IJCTA, 9(10), 2016, pp. 4725-4731
© International Science Press

Semantic Role Labeling for Malayalam

***Jisha P Jayan **J Satheesh Kumar**

Abstract : Semantics, a field of Natural Language Processing (NLP), is concerned with the extraction of meaning from a sentence. The Semantic Role Labeling (SRL) takes the preliminary steps in extracting meaning from a sentence by giving generic labels or roles to the tokens of the text. This paper studies the important task of Semantic Role Labeling. This work proposes SRL approach for Malayalam language. Here Semantic roles are identified using Karaka theory based on the Paninian Grammar, useful for both the syntactic and semantic analysis. To identify these relations between nouns and verbs in a sentence, case markers plays an important role. The results obtained are very much promising and further enhancements to this can be made in future.

Keywords : Semantic role labeling, kaarakas, case roles, parsing.

1. INTRODUCTION

Semantic parsing of sentences is admitted to be an important task on the road to natural language understanding, and has immediate applications in some of the tasks such as information extraction and question answering. SRL is a shallow semantic parsing task, where for each predicate in a sentence, the aim is to determine all the constituents that fill a semantic role, and to determine their roles and their adjuncts. Semantic Role Labeling is the task of assignment of the semantic roles to the constituents of the sentence. In linguistic theory, these roles are one of the oldest classes of constructs. The Paninian karaka theory is considered as one of the oldest works in this field [1]. Many varieties in semantic roles exist today. The roles of semantic could be mostly domain-specific or generic in nature. Fillmore [2] gives a hierarchical classification of semantic roles. Present approaches towards semantic role labeling are based on supervised machine learning, normally using the FrameNet and PropBank resources to define the specific set of roles used in the task, and providing the training and test sets [3].

The Semantic Role Labeling is identified as a crucial problem in almost all the NLP applications where some kind of semantic interpretation is required. The CoNLL 2004 and CoNLL 2005 Shared Tasks defined the task of Semantic Role Labeling as "analyzing the propositions expressed by some target verbs of the sentence. In particular, for each and every target verb in the sentence, all the constituents which fill a semantic role of the verb have to be recognized". SRL finds its importance in many NLP applications like Question Answering, Information Extraction, Text Summerization, Machine Translation etc.

This paper presents the work in different stages, beginning in Section 2 with the major works related to SRL. Section 3 introduces the semantic role labeling for Malayalam. Section 4 describes the architecture and the results obtained from this study. Finally, Section 5 concludes the paper.

2. STATE OF ART OF SRL

Shallow semantic parsing is an approach by Gildea and Jurafsky [4] in which the semantic relationships or roles identification with the help of constituents of a sentence within the semantic frame are described. They proposed a statistical approach for identifying the semantic roles of pre-segmented constituents and claim an accuracy of 82%. They used the sentences annotated with the semantic roles of FrameNet project [2].

* Research Scholar Bharathiar University jishapjayan@gmail.com

** Assistant Professor Department of Computer Applications Bharathiar University jsathee@rediffmail.com

The challenge for CoNLL-2004 shared task of Xavier Carreras [5] was to think of machine learning techniques, which address the SRL problem on the premise of only partial syntactic information, avoiding the usage of full parsers and external lexico-semantic knowledge bases. The annotations provided for the development of systems incorporate, besides from the argument boundaries and role labels, the levels of processing treated in the past editions of the CoNLL shared task, *i.e*., words, parts-of-speech tags, base chunks, clauses and named entities.

Most frameworks for automatic SRL practices a full syntactic parse in order to define argument boundaries and to extract suitable information for training classifiers to disambiguate between role labels of the sentence. Thereby the task has been usually approached as a two stage procedure that consists of recognition and labeling of the arguments. The CoNLL-2004 shared task mentioned various models for the learning component including pure probabilistic models [1], [6] and [7], Maximum Entropy models [8], generative models [9], Support Vector Machines (SVM) [10], Decision Trees [11] and [12]. Nianwen [13] defined the method of calibrating features. The Co-NLL 2005 conference on Semantic Role Labeling provided a vision on several statistical methods followed, prominently the Maximum Entropy classifier of Wanxiang [14]. The finest results were obtained by systems using Support Vector Machines [15].

