Clause Boundary Identification for Different Languages: A Survey

Sanjeev Kumar Sharma
Assistant Professor
DAV University, Jalandhar

ABSTRACT: The problem of computing complex sentences in natural language processing is to make sentences simple to understand, by identifying clause boundaries as complex sentences are divided into clauses. Clause identification is very important task as many applications need simple sentences for computation. This paper reports about the identification of clauses in various languages. Each language uses its own methods and techniques for identification and making sentences simpler for use. Various techniques such as Conditional Random Field (CRF), Support Vector Model (SVM), Hidden Markov Model (HMM), Rule based approaches have been used for identification of clauses.

KEYWORDS:CRF, SVM, NLP, Stochastic, Corpus, Morphology.

1. INTRODUCTION

Clause identification is one of the major tasks in NLP, which is important in many NLP applications i.e. Machine Translation, Grammar Checking, Summarization etc. Clause is a combination of twowords including a subject and predicate i.e. a verb phrase or verb and forms part of a sentence. Clause identification is helpful in processing of complex sentences which is very challenging task [19]. The main goal of clause identification is to divide the sentences into clauses. The clauses can be identified as Principal clause i.e. a clause in a sentence do not depend upon another sentence and Sub-ordinate clause i.e. a clause which depends on the principal clause in a sentence [14]. There are three main types of Sub-ordinate clause and are Noun clause, Adjective clause and Adverb clause. Adverb clause is further divided into many types: Time, Place Manner, Reason, Purpose, Consequence (Result), Comparison, Condition and Contract (Supposition). Numerous techniques are there to identify the clause boundaries and type of clauses for different languages.

2. CLAUSE BOUNDARY IDENTIFICATION FOR DIFFERENT LANGUAGES

2.1 Hindi

Hindi is one of the widely spoken languages of India and is mainly spoken in Northern part of India. More than 258 million people on India proclaim Hindi to be their residential language. It is phonological rich language. A hybrid approach has been proposed by authors [13] for clause boundary identification in Hindi. In this work, hybrid approach locates the clause(s) in the sentence and marked the 'Clause Start Position' and 'Clause End Position'. This system comprises of two main parts; a stochastic model trained with 14500 sentences and hand crafted rules. In first module two different models i.e. step-by-step and merged models, using CRF machine learning approach was used. Both the models take word, its POS tag and its suffix as words feature for training. Common features used in identifying boundaries are:Present word's Lexicon, POS tag, last character, last two character and last three character, Previous four words' Lexicon and POS tags, Next four words' Lexicon and POS tags, Next three words' last character, last two character and last three character. Different rules are formalized to mark the boundaries of clauses to identify them. Out of 16000 sentences used by authors, 1500 sentences was selected, which are further divided into two more sets; development set consists of 500 sentences and testing set consists of 1000 sentences. Evaluation of the system was done with both models along with rules and perceived that system with step-by-step model performs well than system with merged model. Average accuracy shown by step-by-step model was 91.53% and the merged model shows 80.63%.

2.2 Bengali

Bengali is an Indo-European language. It is the national language of eastern South Asia and is one of the most spoken languages in the world. It has very long and rich literary tradition. Two different approaches were used for clause boundaryidentification [4]. A rule based model has been used to identify the clause boundary and CRF was used for clause identification. "Sanchay" was used as annotation tool for this work to identify the clause boundaries as well as the types of clauses. "Shallow Parser" was used as a linguistic tool for linguistic analysis of this paper. Dataset of 980 sentences have been used for training of the system, out of which 100 sentences were used for training. This system first identifies the verbs present in a sentence and then identifies the clause. After clause boundary identification, type of clauses was identified. Some features are used to identify type of clauses that are chunk label, chunk heads and word. Morphological information along with position of sentences was useful in clause type identification. In this work "Quadgram" technique was used. The clause information was in begin/inside/end i.e. B-I-E format. According to the results the accuracy of rule based clause boundary identification system was 73.12% and clause type identification was 78.07%.

2.3 Urdu

Urdu is the state language of Pakistan but also spoken by people in different countries all over the world. It is an Indo-Aryan language and is approximately twentieth most popular spoken language. A hybrid approach was used for clause boundary identification which includes two techniques i.e. Rule based and machine learning [10]. CRF has been applied in this paper. Clause boundary detection is a shallow-parsing technique. POS tagging and chunking have been done manually. The resulted corpus has been used as input data.

