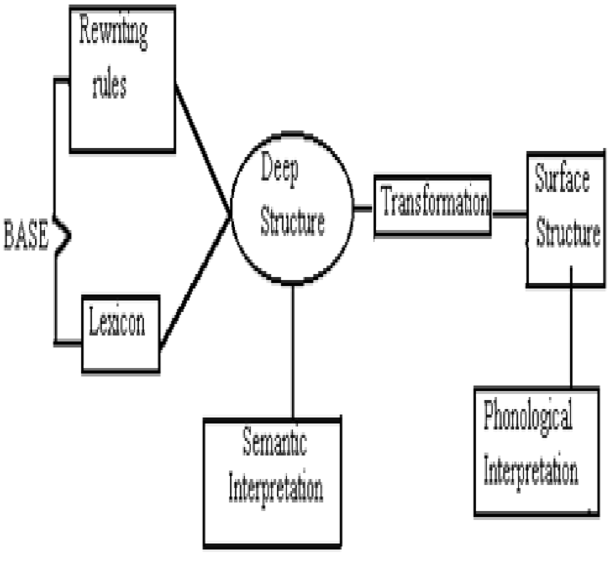

Fig 1: Semantic Interpretation

5. THE APPROACHES AND METHODS USED IN DEEP NLP

Deep semantic analysis of texts may prove more appropriate to extract not only concepts but also relationships and axioms. Shallow methods may fell short when confronted with more complex understanding of texts. Computational semantics deals with such aspects, and aims at producing meaning representations while tackling very fine-grained aspects of the language such as anaphora resolution, quantifier scope resolution, etc. For example, anaphora resolution identifies the entities referred by expressions such as pronouns. In general, computational semantics aims at grasping the entire meaning of sentences and discourse, rather than focusing on text portions alone. Computational semantics provides a method for extracting representations from natural language. Following below is a description of some of the approaches and methods used in Deep NLP.

5.1.Syntactic Parser

The first essential component for a deep analysis of texts is a syntactic parser based on syntactic grammars and drawing inferences based on these representations. Syntactic parsing is performed using a set of grammar rules that assign parse trees to sentences. This set of rules is known as a syntactic grammar. Traditional syntactic parsing relies on a lexicon that describes the vocabulary that may be used in the parsed sentences. For instance, Word Net can be considered as a lexicon. There are also statistical parsers [5] which learn their knowledge about a language using hand labeled sentences and produce the most likely analyses when parsing sentences. The output representations can take the form of phrase structure tree representations or dependency parses. Phrase structure parses associates a syntactic parse [10] in the form of a tree to a sentence, while dependency parses creates grammatical links between each pair of words in the sentence. Phrase structure grammars and dependency grammars cannot be considered as opposite approaches but rather as complementary. In fact, many syntactic theories make use of both formalisms. However, despite the fact that these two representations differ from each other only by what is explicitly encoded. Practical experience has shown that it is a non-trivial task to perform an automatic conversion [3] from one type of representation to the other. Dependency parsing seems to regain its central place in the research community with many researchers arguing that dependencies model predicate argument structures in a more convenient or intuitive form for further semantic analysis and that dependency grammar has led to the development of accurate syntactic parsers using machine learning on 'Treebanks' . The majority of the approaches are based on the exploitation of the syntactic parses to extract relevant structures, using patterns and machine learning. These approaches use syntactic parses for a more refined term extraction, relation extraction and axiom learning.

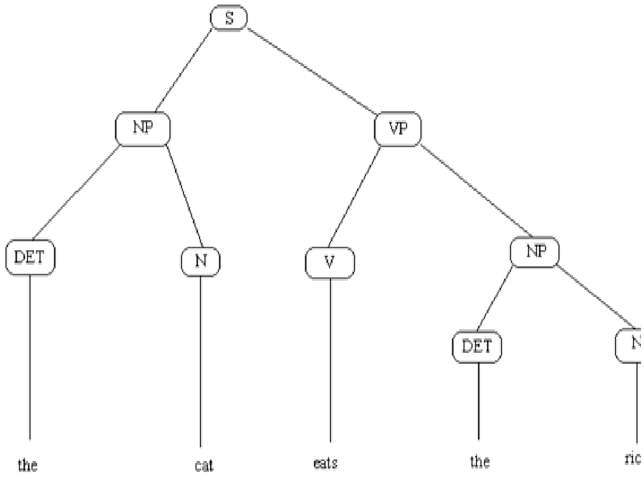

Fig 2: Syntactic Parser

5.2. Linguistic Grammars

Traditionally, the use of grammars relies mainly on lexical items. Linguistic grammars here refer to a number of syntactic rules and a lexicon to produce some kind of representations of the sentence or the discourse. These grammars include HPSG grammars and CCG grammars, which are among the most used formalisms in the linguistic community. Linguistic grammars rely basically on two main methods: the unification-based approaches and the lambda-calculus approaches, which may be considered as drawing their roots in formal semantics. Many grammars, such as HPSG, are based on the unification mechanism. The unification based-approaches rely on representations called feature structure that expresses valid phrase structure rules between syntactic categories like NP and VP and use lexical entries which is also described as feature structures as their content. The phrase structure rules have the form LHS/RHS where LHS is a non lexical category and RHS might be a set of lexical and non lexical categories. These rules guarantee the correct unification of all the semantic features with instances that respect the syntactic features. A lexical item is represented as phoneme coupled with a syntactic and a semantic part.