Many researches have focused on describing the various expressions of verb arguments within the syntactic positions [16]. The conclusion of this work states that the pattern of syntactic alternation reflects the semantic similarity among the verbs, forming the basis for verb classes. Such verb classes have proven useful in a number of NLP tasks [17] [18], including SRL [19] and have provided the foundation for VerbNet , a computational verb lexicon [20].

In Shallow Semantic Parsing by [21] first they placed their algorithm based on the statistical classification with one that uses Support Vector Machines and then added to the existing feature set. Then they evaluated the results using both the hand-corrected syntactic parses using TreeBank, and actual parses obtained from the Charniak parser.

In Semantic role labeling using syntactic chunk by Kandri Hacogolu [22] presented a semantic role labeler (or chunker) that clubs the syntactic chunks (*i.e*. base phrases) into the arguments of a predicate. This is done by casting the semantic role labeling as the classification of syntactic chunks such as NP-chunk, PP-chunk into one of the several classes such as the beginning of an argument, inside an argument and outside an argument. This accounts the tagging of syntactic chunks with semantic labels using the IOB representation.

The use of dependency parsers for semantic role labeling was introduced in the CoNLL-2008 shared task [23]. Recent works on this focuses on deep neural networks [24]. Semantic roles can act as a major midway representation in various statistical MT systems or in areas of automatic text summarization and in the emerging field of text data mining applications [25]. Finally, incorporation of the semantic roles into probabilistic models of language may gradually yield accurate parsers and better language models for several speech recognition systems.

3. SRL FOR MALAYALAM

The semantic roles range from the very particular to the very general, and numerous have been utilized as a part of computational implementations of one type or another. The study that relates with the different roles connected with each verb and across various verb classes is called thematic role analysis or case role (karaka) analysis. In Malayalam, the root or the stem word by itself gets inflected to change its meaning, or to combine the word with other words. Malayalam language follows a relatively free word order.

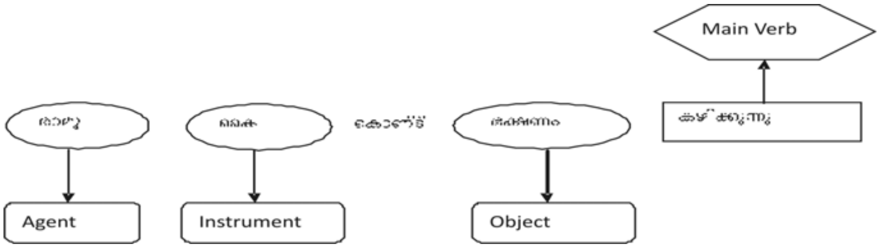

Fig. 1. SRL for a Malayalam sentence.

The feature of inflection and agglutination nature of Malayalam makes computer-based language processing a challenging task. Amid the semantic analysis, verb is taken as the central, element of the sentence. According to Paninian viewpoint, there are four levels in understanding any sentence [26] namely the surface level (uttered sentence), the vibhakthi level, the karaka level and the semantic level. The karaka level has relationship to semantics at one side with the syntax on the other side. Karaka relation can be identified from markers or case endings after noun. These markers and case endings are called vibhakthi. In Malayalam, karaka relations are analyzed from vibhakthi point of view and the postpositions that associate with it. Consider the following example:

രാമു കൈ കൊണ്ട് ഭക്ഷണം കഴിക്കുന്നു.

raamu kai koNT bhakshaNaM kazhikkunnu

ramu hand with food eat

Ramu eats food with hand.

Figure 1 depicts the diagrammatic representation for SRL for the above example. Kaarakas gives the vital information with respect to a verb by giving the relations that exists between the nouns and the verbal stem in a sentence. It specifies a relation between a verb, which stands for an action, and nominals in the sentence. Consequently, the verb decides the kaaraka of nominal words that are used in a sentence. Verbs are related to the nominal words in various ways based on which the kaaraka varies. Hence, for any verb, distinctive kaarakas may result. In view of the semantic relation between the nouns and verbal stem, the Kaaraka relations are decided. Therefore, the Kaaraka relation caters the syntactic-semantic relationship that exists between the different words of the sentence. The karakas are as follows:

Table 1. Kaarakas with Cases.

