Evaluation of Parsing Techniques in Natural Language Processing

Ankita Nohria[1], Harkiran Kaur[2]
[1]Research Scholar, Department of Computer Science and Engineering, TIET, Patiala (PB), India
[2]Faculty,Department of Computer Science and Engineering, TIET, Patiala (PB), India
Department of Computer Science and Engineering
Thapar Institute of Engineering and Technology (Deemed to be University), Patiala (PB), India

Abstract

Human languages are handled in different ways at different levels, such assyntactic analysis at sentence level, semantic analysis at meaning level, discourse analysis at text level and morphological analysis at word level. Syntactic analysis or parsing is an important application in field of Natural Language Processing (NLP) as it helps in determining underlying meaning and depicting the output as linking of sentences to each other. Parsing is performed to determine the grammatical structure, marking the parts-of-speech and specifically to remove ambiguity. Though there is no way to remove ambiguity completely, parsers partially remove it. There are various types of ambiguity and for each type; there are different ways or methods to handle. This paper presents the different approaches followed for parsing across different languages such as Chinese, Arabic, Mongolian and Hindi. Some of them followed the static approach while some followed dynamic approach. In case of machine learning models, some of these languages followed supervisedlearning model while others followed unsupervised model. Their comparisons have been laidto create a parser or word sense classifier or translator or Finite State Automata(FSA) parser, each aiming to achieve maximum accuracy and optimum results.

Keywords

Natural Language Processing, Parsing, Bottom-up Parser, Rule-based approach, Statistical Parser

I. INTRODUCTION

Humans need a means to communicate and use language as primary means of communication. In order to express their feelings and emotions, theyuse words or gestures to make others understand the meaning. They have numerous ways to represent their feelings that they hardly heed any thought of the process going simultaneously. In a similar way, machines also need to be trained to understand their language in this modern era. Natural Language Processing is one such way to compute the models to understand human language. The computational models help in developing automated tools to have better understanding of human language. The basic application of NLP is training the machine in natural languages so that it can effectively behave more in humanly manner. After training, the processed data is represented using finite state automata, parse trees, etc. However, while representing, there can be ambiguity at word level or phrase level. This isbecause; in general,there can be more than one meanings of single word or vice-versa. There are three types of ambiguities identified namely lexical ambiguity which arises at word level, syntactical ambiguity which arises at sentence level and referential ambiguity when meaning is not well referred.

The word "syntax" in natural language, "refers to the grammatical arrangement of words inasentence and their relationship with each other". The main aim is to find the syntactic structure of the sentence through syntactic analysis. The syntactic structure is generally represented in the form of tree who senodes are the phrases and the leaves corresponds to words of the languages. The process of identifying the syntactic structure of the sentence is called syntactic parsing or simply parsing. Syntactic parsing can also be defined as the process of assigning 'phrasemarkers' to a sentence. Syntactic analysis or parsing is used to determine the meaning of sentence.

II. LITERATURE SURVEY

Zhang, C. - X.et.al (2014)highlighted the importance of Word Sense Disambiguation(WSD) in machine translation. They have proposed a word sense classifier for Chinese language whose discriminative features are extracted from parsing tree using unsupervised approach to identify noun sense changes.The Bayesian classifier is also used for this purpose. Thus, the classifier has improved the accuracy and translation quality of machine. The authors have compared supervised and unsupervised methods for probabilistic measures and to identify noun sense changes. Various words and sentences have been cited in Chinese language whose parse trees have been generated using syntactic analysis. Part-of-Tagging Model, Constraint Word Extraction model and Syntactic Analysis models are discussed in brief for ambiguous words because Chinese words are processed through these models. The WSD is then integrated into machine translation. Consequently,

Chinese words are translated into English language. In order to evaluate the proposed method, they collected 120 Chinese sentences from Word Sense Disambiguation Corpus and divided the set into two parts. Then they conducted two experiments. In first experiment, words to the left and right of ambiguous words were extracted while in second experiment, the method was practiced to train the classifier. The accuracy for both experiments was checked and results were evaluated [3].

