Semantic Parsing Approach in Malayalam for Machine Translation

Shabina Bhaskar
MPhil CS
IIITMK-Trivandrum

Sandeep Chandran
Assistant Professor
LBS Center for Science and Technology
Trivandrum

Abstract **- Semantic parsing is the process of mapping a natural language sentence into a formal representation of its meaning. Shallow form of Semantic Parsing is proposed in this paper. Here Semantic roles are identified using Karaka theory in Paninian Grammar. It is useful for both the syntax analysis and semantic analysis of Malayalam sentences. And we proposed an algorithm for finding semantic relations using Karaka theory**.

I. INTRODUCTION

Semantic parsing maps text to formal meaning representations. There are two approaches in semantic parsing deep parsing and shallow parsing. Deeper semantic analysis means representation of a sentence in predicate logic or other formal representation of meaning.

Shallow parsing means case role identification. The study of roles associated with specific verbs and across classes of verbs is called thematic role analysis or case role (karaka) analysis. The proposed approach is shallow parsing approach in Malayalam. We can consider a sentence compose of entities and interactions between entities. Here nouns and verbs take as entities and relation (karakas) between verbs and nouns as interactions. Here we identify abstract semantic roles or thematic relations such as Agent, Patient etc.

According to Paninian perspective there are four levels in the understanding process of a sentence [4]. They are surface level (uttered sentence), vibakthi level, karaka level and semantic level. The karaka level has relationship to semantics on one side and to syntax on the other side. Karaka relation can be identified from post position markers after noun or surface case ending of noun. These markers and case endings are called vibhakthi. In Malayalam karaka relations are analyzed from vibakthi and post position markers. An algorithm is implemented for finding case roles of nouns and machine learning approach is used for finding the roles of adjectives, adverbs and postposition markers. So the system proposes a hybrid approach for Semantic Parsing.

II. RELATED WORKS

Shallow semantic parsing is the process of assigning a simple WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW, etc. and this semantic parsing approach described by Gildea and Jurafsky[1].

In Shallow Semantic Parsing using Support Vector Machines [2] first they placed their statistical classification algorithm with one that uses Support Vector Machines and then added to the existing feature set. Then evaluate results using both hand- corrected TreeBank syntactic parses, and actual parses from the Charniak parser.

In Semantic role labeling using syntactic chunk by Kandri Hacogolu presented a semantic role labeler (or chunker) that groups syntactic chunks (i.e. base phrases) into the arguments of a predicate [3]. This is accomplished by casting the semantic labeling as the classification of syntactic chunks (e.g. NP-chunk, PP-chunk) into one of several classes such as the beginning of an argument (B-ARG), inside an argument (I-ARG) and outside an argument (O). This amounts to tagging syntactic chunks with semantic labels using the IOB representation. The chunker is realized using support vector machines as one-versus-all classifiers.

III. METHODOLOGY

The proposed system architecture contains two phases syntactic parsing and semantic parsing. In syntactic parsing the steps involved are tokenization, POS tagging and morphological analysis .In Semantic Parsing are Semantic role labeling and conceptual graph representation.

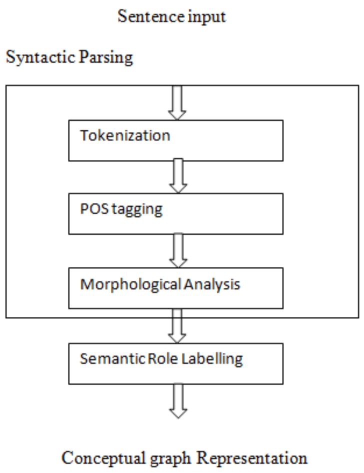

Fig.1: System Architecture

A. Syntactic parsing

Syntactic parsing start with tokenization. As Malayalam is a highly inflected and agglutinative language, tokenization cannot be done blindly. We used a supervised machine learning approach which classifies the tokens into two groups: the words to be split (compound word) and not to be split (Simple word).Compound word is splitted using sandhi rules. After tokenization we perform parts of speech tagging. For tagging we used a supervised machine learning approach which identifies the syntactic structure of the sentence. Next phase is Morphological analysis and it is the process of recognizing the root or stem and the categorical information of the items that may accompany the root or the stem. In the case of Noun, the root word and it categorical information such as person, number gender information and in the case of verb, tense, aspect and modality are included. And in this system we add an additional field which specifies the Semantic Property associated with nouns and verbs. And it is stored as

- Noun {root, {suffix, vibakthi}, gender, number, person, semantic property}
- Verb {root, {suffix, verb forms},semantic property}
- For example The noun "രാധ " is stored as

  (radha,O:nirdeshika എ):prathigrahika, ഓട്:samyojika,ക്ക്:udeshikaആൽ:prayogika, ഇൽ:adharika,ഉടെ:sambandhika,female,singular,third, {human})

Semantic property of verbs such as Artistic Performance Verbs (പാടുക ,വരയ്ക്കുക) Combine verbs(കെട്ടുക ,കൂട്ടിചേര്ക്കുക) and Nouns such as Animals(പശു, പൂച്ച) Birds(കോഴി ,മയിൽ) ,Buildings(കോളേജ് ,വീട്) are used in this work.