Karakas	***Role***	***Case***
Karthru Karakam (ക ത്തൃകാരകം)	Subject/Agent (കർത്താവ്)	Nominative (നിർദ്ദേശിക) Instrumental (പ്രയോജിക)
Karma Karakam (ക മ്മകാരകം)	Object (കർമ്മം)	Accusative (പ്രതിഗ്രാഹിക)/ Nominative (നിർദ്ദേശിക)
Karna Karakam (കരണകാരകം)	Instrument (കർണം)	Instrumental (പ്രയോജിക)
Kaarana Karakam (കാരണകാരകം)	Instrument (കാർണം)	Instrumental (പ്രയോജിക)
Sakshi Karakam (സാക്ഷികാരകം)	Experiencer (സാക്ഷി)	Sociative (സംയോജിക)
Swami Karakam (സ്വാമികാരകം)	Beneficiary (സ്വാമി)	Dative (ഉദ്ദേശിക)
Adhikarana Karakam (അധികരണകാരകം)	Locative (അധികരണം)	Locative (ആരാധിക)

4. IMPLEMENTATION AND RESULT

The basic architecture for the proposed semantic role labeling for Malayalam is shown in the Figure 2.

The architecture consists of different stages like tokenization, morphological analyzer, parts-of-speech tagging, chunking, case analysis and finally semantic role identifier. The roles taken for this study are Agent, Patient, Instrument, Beneficiary, Experiencer, Causer, and Recipient.

Agent - performs an action.

റാം തന്റെ സൂപ്പ് പതിയെ കഴിച്ചു.

raam tan~Re suupp patiye kazhiccu.

Ram ate his soup slowly.

Experiencer - one that receives an emotional affect.

ഞാന് കരഞ്ഞു.

njaan~ karanjnju.

I cried.

Causer - cause of an action.

രാമു വിശപ്പ് കാരണം കേക് കഴിച്ചു.

raamu viSapp kaaraNaM keek kazhiccu

Ram ate cake due to hunger.

Patient - action that changes its state.

രാമന് രഘുവിനെ അടിച്ചു.

raaman~ raghuvine aTiccu.

Ram beat Raghu.

Instrument - tool that cause an action.

ആരോ കത്തി കൊണ്ട് ബ്രെഡ് മുറിച്ചു.

aaroo katti koNT breD muRiccu.

Somebody cut bread with a knife.

Beneficiary - benefit of the action that takes place.

അവന് ഒരു കാര് എനിക്ക് വേണ്ടി ഉണ്ടാക്കി .

avan~ oru kaar~ enikk veeNTi uNTaakki

He made a car for me.

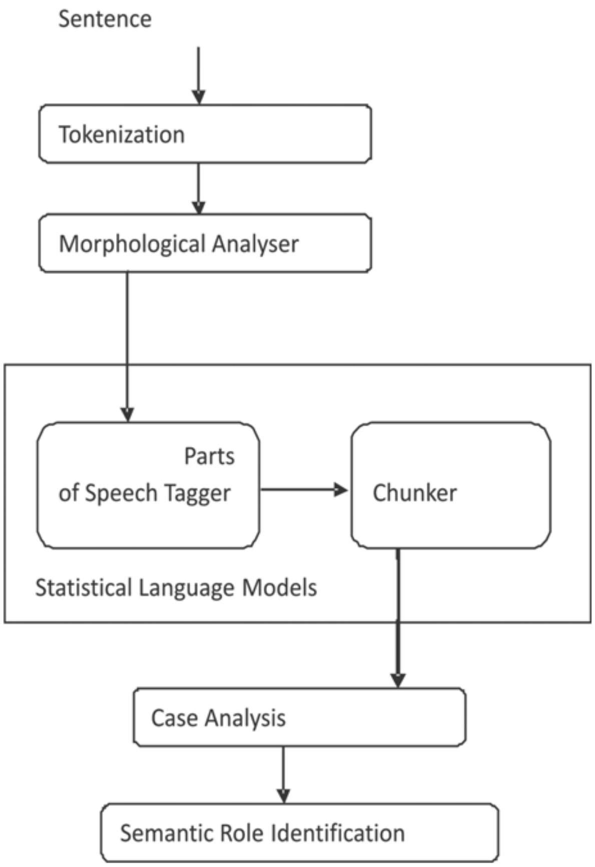

Fig. 2. Architecture of SRL for Malayalam

Recipient - one whor receives the goal of an action.

