Clause Identification in English and Indian Languages: A Survey

Misha Mittal[1],[2]Abhilasha
GZS,Bathinda,Associate Professor(CSE),GZS Punjab Technical University,Bathinda.

ABSTRACT-*The problem of identification of clauses in natural language processing is to find away to identify the each type of clause, e.g., Dependent clause, independent clause. Clause identification plays very important role in preprocessing task for language processing activities. This paper reports about the Clause Identification proposed for various Languages like English, Malayalam, Bengali and Urdu. Various clause identification approaches like Hidden Markov Model (HMM), Support Vector Model (SVM), and Rule based approaches, Maximum Entropy (ME) and Conditional Random Field (CRF), Entropy Guided Transformation Learning (ETL), have been used for POS tagging.*

Introduction

Clause identification is a one of the shallow semantic parsing tasks consisting of splitting a sentence into clauses, which is important in various NLP applications such as Machine Translation, parallel corpora alignment, Information extraction and speech applications. Grammatically a clause is a group of words having a subject and a predicate of its own, but forming part of a sentence. Clause identification is a special kind of shallow parsing, like text chunking Nevertheless, it is more difficult than text chunking, since clauses can have embedded clauses. Clause information is important for several more elaborated tasks such as full parsing and semantic role labeling. Clause boundary identification of natural language sentences poses considerable difficulties due to the ambiguous nature of natural languages. Clause classification is a convoluted task as natural language is generally syntactically rich in formation of sentences or clauses.

Clause Identification is one of the activities performed in a typical word processing application.

Sentence Identification means to ensure that a given piece of text follows the grammar rules of the language in which it is written. Clause Identification systems for major foreign languages have been directly or in directly available but at present, no such system is available for any of the Indian languages. With the computers being widely used for day-to-day tasks of word processing, need for clause identification is being felt earnestly. Indian languages differ in many aspects from English and other European languages when it comes to the language structure. Natural language processing is a very young discipline in Punjabi. Therefore, there is a lack of basic resources and tools for processing the Punjabi language. The research work under study investigates the effectiveness and viability of using rule-based methods for analyzing the Punjabi language at morphology and syntax levels. For this purpose, and because this work has a practical focus, various developed tools, techniques, and resources for processing the Punjabi language have been used. The main tools that will be used include the morphological analyzer and generator, the part-of-speech tagger, the phrase chunker (modified).

Recent Work on Clause Identification Indian and English Languages:

Urdu : It is the language of Indo-Aryan group within the Indo-European family. It is spoken by more than 100 million peoples. It is the national language of Pakistan. It is widely spoken in many areas of India. Also it is officially recognized by the constitution of India [17]. Many researchers are working on the development of Urdu language. Many tools like POS tagger, phrase chunker etc has been developed and many tools yet to be developed. In case of clause identification and sentence simplification some work has been done. Mostly statistical techniques have been used for developing the tools.

D. Parveen et al.(2009)[5] have proposed Clause Boundary Identification using Classifier and Clause Markers in Urdu Language.They presented the identification of clause boundary for the Urdu language.They used Conditional Random Field as the classification method and the clause markers. The clause markers play the character to detect the type of sub-ordinate clause that is with or within the main clause. If there is any mis-classification after testing with different sentences then more rules are identified to get high recall and precision. Obtained results indicate that this approach efficiently determines the type of sub-ordinate clause and its boundary.POS tagging and chunking are the preprocessing steps which have been done manually here, so contain a great accuracy. The part-of-speech (POS) and chunked tagged corpus has been considered as input data. Initially machine learning approach is applied, within which linguistic rules are used.

D.Parveen have also done work on clause identification in 2011 by using CRF techniques.

D. Parveen et al.(2011) [4] have proposed a method based upon Conditional Random Fields(CRF) techniques for the identification of Clause boundary for Urdu language. In this work, the feature like current word, next word, previous word and Part-Of-Speech (POS) tag of current, next and previous words are used. To extract these features,linguistic rules are followed. Nine types of clause marker for marking nine different types of subordinate clauses are employed.Linguistic rules were used for misclassification. Author used a tagger and chunker for preprocessing of corpus.Author developed an algorithm that could mark the main clause as well as subordinate clauses.

English: It is the member of West Germanic language of the Indo-European language family . This language is closely related to Frisian, German, and Dutch languages. It was originated in England but now it is spoken on six continents. It is the primary language of the United States, the United Kingdom, Canada, Australia, Ireland, New Zealand, and various small island nations in the Caribbean Sea and the Pacific Ocean. It is also an official language of India, the Philippines, Singapore, and many countries in sub-Saharan Africa, including South Africa. English is the first choice of foreign language in most other countries of the world, and it is this status that has given it the position of a global lingua franca. It is estimated that a third of the world’s population, some two billion persons, now use English [16]. From all the known languages in the world, maximum work in the field of NLP has been done on English language. Most of the tools with different possible techniques have been developed for English language. In clause identification a lot of work has been done in English language.