B. Semantic Parsing

In Semantic Parsing we introduce Shallow Semantic Parsing approach in Malayalam. Here Semantic Roles of noun, adjectives, adverbs and postposition markers are considered. Semantic Roles of Nouns are identified using the Karaka Theory which has the basis in Paninian Grammar. Karaka theory specifies the relation between noun and verb. And the roles of adjectives, adverbs and postposition markers identified using a machine learning approach.

C. Semantic Role Labeling

Malayalam follows the system of marking grammatical relations and semantic roles through a set of case suffixes. According to Karaka Theory the Karaka Relations (Semantic relations) can be identified from vibakthi. In semantic parsing for semantic role (karaka) labeling is based on vibakthi prathyam. From morphological analysis the case markers of nouns and pronouns are obtained from which the vibakthi class is identified. In Malayalam there are 7 vibakthi.

Table 1: Vibakthi list

Vibakthi	Suffix	
Nirdeshika		കവി
Prathigrahika	എ	കവിയെ
Samyojika	ഓട്	കവിയോട്
Udeshika	ക്ക് ന്	കവിയ്ക്ക്
Prayogika	ആൽ	കവിയാൽ
Sambanthika	ഉടെ	കവിയുടെ
Adarika	ഇൽ കൽ	കവിയിൽ

D. Semantic roles

Semantic roles such as agent patient, instrument, beneficiary, experiencer, recipient etc are identified.

- Agent: Deliberately performs the action
- Experiencer: The entity that receives sensory or emotional input
- Theme: Undergoes the action but does not change its state .
- Patient: Undergoes the action and changes its state
- Instrument: Used to carry out the action
- Beneficiary: The entity for whose benefit the action occurs

Karaka Relations in Malayalam

1. Kartha Kararaka

The activity actually resides in or springs from karta. The result of the verb is reflected in Kartha karakamVibakthi corresponding to kath karaka is Nirdeshika and prayogika.

Nirdeshika (Nominative)

Role can be agent , experiencer or causer.

രാമു ഓടി

Here the sentence is subject+verb form and രാമു is Agent.

എൽസി കരഞ്ഞു

If the verb is emotional verb then the role of എൽസി is experiencer.

രാമു രാജുവിനെ കരയിച്ചു

Here the sentence is subject+object+verb form and if the verb is causative verb then the noun with nirdeshika vibakthi is take the role as Agent.

2. Karma Karakam

It denote the object of the sentence.vibakthi is Prathigrahika (Accusative) or Nirdeshika

അപ്പു(Agent) തത്തയെ(Patient) പിടിച്ചു

The sentence is subject+object+ verb form then the Noun with Prathigrahika vibakthi is take the role of Patient. And if the verb is causative verb then the role is Experiencer.

രാമു(Agent) രാജുവിനെ(Experiencer) കരയിച്ചു

Nirdeshika(Nominative)

The object denotes Patient.

രാമു(Agent) മാങ്ങ(Patient) കഴിച്ചു

3. Karana Karakam

It denotes the instrument for performing the action. It will be in the prayojika (Instrumental) vibakthi form.

വടിയാൽ(Instrument) അടിച്ചു

4. Hethu Karakam

It denotes the reason for verb and it is in Prayogika vibakthi.The role is agent.

5. Sakshi Karakam

Sakshi Karakam will be in Samyojika(Sociative) vibakthi and it denote role as Participant.

രാമു(Agent) രാജുവിനോട്(Participant) പറഞ്ഞു

6. Swami Karakam

Swami karakam denote the beneficiary of the activity and vibakthi is Udeshika(Dative).

രാമു (Agent) രാജുവിന് (Beneficiary) പണം (Patient) നൽകി

7. Adikarana Karakam

Here the vibakthi is Adaraika(Locative) vibakthi and it denote location.