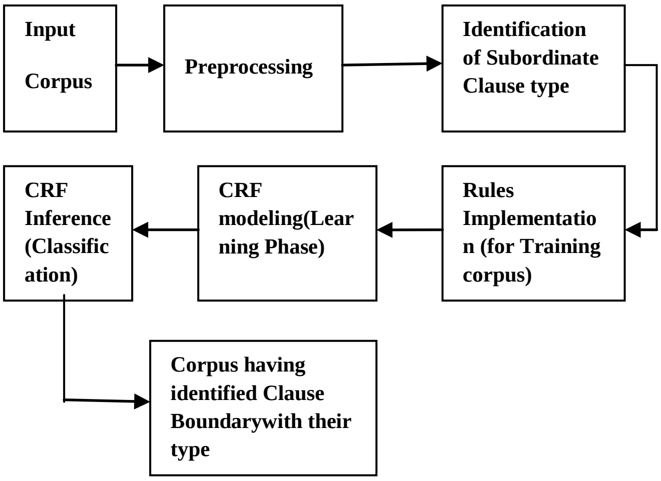
Input
Corpus
Preprocessing
Identification of Subordinate Clause type
Rules Implementation (for Training corpus)
CRF modeling(Learning Phase)
CRF Inference (Classification)
Corpus having identified Clause Boundarywith their type

Fig 1: Workflow for clause identification

Two phases for clause boundary identification has been used. Inference phase where modeling takes place from given input data set and Inference phase where classification of data set takes place. Different types of clause markers are used which help in making relations between the sentences and combining two Urdu sentences. In Urdu, linguistic rules were used for identification of beginning and ending of clauses. If there are any errors then corrections are done through these rules. Data set which is used as input encompasses different types of sub-ordinate clauses. Results are acquired using clause markers and are 89.2% precision and 90.0% recall.

2.4 Malayalam

Malayalam is one of the four Dravidian languages which is mainly spoken in Kerala and is derived from Ancient Tamil. In this language verbs are present at the end of the sentence and Urdu is a ‘Left Branching Language’. Here CRF is used for clause identification [12]. Three types of clauses are identified in this paper i.e. Relative Participle clause, Conditional clause and Main clause. Two types of features are used. First is word level feature in which three things are considered i.e. lexical word, its part-of-speech and chunk. Second feature used is structural level which is grammatical rules. Grammatical rules of Malayalam language have been considered as major feature. The first step of basic method used was preprocessing. Data presented here was in column format. Window of size five have taken and first column represents words, second column shows the pos tags, in third column chunk information was given, Boolean entries forms the fourth column and at last fifth column was of clause boundary information. In this work 3638 sentences was trained and 801 sentences was used for testing. According to results calculated conditional tags were more accurate and approximate accuracy of conditional closing tags was 98.53%.

2.5 Portuguese

Portuguese belong to the Indo-Aryan family of languages and from a group of languages called Romance. It is the sixth most popular language spoken in the world. A supervised machine learning approach was used for clause identification in Portuguese. This work was based on [3] Entropy Guided Transformation learning. POS and PCL tags were used as input features. Corpus for training was taken from Bosque corpus. Clauses were identifiedin three steps i.e. clause start identification, clause end identification and complete clause identification. Authors proposed a simple heuristic to derive a phrase-chunk like feature from phrases in the Bosque corpus. Three types of phase chunks were considered i.e. verbal, nominal and prepositional. IOB2 tagging style was used to codify these features. ETL algorithm was used. The three steps used for clause identification was performed sequentially so that information of previous step act as input for next step. They adopt simple baseline system proposed in CONLL’2001. The first two steps identify the tokens and last step splits the given sentence into clauses. Two modeling approaches were taken for the last sub task and are ETL-Token and ETL-Pair. For start, end and ETL-Token window size of 7 was set and for ETL-Pair window size of 9 was set. Three versions of system have been taken to evaluate the potential and impact of PCL feature. Result differs according to each version. According to first version i.e. using no information of PCL feature the accuracy was 66.95%. The accuracy with respect to second version i.e., using automatic values of PCL feature was 69.31% and according to third version i.e. using golden values of PCL feature accuracy was 73.90%.