KatsuhitoSudoh et al. (2010) [10] experimented that English is a typical Subject Verb Object

(SVO) language, while Japanese is a distinctive Subject Object Verb (SOV) language.However, Statistical Machine Translation (SMT)-0based translation form an SVO language to an SOV language does not work well because their word orders are completely different. The author proposed a method to rewrite SVO sentences to derive more SOV-like sentences by using a set of handcrafted rules and proposed an alternative single re-ordering rule: Head Finalization. This was a syntax-based preprocessing approach that offers the advantage of simplicity. Author did not concern about part-of-speech tags or rule weights because the powerful Enju parser allows him to implement the rule at a general level. The experimental results show that Head Final English (HFE) follows almost the same order as Japanese. Author also showed that this rule improves automatic evaluation scores.

Vinh Van Nguyen et al. [12] have proposed Conditional Random Fields (CRFs) framework for clause splitting problem. Two types of feature sets were used one is at word level and second at sentence level. In word level feature set a window of size 5 was used. This window may contain word form and Part-Of-Speech (POS) tag and chunking tag. In sentence level features score function of clause candidates which includes Punctuation marks, Coordinate Conjunctions, Relative Pronouns was calculated. On the basis of this score the clause is identified and separated. They used Penn Treebank corpus and CRF++ toolkit for this purpose. The output shows a Precision of 90.01%.

Abney (1990) [3] used a clause filteras a part of his CASS parser. It consists of two parts: one for recognizing basic clauses andone for repairing difficult cases (clauses without subjects and clauses with additional VPs). Ejerhed (1996) showed that a parser can benefit from automatically identified clause boundariesin discourse. Papageorgiou (1997) used a set of hand-crafted rules for identifying clause boundaries in one text. Leffa (1998) wrote a set of clause identification rules and applied them to a small corpus. The performance was very good, with recall rates above 90%. Orasan (2000) used a memory-based learner with post-processing rules for predicting clause boundaries in Susanne corpus. His system obtained F rates of about 85 for this particular task.

Georgiana Puscasu (2004) [9] proposed a multilingual method for detecting clause boundaries in unrestricted texts. Author used combination of language independent machine learning techniques and language specific rules in order to take the first step in building the hierarchical structure of sentences. Annotated corpus was used for the training of machine learning algorithm. The result obtained from this machine learning algorithm was then processed by a rule-based module. This module dealt with clause boundaries not included in the learning process author used formal indicators of coordination and subordination, together with verb type information (finite or non-finite) for identification of clause boundaries. Author tested this method on Romanian and English and the F-measure for clause start detection was 95% for Romanian and 92% for English.

ErifK.Tjong et al. [6] have proposed a memory based learner to CONLL-2001 shared task and the task has been divided into three parts in which first two parts identify the position of clause starts and ends given a word,with its part of speech tag and base chunk.Third part is for identifying embedded clauses. A list of heuristic rules is used for converting these positions to a consistent embedded clause structure.F value obtained is 66.67 on the test data of the third part of the shared task.

Erik F. Tjonget al. (2001) [7] have proposed a system for identification of clauses. Author employed machine learning approach for finding the clause boundaries in the text. The proposed system worked in three steps. First step is identification of the start of the clause, second step is the identification of the clause and the result obtained from these two steps is used to find the complete clause in the third step. On testing this system the author obtained an F rate of 78.63 for the third part of the shared task.

EraldoFrenandeset al. [8] have proposed Entropy guided transformation learning(ETL) method for clause identification. This was a machine learning method and was an extension of Transformation Based Learning(TBL) as it provides automatic template generation.Authors used the English corpus of CONLL-2001 task but the performance of ETL based system was not upto the mark as of the state-of-the-art approaches.Moreover modelingstrategy was also very simple as compared to state-of-the-art approaches.ETL uses information gain measure that select feature combinations to provide good template sets.ETL process takes place in two steps:In the first step,ETL uses decision tree induction to perform entropy guided template generation and in 2nd step,TBL algorithm was applied to learn a set of transformation rules.

Zhemin Zhu et al. (2010) [14] Author considered sentence simplification as a special form of translation with the complex sentence as the source and the simple sentence as the target. Author proposed a Tree-based Simplification Model (TSM), which, to his knowledge, was the first statistical simplification model covering splitting, dropping, reordering and substitution integrally. Author also describedan efficient method to train his model with a large-scale parallel dataset obtained from the Wikipedia and Simple Wikipedia. The evaluation showed that his model achieved better readability scores than a set of baseline systems.

Ani Thomas et al. (2011) [1] in tune with the recent developments in the automatic retrieval of text semantics, this paper was an attempt to extract one of the most fundamental semantic units from natural language text. The context was intuitively extracted from typed dependency structures basically depicting dependency relations instead of Part-Of-Speech tagged representation of the text. The dependency relations imply deep, fine grained, labeled dependencies that encode long distance relations and passive information. Apart from the typed dependencies, the present work did not take the help of Noun phrase Chunking tool or Part of speech Taggers for the compound noun phrase extraction.