Hmeidi I. et. al (2016)attemptedto develop a tool for Arabic language to convert simple present and simple past sentences of Arabic into defined English called bi-lingual Machine Translation tool by using a dictionary to translate a word into desired output. Since there is great difficulty to remove the ambiguity in Arabic language and in the past years it has not been developed much because of syntactic and lexical differences, so the authors tried to automate the machine translation system for the user interfaces of same language. The algorithm they applied is that they divided the sentence into words according to space between them and stored them as primary entity. In the process followed which was word analysis, each word was analyzed to find similarity between given word and in lexicon, to compute present tag and to find directory to which it belongs to find its meaning directly using Bottom-up parser. Then the system finds right and apt forms of noun, verb and adjective phrases to combine them together to form correct sentences using Top-down parser. Then they have compared their translator with Al_Qafi translator which translated a sentence word by word and not at sentence level [4].

Wu R. et. al(2016) have proposed a "from-bottom-to-top" method for Mongolian rules to analyze the constituent of a sentence. The first and foremost step to analyze a constituent of sentence was Part-of-Speech (POS) tagging that was classifying the phrases and words on the basis of dictionary library and rule base. After tagging and all the preprocessing was done, the sentence was broken down into various modules on the basis of phrases, case, keywords and so on. All the modulesthen used "from-bottom-to-top" approach to identify and classify the sentence component. They followed the rule-based approach and the principal idea controlling this syntactic parsing method was to construct the Knowledge Base of Grammar by using dummy rules and executing the eradication of ambiguity of syntactic structure by constraint and checking [1].

Singla, D. and Kumar, P. (2017)have discussed discourse analysis at text level in Hindi language. One such process is Anaphora Resolution. It is a process in which "interpretation of an occurrence of one expression depends on the interpretation of an occurrence of another"to obtain right explanation of the text. The explanation of Entity anaphora includesrecognition of rightantecedent of a pronoun amongfeasible noun phrases. Using anentire table of pronominal form, the class of the pronoun was recognized.Then after the categorization, a set of rules which has been defined earlier was used to locate the antecedent. The five categories of pronominal forms included - Personal pronoun (the words which substitute for noun), Relative pronoun (words used to link two clauses sharing common word), Reflexive pronoun (pronouns preceded by referent in same clause), Indefinite pronoun (pronouns that refer to unspecified nouns) and Place pronoun (pronoun to refer to a place). The process workflow consists of five steps: processing the sentence, identifying anaphors, then applying rule based resolution, finding referents and then resolving the output. The categories pronominal references can be easily resolved using dependency structures. Then the rules for reflexive pronoun, spatial pronoun, first and second pronoun and relative pronoun have been formulatedcorresponding to the dependency structure which is described with various examples in the paper [2].

Zaki Y. et. al(2017)have intended to develop a new statistical parser for Arabic language. There were many parsers developed in same language but some of them followed the statistical approach because these parsers were based on utilizing supervised learning methods. Statistical parsing follows two steps - learning and parsing. A total of 5000 words Treebank 'ArabTag' had been created using annotation tool. Further, a morphological analyzer named tree adjoining grammar formalism had been used for syntactic parsing phase. This complete information was encapsulated inside a model of syntactic tree. After the encapsulation,these patterns were derived from the Treebank and stored in a patterns' base. An algorithm was developed by Zaki Y. et. al, in order to buildtheir parser.This algorithm used the patterns' base to identify and measure input sentences, then create the syntactic descriptions with precision f-score of 86.81%. A shallow parser Base Phrase Chunker was created by the authors (method of classifying adjoining words of same sentence consecutively). They used YAMCHA sequence model for POS taggingbased on SVM classifier. Then the model was trained and the accuracy of model was checked and it came out to be 96.13%. Three of them created a statistical parser which used Treebank based LFG resources leading to creation of another parser based on Supervised Learning Model thatwas named ARSYPAR. It used Support Vector Machine (SVM) and fewof the features from Corpus to learn Grammar rules. They did not want to use Annotated Corpus so they used the Comprehensive Arabic Corpus to form rules to analyze input sentences and used LSV algorithm for word segmentation[5].