2.6 English

It is the most important languages in the world. It is a member of the Indo-European family of languages. A Rule based technique was proposed [1] to simplify the complex sentences based on relative pronouns, coordinating and subordinating conjunctions. Based on rules, the sentences were simplified. Most of the sentences contain conjunction and sentences were split based on conjunctions. The proposed algorithm follows following steps: First, split the sentences from the paragraph based on delimiters such as "." And "?". Second, delimiters such as comma are ignored from the sentences. Third, individual sentences are split based on coordination and subordinating conjunctions. Presence of delimiters was an important pre-requisite as the initial splitting was done based on them.

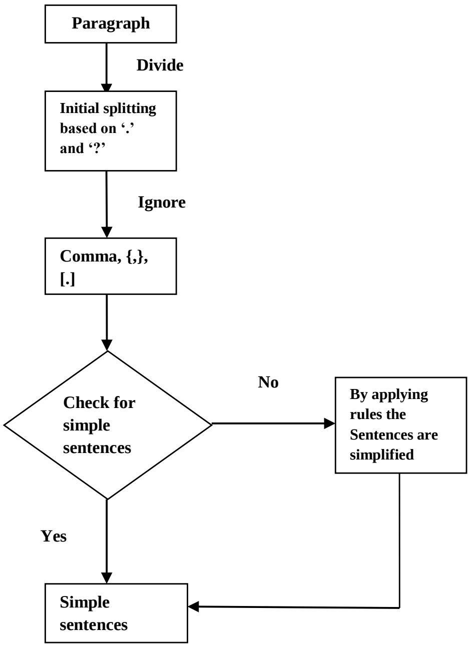
Paragraph
Divide
Initial splitting based on '.' and '?'
Ignore
Comma, {,}, [.]
Check for simple sentences
No
By applying rules the Sentences are simplified
Yes
Simple sentences

Fig 2: Framework of authors work

Two techniques were used splitting and simplification. Splitting was used to break the sentences containing coordinating and subordinating conjunction whereas simplification was used to simplify the sentences which contain relative pronouns. Sentences are simplified using above algorithm. Out of 200 sentences which was tested 115 sentences was translated correctly after simplification.

2.7 Chinese

It is the most widely spoken language in the world. It is the fifth official language of United Nations and also spoken in Singapore. Chinese language belongs to the Sino-Tibetan language system. Chinese languages are tonal i.e. their meaning varies by the way words are pronounced. SVM model was applied to Chinese clause splitting [15]. Clause splitting is important task for various tasks such as machine translation, sentence simplification, syntactic parsing etc. Chinese clause splitting aims at recognizing the high-level clauses. The input given was complete Chinese sentences with proper tagging. The algorithm identifies the left and right boundaries of every clause to form sequence of clauses. In Chinese clause splitting chunk tags are not provided due to which Chinese clause splitting is a very difficult task. Chinese clauses are often split at the locations of punctuators so we formulate the clause splitting task as a sequence labeling problem. Authors employ SVM that see tag sequences called an, SVM Hidden Markov Model.In SVM-HMM model, the features should be divided into two types: emission features and transition features. But while applying SVM-HMM in Chinese clause splitting suitable feature vector and loss function was considered. Authors used lexical and pos information within a fixed window based on blocks. The different features used are: word features, pos features, punctuator features, combinatorial features, the features of inclusion of verb and the number of words appearing in the block. After words are separated they are examined whether the block forms an independent clause to give the right boundary of Chinese clause, left boundary is implicitly decided. Then post processing was done based on various rules. Qinghua Treebank was used as data for training and testing. To compare the effectiveness of their approach authors developed two baseline systems. Experimental results showed that SVM-HMM approach used in this paper achieved much better results. The results are 76.36% precision and 80.02% recall for 3th order model.