NaushadUzZamanet al.(2011) [11] proposed a rule based system for the simplification of the sentences. This simplification was required to improve the machine translation system. The machine translation system from English to Tamil was developed by the author. This system lacks in accuracy because of problem in translating compound and complex sentences from English to Tamil language. To overcome this difficulty author proposed a system that will first identify the compound and complex sentences and then simply convert them to simple sentences. Handmade rules were used to develop this system.

Xavier Carreras et al. [13] proposed an approach consists in decomposing the clause splitting problem into a combination of binary “simple” decisions that can be solved using Ada Boost Learning Algorithm. Author decomposed the whole problem in two levels, with two chained decisions per level. Ada Boost is a general method for obtaining a highly accurate classification rule by combining many weak classifiers.

Bengali : Like other Indian languages Bengali is also a member of the Indo-Aryan group of the Indo-Iranian branch of the Indo-European language family. It is spoken by more than 210 million people as a first or second language, with some 100 million Bengali speakers in Bangladesh; about 85 million in India, primarily in the states of West Bengal, Assam, and Tripura; and sizable immigrant communities in the United Kingdom, the United States, and the Middle East. It is the state language of Bangladesh and one of the languages officially recognized in the constitution of India [15].

Aniruddha Ghosh et al. (2010) [2] have proposed a clause identification and separation system in Bengali language.A rule based approach was used for the identification of clause boundary and a conditional Random Field (CRF) based statistical model was used for identification of clause type. Four types of clauses i.e. Principal Clause, Noun Clause, adjective clause and Adverbial Clause were identified. The features used in the CRF were chunk label, chunk head and word itself. Authors were able to achieve an accuracy of 73.12 for clause boundary identification using rule based techniques and 78.07 for clause classification using CRF.

Scope of work:

Clauses are the building block of a sentence. Identification of clauses is one of the major necessities for identification of sentences. As every language have different types of sentences i.e. simple, compound and complex sentences. All these sentences are composed of different types of clauses. These different types of sentences can't be identified without identifying the subpart of sentence i.e. clause. The clauses also play a major role in machine translation. As in machine translation complex sentences are the most difficult to translate. These complex sentences need to be simplified before translation. So simplification of such complex sentences in to simple sentences requires identification of different types of clauses.

References

1. Ani Thomas,Kowar M K,Sharma Sanjay and Sharma H R (2011). "Extracting Noun Phrases in Subject and Object Roles for Exploring Text Semantics" , International Journal on Computer Science and Engineering (IJCSE) vol-3.

2. Aniruddha Ghosh, Das Amitava,BandyopadhyaySivaji(2009)."Clause Identification and Classification in Bengali" ,Department of Computer Science and Engineering Jadavpur University.

3. Abney(1990)

4. Daraksha Parveen, SanyalRatna, and AnsariAfreen (2011). "Clause Boundary Identification using Classifier and Clause Markers in Urdu Language".

5. DarakshaParveen, SanyalRatna, and AnsariAfreen (2009xc xc). "Clause Boundary Identification using Classifier and Clause Markers in Urdu Language",

6. ErifK.Tjong and Sang Kim (2001),"Introduction to the CoNLL-2001 Shared Task: Clause Identification".

7. Erik F. Tjong, SangKim,D´ejeanHerv´e(2001). "Introduction to the CoNLL-2001 Shared Task: Clause Identification",Proceedings of the 5th Conference on Natural Language Learning (CoNLL-2001) at the 39th Annual Meeting of the Association for

Computational Linguistics (ACL 2001)', Toulouse, France, pp. 52–57.

8. Eraldo R. Fernandes, A. Pires Bernardo,N. dos SantosC´ıcero,Milidi´uRuy L. (2009). "Clause Identification using Entropy Guided Transformation Learning".

9. Georgiana Pu¸sca¸su(2004). "A Multilingual Method for Clause Splitting".

10. KatsuhitoSudoh, WuXianchao, DuhKevin, TsukadaHajime, NagataMasaaki(2011), "ExtractingPre-orderingRulesfromPredicate-ArgumentStructures",Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 29–37

11. NaushadUzZaman , Jeffrey P. Bigham and James F. Allen (2011) 'Multimodal Summarization of Complex Sentences', IUI 2011, February pp. 13-16.

12. Vinh Van Nguyen, Nguyen Minh Le and ShimazuAkira, "Using Conditional Random Fields for Clause Splitting",Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 58–65.

13. Xavier Carreras and LluisMarquez(). "Boosting Trees For Clause Splitting".

14. Zhemin Zhu, Bernhard Delphine and GurevychIryna (2010), "AMonolingualTree-basedTranslationModelforSentenceSimplification" ,Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1353–1361,

15. http://www.britannica.com/EBchecked/topic/60785/Bengali-language

16. http://www.britannica.com/EBchecked/topic/188048/English-language

17. http://www.britannica.com/EBchecked/topic/619612/Urdu-language