5.3.The Use of Lambda-Calculus

The use of lambda-calculus is also very popular in this connection. It is used as the glue to combine semantic representations, and solve some of the problems of the unification-based approaches such as coordination processing. In lambda-calculus, all the formulas are regarded as lambda expressions, as well as the combination of variables and lambda expressions. All the approaches described so far tackle the semantic analysis at the sentence level. The Discourse Representation Theory (DRT) was invented to address this shortcoming by providing means to parse a discourse. More precisely, DRT enables resolving pronouns to their textual antecedents. The basic unit of the DRT is the Discourse Representation Structure (DRS) which map sentence constituents to objects, properties and predicates and provides discourse referents or variables to represent these objects. These variables are then used to resolve a pronominal anaphora. The DRT provides triggering rules associated to particular syntactic configurations as well as transformation methods that output semantic representations.

6. SIGNIFICANCE OF DETAILED SYNTACTIC REPRESENTATION

It is true that the applications that are used for natural language understanding or inference will ultimately need detailed syntactic representations from which proper semantic interpretations can easily be made. There is already some evidence that the current, popular deep techniques can, in some cases, surpass shallow approaches. Several workshops and projects are going on demonstrating this in question answering, targeted information extraction and mention can be made to recent textual entailment recognition task, which is perhaps most notable in machine translation. In the area of machine translation, after a period of little use of linguistic knowledge, deeper techniques are beginning to give better performance, for example, redefining phrases by syntactic “treelets” rather than adjoining word sequences, or including a syntactic component in the probability model, or syntactic preprocessing of the data. In recent times, the divide between “deep”, rule-based, methods and “shallow”, statistical, approaches, is diminishing from both sides. Recent advances in case of Treebanks to extract more expressive grammars and development in framework-specific Treebanks [4] have made it possible to obtain similar coverage, sturdiness and accuracy for parsers that are used in richer structural representations. It is seen from the current research work that a large proportion of current deep systems have statistical components to them, for example, as pre- or post-processing to control ambiguity, as means of acquiring and extending lexical resources, or even use machine learning techniques to acquire deep grammars automatically. Moreover, many of the purely statistical approaches are using increasingly richer linguistic features and are taking advantage of these expressive features to tackle problems that were traditionally thought to require deep systems, such as the recovery of traces or semantic roles.

7. THE NEED FOR MANUAL GRAMMAR DEVELOPMENT

Although statistical techniques are becoming commonplace even for systems built around handwritten grammars, there is still a need for further linguistic research and manual grammar development. For example, supervised machine-learning approaches rely on large amounts of manually annotated data. Where such data are available, developers of deep parsers and grammars can exploit them to determine frequency of certain constructions, to bootstrap gold standards for their systems, and to provide training data for the statistical components of their systems such as parse disambiguators. But for the majority of the world’s languages, and even for many languages with large numbers of speakers, such corpora are unavailable. Under these circumstances, the need for manual grammar development is unavoidable, and recent progress has allowed the underlying systems to become increasingly better engineered, allowing for more rapid development of any given grammar, as well as for overlay grammars that adapt to particular domains and applications and for porting of grammars from one language to another. Despite recent work on multilingual parsing, it is still the case that most research on statistical parsing is done on English, a fixed word-order language where simple context-free approximations are often sufficient. It is unclear whether our current models and algorithms carry over to morphologically richer languages with more flexible word order, and it is possible that the more complex structural representations allowed by expressive formalisms will cease to remain a luxury.

8. CONCLUSION

To conclude, it can be said that further research is required on all aspects of deep linguistic processing, including noble linguistic analyses and implementations for different languages, formal comparisons of different frameworks, efficient parse and learning algorithms, better statistical models, innovative uses of existing data resources, and new inventions of tools and methodologies.

9. REFERENCES

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

[2] D.W. Aha, D. Kibler, and M. K. Albert, 1991 "Instance-based Learning Algorithms: Machine Learning", ISSN 0885-6125.

[3] C. Apt´e, F. Damerau, and S. M. Weiss, 1994 "Automated Learning of Decision Rules for Text Categorization", ACM Transactions on Information Systems, ISSN 1046-8188.

[4] E. Charniak, 1996 "Tree-bank Grammars", Technical report, Department of Computer Science, Brown University.

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

[6] M. Collins, 2003 "Head-driven Statistical Models for Natural Language Parsing", Computational Linguitics.

[7] M. Davy and S. Luz, Dec. 2008 "An Adaptive Pre-filtering Technique for Error-reduction Sampling in Active Learning", International Conference on Data Mining Workshops, pp 682–691, IEEE Press, Pisa.

[8] G. Escudero, L. M`arquez, and G. Rigau, 2000 "Boosting Applied to Word Sense Disambiguation", R. L. D. M´antaras and E. Plaza, eds, Proceedings of ECML-00, 11th European Conference on Machine Learning, pp 129–141, Barcelona, Springer Verlag.

[9] G. Forman, 2003 "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research, ISSN 1533-7928.

[10] M. Haruno, S. Shirai, and Y. Ooyama, 1999 "Using Decision Trees to Construct a Practical Parser", Machine Learning, pp. 131–149.

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

[12] T. Joachims, 1998 "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", in Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137–142, Chemnitz.

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

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