2.8 Romanian

Romania is a modern romance language like French, Portuguese, and Spanishetc. This is the primarily spoken language of Romania and Moldova. A multilingual method was used for detecting clause boundaries in Romanian language. Paper [11] presented a clause splitting method which combines the benefits provided by machine learning with a small set of human-written rules. Machine learning was used for categorizing the coordinating conjunctions and punctuation marks as clause boundaries or not. The rules were used for identifying the finite verbs which was the central part of each clause and also for finding other clause boundaries not included in the learning process. A classifier is used to decide which coordinating conjunction and punctuation marks were clause boundaries. The resultant clause boundaries were passed onto the rule-based module for performing consistency checks. Authors divide the clause boundaries to be identified into different types. Firstly, Coordinating conjunctions and punctuation marks because according to statistics author show that approximately 75% of the English clauses either starts with a coordinating conjunction, end with a punctuation mark, or both. The machine learning algorithm used was memory-based learning. The memory-based learning method implemented by the author was k-nearest neighbors. Second type was Subordinators. Third were other clause boundaries. In this type the rule-based module will take pair of two from ordered list of previously detected clause boundaries and check how many predicates lie in-between. If more than two predicates are found then a new clause boundary was searched. The corpus consists of 118,360 words and 16.012 clauses were annotated. The results were 89.03% precision and 88.51% recall at the level of complete clauses retrieved and 95.59% precision and 95.03% recall at detecting only the start position of each clause.

2.9 Dutch

Dutch is a member of the western group of the Germanic languages and is primarily spoken in Netherlands as well as major parts of both Belgium and Surinam. This language is closely related to English and German. In this work authors [2] compared two approaches to sentence simplification for TV programme transcripts and the related subtitles. The first was machine learning approach and second was knowledge based approach relies on hand-crafted phrase deletion rules. For machine learning, they had represented the summarization process as a word transformation task. Words can be copied, deleted or replaced. A corpus of 12,535 sentences or 108,015 words was taken. 90% of data was kept as training material for machine learner and 10% data was used as test material. A memory-based learner was applied to the training data which was fed with words, lemmas, part-of-speech tags, chunk tags, relation tags and proper name tags. A bidirectional hill-climbing method was used foe feature selection for determining optimal set of features. Namefeatures and relation features were not used by the selection process. The machine learner obtains a 92.3% compression rate but the target was 81.3% due to which this approach did not perform well so to get better results authors manually compile phrase deletion rules. Their goal was to perform phrase deletion in two steps: first selecting those phrases that are more or less redundant and second choosing some of the phrases for deletion so that the required compression rate was met. The phrase deletion rules were applied. After all eligible rules had been applied, a final check was done. Two different selection methods were chosen for phrase deletion i.e. by length and sentence position. The deletion process was continued until the required compression rate was obtained. The deletion rules obtain the compression rate of 74.3% which was required.

2.10 Estonian

This language belongs to the group of Finno-Ugric languages. This is the only spoken language of Estonia and is widely used by the locals. This paper [5] describes the rule-based system for tagging clause boundaries. This system had taken morphologically annotated text as input. The system also identifies parenthesis and embedded clauses i.e. clauses that are inserted into another clauses. Corpus of the University of Tartu was taken which is a collection of written texts containing 245 million words. Finite clauses and subclass of finite clauses were targeted. Type of clause was not being identified by the system. The algorithm used was processed in various steps. A sentence was traversed several times. An algorithm proceeds as follows. First, text was placed in brackets which were fail-safe indicators. Second, the verb forms that might be suitable for acting as main verbs in clauses were tagged. Third, conjunctions and punctuation marks were marked as potential clause boundaries. Fourth, the start and ending of direct speech were tagged. Fifth, the data between the quotation marks was tagged. Sixth, colon and semicolon were tagged. At last remaining clause boundaries was considered and if they had possible verbal centre on both sides, tag the boundary. Then these clause boundaries were classified and then relative clauses were marked ad embedded clauses.A 16,000 word test corpus, consisting data from fiction, newspaper and popular science texts was used for the evaluation. So the result combining all three data sets was 95% recall and 96% precision.