പുസ്തകം മേശയിൽ (Location) ഉണ്ട്

E. Semantic Role Labeling Algorithm

Based on the above Karaka Relations we developed an algorithm for Semantic Role Labeling.

```
SRL Algorithm for a sentence
Step1: identify the verb from the POS tagging and chunking.
Step2: select the NP chunks and find the vibakthi from morphology.
Step3: if the sentence is Noun+verb form
    Step3.1: if nirdeshika vibakthi then check the verb class
        Step3.2: if its causative then it is labeled as causer.
        Step3.3 : if it is emotional verb then it is labeled as experiencer .
        Step3.4: otherwise labeled as agent.
Step3.2: if it is prathigrahika vibakthi it is labeld as Patient.
Step3.3: if it is udeshika vibakthi then it is labeled as Beneficiary.
Step3.4: if it in samyojika vibakthi it is labeled as patient.
Step4: if the sentence is noun1+noun2+verb form
    Step4.1: one noun is nirdeshika
    Step4.1.1. if other is also nirdeshika and its semantic class is non human then nirdeshika noun is agent and other is patient.
    Step4.1.2: if other noun is prathigrahika or samyojika then check verb class then it is causative verb the nirdeshika noun is causer otherwise it is agent and other is patient.
    Step4.1.3: if other noun is udeshika verb class is beneficiary then udeshika noun is beneficiary other is agent. Otherwise udeshika noun is patient and other is agent.
Step5: if the sentence is noun1+noun2+noun3+verb form
    Step5. 1:if noun is nirdeshika and its semantic class is non human then nirdeshika noun is agent.
    Step4.1.2: if other noun is prathigrahika or samyojika then check patient.
```

Step4.1.3: if other noun is udeshika verb class is beneficiary then udeshika noun is beneficiary.

Step6: Stop

F. SRL using YamCha Toolkit

If the above algorithm fails to find the semantic roles then we use machine learning approach for SRL. Here we used Yamcha for SRL. YamCha (Yet Another Multipurpose Chunk Annotator) is a generic, customizable, and open source text chunker oriented toward a lot of NLP tasks, such as POS tagging, Named Entity Recognition, base NP chunking, and Text chunking. YamCha is using the machine learning algorithm called ‘Support Vector Machines’ and it is exactly the same system which performed the best in the CONLL2000 Shared Task, Chunking and BaseNP chunking task. The features of YamCha include its high performance chunker based on Support Vector Machines and partial chunking. It is independent from the given task, training/testing with any data which can be seen as a “generic” text chunking task. It uses PKE/PKI which make the classification (chunking) speed faster than the original SVMs. YamCha can also redefine feature sets (window-size), parsing-direction (forward/backward) and algorithms of multi-class problem (pair wise /one vs rest) and it also uses C/C++ library.

Here we used three column format.1st column is 'word', second column is 'POS tag' third column is 'semantic roles' .The last column represents a true answer tag which is going to be trained by SVMs. First of all, run **yamcha-config** with **--libexecdir** option. The location of Makefile which is used for training is output. Please copy the Makefile to the local working directory. There are two mandatory parameters for training.

CORPUS: The location of file which is written in the training/test format.

MODEL: Prefix name of model file(s)

D. Conceptual graph representation

Conceptual Graph (CG) expresses meaning in a form that is logically precise, humanly readable, and computationally tractable. With their direct mapping to language, conceptual graphs serve as an intermediate language for translating computer-oriented formalisms to and from natural languages. With their graphic representation, they serve as a readable, but formal design and specification language. CGs have been implemented in a variety of projects for information retrieval, database design, expert systems, and natural language processing. In diagrammatic representation of CG rectangular box represents the concept node and oval represents the semantic relation.

For example

ഗോവിന്ദൻ സീതയ്ക്ക് പണം നല്കി

CG representation

{നല്കി (ഗോവിന്ദൻ: Agent) (പണം: Patient) (സീത: Beneficiary)}

IV. CONCLUSION

The Proposed system is tested using 1000 sentence and we get 90% accuracy. The system can be utilized for the applications such as document summarization and natural language generation etc.

REFERENCES

[1] Gildea and D. Jurafsky, “Automatic labeling of semantic roles”, In Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), pages 512520, Hong Kong, October 2000.

[2] Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James Martin, and Dan Jurafsky, “Shallow Semantic Parsing Using Support Vector Machines”, CSLR Tech. Report, CSLR-TR-2003-1.

[3] Kadri Hacioglu and Wayne Ward, ” Target word detection and semantic role chunking using support vector machines”, In HLT-03.

[4] Akshar Bharati, Vineeth Chaithanya, Rajeev Sangal,”Natural Language Processing: A Paninian Perspective”, Prentice-Hall of India, New Delhi .

[5] A.R.RajarajaVarma , Keralapanineeyam, Sahitya Pravarthaka ,C S Ltd., Kottayam,1968

[6] Jisha P. Jayan and Rajeev R. R, ”Parts of Speech Tagger for Malayalam”, IJCSIT International Journal of Computer Science and Information Technology, Vol.2, No.2, December 2009, pp.209-213.

[7] Saranya S. “Morphological Analyzer for Malayalam verbs”, M.Tech Thesis, Amrita School of Engineering, Coimbatore, 2008.

[8] A.R.RajarajaVarma, Keralapanineeyam ,SahityaPravarthaka,CSLtd., Kottayam,1968

[9] Daniel Jurasfky and James H. Martin, An introduction to Natural Language Processing, Computational Linguistics , and Speech Recognition, PearsonEducation , Inc.2000.