III. COMPARATIVE ANALYSIS OF NLP PARSING TECHNIQUES

There are many methods and techniques for parsing which are followed to get desired output and maximum accuracy. Those techniques have been compared on the basis of approach followed and the corresponding processes.

TABLEI
COMPARATIVE ANALYSIS OF NLP PARSING TECHNIQUES

Papers	Approach	Process
[1]	WSD is integrated into machine translation.	collected 120 Chinese sentences from WSD corpus and divided the set into two parts. Conducted two experiments. In first, words to the left and right of ambiguous words were extracted while in second experiment, the method was practiced to train the classifier.
[2]	Bottom-up parser to find meaning and top-down parser to put phrases in correct form.	First, did the word analysis at sentence level using a dictionary and simple lexicon to translate a single word of Arabic language.
[3]	Followed Rule-based approach and bottom-up method	Followed "bottom-to-top method" to identify the constituents of sentence using traditional Mongolian character "that the predicate was generally at the end of the sentence".
[4]	Followed rule-based approach and used CPG based Anaphora resolution approach.	Firstly, sentence was passed into one side of Hindi Shallow parser which gives anaphors as output on the other side which were then detected using parser for which antecedents were recognized on the basis of rules defined.
[5]	Statistical parser based on supervised learning which uses SVM and some features from corpus.	First, converts input text to vector format, then use SVM algorithm for creating hyperplane equation needed for classification per group of words.

desired output.

IV. CONCLUSION

There exists a diversified language world which tends to a great possibility that one may not know all of them. So there needs to be some translators which can help people to understand other languages. It is difficult for many to categorize a word into different parts of speech or different forms and structures of sentences. While different approaches and methodologies have been made available for translators, POS taggers, FSA parsing, graphic visualization of parser, anaphora resolution parser, word sense classifier for different languages, but these all have limitations that they can be implemented on particular languages only. Each and every parser created till date, either works for Hindi, English, German, Arabic, Mongolian or other languages. Most of them follow Supervised Learning approach which means they train a model and then test it but it will always leave an exception. There has not been a parser, tagger or classifier yet available, which allows input in any language and output generated is the desired one. All of these work for specific language. However, in coming years, new approaches will be invented and followed which will take any input language and give the

ACKNOWLEDGEMENT

Ms. Harkiran Kaur is presently working as Faculty in Department of Computer Science and Engineering, Thapar Institute of Engineering and

Technology (Deemed to be University), Patiala, since December 2012. She has her contribution in numerous research papers, published in International Journals, National level Conferences and International level Conferences, in the areas of Data Analytics, Databases and NLP. She has also attended a number of workshops and faculty development programs. She can be reached at harkiran.kaur@thapar.edu.

Ms. Ankita Nohria is a BE student in the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala. She has worked on Web developments projects includingCold Storage Management System, Academic Monitoring System and Wildlife Explore and presently working in the area of Natural Language Processing. She can be reached at aankita1_be15@thapar.edu.

REFERENCES

[1] R. Wu, "From-Bottom-to-top" to Analyze Sentence Constituent of Traditional Mongolian Basing on The Rule," in *Proceedings of the IEEE International Conference on Information and Automation Ningbo, China, August 2016*, 2016.

[2] D. Singla and P. Kumar, "Rule Based Anaphora Resolution in Hindi," in *2017 International Conference on Computational Intelligence in Data Science(ICCIDS)*, 2017

[3] C.-X. Zhang, "Integrate Word Sense Disambiguation Based on Parsing Tree Into Machine Translation," in *International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014*, 2014.

[4] I. Hmeidi, "A Simple Present and Past Sentences Machine Translation from Arabic Language (AL) to English language," in *2016 International Conference on Engineering & MIS (ICEMIS)*, 2016.

[5] Y. Zaki, "Towards the Development of a Statistical Parser of Arabic Language," in *Computing Conference 2017 18-20 July 2017 | London, UK* , 2017.

[6] T. C. Tuck, "Natural Language Processing... Understanding what you say".

[7] Syntactic Parsing," Speech and Language Processing. Daniel Jurafsky & James H. Martin. Copyright

[8] "Parsing," Wikipedia