അമ്മ കുട്ടിക്ക് കുട കൊടുത്തു.

amma kuTTikk kuTa koTuttu.

Mother gave an umberlla to child.

The tokenizer does the breaking up of sentences into different tokens. The morphological analyzer involves the identification of root/stem along with the associated grammatical features of the given words. The paradigm based morphological analyser is used [27]. For nouns and pronouns, the morphological analyzer gives person, number and gender information and for verb, it gives tense, aspect and modality.

The part-of-speech tagging assigns the best part of speech/category to the tokens while the chunking involves the identification of the boundary of the chunks and the labels. The tagset used for the part-of-speech tagging and chunking are developed by IIIT-H [28]. Malayalam corpus was manually tagged for both pos. The tagger and chunker are trained using a statistical model based on CRF with an annotated corpus consisting of about three lakh tokens each. Case markers are identified and the semantic roles are labelled accordingly.

Input 1 : അദ്ധ്യാപിക കുട്ടിയെ വടിയാ അടിച്ചു.

addhyaapika kuTTiye vaTTiyaal~ aTiccu.

Output : അദ്ധ്യാപിക(Agent) കുട്ടിയെ(Patient) വടിയാ (Instrument) അടിച്ചു .

addhyaapika(Agent) kuTTiye(Patient) vaTiyaal~(Instrument) aTiccu.

Input 2: അച്ഛ എനിക്ക് ഡെ ഹിയി നിന്ന് ഒരു സാരി വാങ്ങി.

acchan~ enikk Del~hiyil~ ninn oru saari vaangngi.

Output :

അച്ഛ (Agent) എനിക്ക് (Benificiary) ഡെ ഹിയി (Location) നിന്ന് ഒരു സാരി(Patient) വാങ്ങി .

acchan~(Agent) enikk(Benificiary) Del~hiyil~(Location) ninn oru saari(Patient) vaangngi .

Input 3 : ഞാ നിപ്പുവിനെ കൊണ്ട് പുസ്തകം വാങ്ങിപ്പിച്ചു.

njaan~ nippuvine koNT pustakaM vaangngippiccu.

Output :

ഞാ (Agent) നിപ്പുവിനെ (Experiencer) കൊണ്ട് പുസ്തകം (Patient) വാങ്ങിപ്പിച്ചു.

njaan~(Agent) nippuvine (Experiencer) koNT pustakaM (Patient) vaangngippiccu.

The morphological analyser plays an important role in determining the cases. The present study identifies semantic roles with an accuracy of about 70%. The ambiguous words in the sentences affect the performance. The present system can be further improved with the addition of more semantic roles. Usage of semantically tagged corpora with machine learning approaches would help in resolving ambiguity.

5. CONCLUSION

SRL presents an excellent framework to perform research on semantic relations among the different components of a text. It is a shallow level representation of semantics and has wide. With advances in both rule based and statistical techniques Semantic Role Labeling is a rapidly developing area in NLP. The morphological endings or the cases markers of words play an import role in identifying the semantic roles. Noam Chomsky's Selectional Restriction approach can also be applied on this study to enhance the result. The present study gives an accuracy of 70%. With the help of semantically annotated corpora, Machine Learning methods can be introduced as a further improvement to the present study.