3. CONCLUSION

At last we conclude that clause boundary identification is the most important activity of many Natural Language based applications.Accuracy of computing complex sentences is dependent of correctly identified clauses.Different approaches have been used by authors for clause boundary identification for different languages. In case of Hindi, Hybrid approach consists of different models using CRF machines and rule based approach. The accuracy achieved is 91.53% and 80.63%. For Bengali language two approaches have been used and with clause identification clause type was also identified with accuracy of 73.12% and 78.07%. In Urdu there were also two techniques which have been proposed and are rule based and machine learning. The accuracy of was achieved 89.2% and 90.0%. For Malayalam CRF was used with accuracy of 98.53%. In case of Portuguese a supervised machine learning approach have been used. Results differ according to versions used and are 66.95%, 69.31%,and 73.90%. For English Rule based approach has been used in which out of 200 sentences 115 sentences were translated correctly after simplification. For Chinese SVM model has been used and results are 76.36% precision and 80.02% recall. In case of Romanian multilingual method was used. Accuracy achieved is 89.03% precision and 88.51% recall. For Dutch machine learning and rule based approach have been used with 74.3 % accuracy. In case of Estonian a rule based approach has been used with 95% recall and 96% precision accuracy. From this study it is evaluated that many different approaches have been used for clause identification in different languages with proper results.

4. REFERENCES

[1] C Poornima, V Dhanalakshmi, M Kumar Anand, P K Sonam, "Rule based Sentence Simplification for English to Tamil Machine Translation System." International Journal of Computer Applications (0975 – 8887), Volume 25, No.8, July 2011.

[2] Daelemans Walter, HothkerAnja, Sang Erik Tjong Kim, "Automatic Sentence Simplification for Subtitling in Dutch and English." In: Proceedings of the 4th Conference on Language Resources and Evaluation, Lisbon, Portugal (2004) 1045-1048.

[3] FernandesEraldo R, Santos Cicero N. dos, MilidiuRuy L, "A Machine Learning Approach to Portuguese Clause Identification." In: Proceedings of the Computational Processing of the Portuguese Language. pp. 55–64 (2010).

[4] Ghosh Aniruddha, Das Amitava, BandyopadhyaSivaji, "Clause Identification and Classification in Bengali." Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 17–25, the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010.

[5] KaalepHeiki-Jaan, MuischnekKadri, "Robust clause boundary identification for corpus annotation." In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey (2012).

[6] Kumar Dinesh, JosanGurpreet Singh, " Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey", International Journal of Computer Applications (0975 – 8887), Volume 6– No.5, September 2010.

[7] LeffaVilson J, "Clause Processing in Complex Sentences." First International Conference on Language Resources And Evaluation, Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Espanha: 1998. v. 2, p. 937-943.

[8] Nadkarni Prakash M, Machado LucilaOhno, Chapman Wendy W," Natural language processing: an introduction", J Am Med Inform Assoc 2011; 18:544e551. Doi: 10.1136/amiajnl-2011-000464, Published by group.bmj.com on October 5, 2011.

[9]Orasan, C.: A Hybrid Method for Clause Splitting in Unrestricted English Text. In: Proceedings of ACIDCA 2000 Corpora Processing, Monastir Tunisia, pp. 129–134 (2000).

[10] ParveenDaraksha, SanyalRatna, Ansari Afreen, "Clause Boundary Identification using Classifier and Clause Markers in Urdu Language." Polibits Research Journal on Computer Science, 43, pp. 61-65, 2011.

[11] Puscasu Georgiana, "A Multilingual Method for Clause Splitting." In: Proceedings of the 7th annual colloquium for the UK Special interest group for computational linguistics (CLUK 2004), Birmingham, UK.

[12] S Lakshmi, Ram Sundar Vijay, R and Sobha, Devi Lalitha, "Clause Boundary Identification for Malayalam Using CRF." Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 83–92, COLING 2012, Mumbai, December 2012.

[13] Sharma Rahul, Paul Soma, "A hybrid approach for automatic clause boundary identification in Hindi." Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 43–49, Dublin, Ireland, August 23-29 2014.

[14] Siddharthan, A.: An architecture for a text simplification system. In: LREC 2002: Proceedings of the Language Engineering Conference, LEC 2002, pp. 64–71 (2002)

[15] Yin Dapeng, Jiang Peilin, Ren Fuji, Kuroiwa Shingo, "Chinese complex long sentences processing method for Chinese-Japanese machine translation." IEEE, 2007.

[16] Zhou Junsheng, Zhang Yabing, Dai Xinyu, Chen Jiajun, "Chinese Event Descriptive Clause Splitting with Structured SVMs." Springer-Verlag Berlin Heidelberg, Pp. 175-183, 2010.

[17] http://tdil.mit.gov.in/

[18] http://punjabi.aglsoft.com/punjabi/learngrammar/

[19]http://en.wikipedia.org/wiki/Clause