6. REFERENCES

1. D. Gildea and D. Jurafsky, “Automatic Labeling of Semantic Roles”, Computational linguistics, pp 245-28, 2002.
2. C F Baker, C J Fillmore, and J B Lowe, “The Berkeley FrameNet Project”, In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pages 86-90. Association for Computational Linguistics, 1998.
3. JH Martin, and D Jurafsky, “Semantic Role Labeling, Speech and Language Processing”, International Edition, 2000.
4. D. Gildea and D. Jurafsky “Automatic labeling of semantic roles”, In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pp 512-520, October 2000.
5. X Carreras, and L Marquez, Introduction to the CoNLL-2005, Shared Task: Semantic Role Labeling, Proceedings of CoNLL, 2005.
6. D Gildea, and M Palmer, “The Necessity of Syntactic Parsing for Predicate Argument Recognition”, In Proceedings of ACL 2002, Philadelphia, USA, pp. 239-246, 2002.
7. D Gildea and J Hockenmaier, “Identifying Semantic Roles Using Combinatory Categorical Grammar”, In Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, Philadelphia, USA, pp. 57-64, 2002.
8. M Fleischman, N Kwon, and E Hovy, “Maximum Entropy Models for FrameNet Classification”, In Proceedings of the 2003 conference on Empirical methods in natural language processing Association for Computational Linguistics , pp. 49-56, July 2003.
9. C. A. Thompson, R. Levy, and C. Manning, “A Generative Model for Semantic Role Labeling”, In European Conference on Machine Learning, Springer Berlin Heidelberg. Dubrovnik, Croatia, pp. 397-408, September 2003.
10. K Hacioglu and W Ward, “Target word detection and semantic role chunking using support vector machines”, In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2, Association for Computational Linguistics, pp. 25-27, May 2003.
11. M Surdeanu, S Harabagiu, J Williams, and P Aarseth, “Using predicate-argument structures for information extraction”, In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, Japan, pp. 8-15, July 2003.
12. S Pradhan, K Hacioglu, W Ward, J H.Martin, and D Jurafsky, “Semantic role parsing: Adding semantic structure to unstructured text”, Proceedings of the International Conference on Data Mining (ICDM-2003), Melbourne, USA, pp. 629-632, November 2003.
13. N Xue and M Palmer, “Calibrating Features for Semantic Role Labeling”, EM-NLP, pp. 88-94, 2004.
14. T Liu, W Che, S Li, Y Hu and H Liu, “Semantic role lableing system using maximum entropy classifier”, In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, pp. 189-192, June 2005.
15. S L Pandian, and T. V. Geetha. “Semantic role labeling for Tamil documents.”, International Journal of Recent Trends in Engineering, Volume 1, No 1, pp 483-489, May 2009.
16. B Levin, “English Verb Classes and Alternations: A Preliminary Investigation”, The University of Chicago Press, Chicago, IL. 1993.
17. N Habash, B J Dorr, and D. Traum, “Hybrid natural language generation from lexical conceptual structures”, Machine Translation, Volume 18, No 2, pp 81-128, June 2003.
18. L Shi, and R Mihalcea, “Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing”, In Computational Linguistics and Intelligent Text Processing; Sixth International Conference, CICLing 2005, Proceedings, LNCS, vol 3406, Mexico City, Mexico, pp. 100-111, February 2005.
19. R Swier, and S Stevenson, “Unsupervised semantic role labelling”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, Volume 95, pp 102, 2004.
20. K Kipper, H T Dang, and M Palmer, “Class-based construction of a verb lexicon”, In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, TX., pp. 691-696, July 2000.

21. K Hacioglu, S Pradhan, W Ward, J Martin, and D Jurafsky, "Shallow Semantic Parsing Using Support Vector Machines", Technical Report TR-CSLR-2003-1, Center for Spoken Language Research, Boulder, Colorado, 2003.

22. K H and W Ward, "Target word detection and semantic role chunking using support vector machines", In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003, Association for Computational Linguistics.short papers-Volume 2 pp. 25-27, May 2003.

23. M Surdeanu, R Johansson, A Meyers, A., L Marquez, and J Nivre, "The conll-2008 shared task on joint parsing of syntactic and semantic dependencies", In Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL-08), Association for Computational Linguistics. pp. 159-177, August 2008.

24. W R Foland Jr, and J H Martin, "Dependency based semantic role labeling using convolutional neural networks", In Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics, pp. 279-288, August 2015.

25. M A Hearst, "Untangling text data mining", Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, College Park, Maryland pp3-10, June 1999.

26. A Bharati, V Chaithanya, R Sangal, "Natural Language Processing: A Paninian Perspective", Prentice-Hall of India, New Delhi, 1995.

27. J P Jayan, R R Rajn and S Rajendran, "Morphological Analyser for Malayalam - A Comparison of Different Approaches", International Journal of Computer Science and Information Technology (IJCSIT),Vol. 2 No. 2, pp 155-160, December 2009.

28. R Sangal, A Bharati, D M Sharma and L Bai, "Guidelines for POS and Chunk Annotation for Indian Languages", 2006.