Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SEMANTIC ROLE LABELING OF MALAYALAM WEB DOCUMENTS IN CRICKET DOMAIN

[1]SUNITHA C, [2]DR. A JAYA, [3]AMAL GANESH

[1]Research Scholar, Department of CSE, BSAR Crescent Institute of Science & Technology, Chennai, India
[2]Professor, Department of CA, BSAR Crescent Institute of Science & Technology, Chennai, India
[3]Asst. Professor, Dept. of CSE, Vidya Academy of Science & Technology, Kerala, India
E-mail: [1]sunitha@vidyaacademy.ac.in , [2]jayavenkat2007@gmail.com, [3]amal.ganesh@vidyaacademy.ac.in

ABSTRACT

Document Summarization is an ongoing research work in the field of Natural Language Processing which will provide a summary which is almost like a summary generated by a human being with the help of NLP tools and techniques. Since the information used across the digital world is exponentially increasing, automatic summarization techniques gained attention especially abstractive methods. But producing an effective abstractive summary, first, the text documents should be represented semantically. From this representation, important sentences must be selected using some strategies and finally the abstractive summary is generated. Representing the sentences in natural language semantically faces many challenges. Various works have been carried out for extracting the semantics of the sentences. Semantic role labeling is a technique in NLP to detect the semantically related arguments of a predicate or verb in a sentence and their grouping into one of the related roles. So this technique can be used to represent the sentences meaningfully and can be further used in different applications such as question answering system, information extraction, summarization, text categorization etc. Currently, limited works are done in Malayalam towards semantic role extraction. Domain based works will give better results. In this paper the semantic roles of important words in Malayalam Web documents pertaining to cricket domain are identified.

Keywords: *Semantic Role Labeling, Karaka relations, Memory Based Learning, Vibhakthi, Chunking*

1. INTRODUCTION

In this modern era, data retrieval across websites and other informative media are used everywhere irrespective of the languages we speak. This led to the rapid growth of information. These enormous volume of information made the necessity of having NLP applications like summarization. Summarization is the task of abstracting the information related to a particular area/domain/topic from various reliable sources. Summarization techniques are broadly categorized into two groups namely, extractive summarization techniques and abstractive summarization techniques [1] [2]. Extractive summarization techniques produce summary based on the key features in the input text. Statistical methods such as term frequency, location, cue method, title/headline word, sentence length, similarity, proper noun, proximity etc. are used to select the sentences needed for summary generation. Whereas in abstractive summarization techniques, the underlying semantics of the input text is conceptualized and based on that summary is generated.

Representing a natural language text semantically, is a challenging task particularly in the case of Indian languages like Malayalam because of its free word order property. Mainly template based and ontology based representations are used for semantic representation. In template based representation, humans manually create linguistic patterns and extraction rules which are used for template slots. Since this approach rely on manual effort, it is very time consuming and will not be able to capture similarity information between documents. Ontology based representation also rely on manually built ontology, which again depends on human experts and is also time consuming. The drawbacks of these semantic representations will be handled by semantic role labeling [SRL] because it relies on well known Panini's Karaka theory [4]. By incorporating semantic role labels, the underlying meaning of the documents can be efficiently extracted.

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Semantic arguments related with the predicate or verb in a sentence and their grouping into the defined roles can be done with semantic role labeling[2]. Semantic roles are one among the linguistic constructs based on Panini's Karaka theory [4]. This is one of the important step towards identifying the meaning of a sentence. So the semantic roles can be effectively used in various NLP applications. This sort of semantic representation is actually a higher-level of abstraction than a lower level syntax tree.

Even though works are carried out in foreign languages, limited works are done in Indian languages especially in Malayalam. In Dravidian languages like Malayalam, both the syntax analysis and semantic analysis of sentences can be effectively done by using Karaka theory.

Semantic roles build a structure for representing the semantics for a given Malayalam input text by extracting karaka relations. The purpose of the Paninian approach is to develop a theory of human natural language communication. The grammar pertaining to this theory of communication can be represented by a set of rules. These rules will help to establish a relation between what the speaker intends to say, his utterance and what information the hearer hears, the meaning he extracts. Any action will contain an activity and a result which is represented by the root verb. When we reach the result state, the action is complete. Thematic role analysis can be defined as the study of roles related to specific verbs and the different classes of verbs. It is also known as case role (karaka) analysis. In Malayalam, the meaning of the root word is changed by using inflection property, The inflection can also be used to relate the word with other words. These inflections can be distinguished by vibhakthis [5]. The figure 1 shows the vibhakthis used in Malayalam language.

വിഭക്തിയുടെ പേര്	പ്രത്യയം	ഉദാഹരണം
നിർദ്ദേശിക	പ്രത്യയമില്ല	കുട്ടികൾ
പ്രതിഗ്രാഹിക	എ	മനുഷ്യരെ
സംയോജിക	ഓട്	മനുഷ്യരോടു
ഉദ്ദേശിക	ൻ, ക്ക്	രാമനു, സീതയ്ക്ക്
പ്രയോജിക	ആൽ	മനുഷ്യരാൽ
സംബന്ധിക	ന്റെ, ഉടെ	മനുഷ്യരുടെ
ആധാരിക	ഇൽ, കൽ	മനുഷ്യരിൽ

Figure 1: Vibhakthis of Malayalam

Based on Paninian theory, the process of understanding a sentence can be defined at four levels. They are semantic level, karaka level, vibakthi level and surface level. The karaka level is in between semantics level and syntax level and therefore karaka level is related to both semantics and syntax. The post position markers after noun or surface case ending of noun can be used to identify Karaka relation. These markers and case endings are called vibhakthi. In Malayalam karaka relations are analyzed from vibakthi and post position markers.

In Vibhakthi processing, the verb of a sentence is the central, binding element of the sentence during the semantic analysis. This idea is used by the sanskrit linguists like Panini in their work. The noun and verb of a Malayalam sentence are very important and related to each other. Karakam is used to define the relation between noun and verb [6] [7] [8]. Seven karakas are there in Malayalam :

1. Karthu Karakam - Subject
2. Karma Karakam - Object
3. Karana Karakam – Instrument
4. Kaarana Karakam - Instrument
5. Sakshi Karakam - Experiencer
6. Swami Karakam - Beneficiary
7. Adhikarana Karakam – Locative

As karakas define the relationship between nouns and verbs, extracting karakas from sentences will be very useful for representing the semantics of the sentences.

2. MEMORY BASED LEARNING

Memory-Based Language Processing (MBLP) is one of the efficient approaches used in NLP [25]. The process is based on the technique called Memory-Based Learning (MBL). This technique learns from experiences instead of extracting rules or make abstract representations. That is this method uses the concept of reusing memory for remembering that experiences directly. Memory-based learning and the method of problem solving works on two important principles: one is learning and the other is reuse. Learning is done by simply storing the experiences in memory, and later an unsolved problem is solved by reusing solutions obtained from already solved similar type of problems.

An MBLP system has two components: one is a learning component and the other is a performance component. Learning component is memory-based and performance component is similarity-based. The learning component stores

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

examples in memory without abstraction, selection, or restructuring. In the performance component contained in an MBLP system, the stored examples are used as a basis for mapping input to output; input instances are classified by assigning them an output label. During classification, a previously unseen test instance is given as an input to the system. MBLP approach has already been applied successfully to many problems in NLP, such as morphological analysis, POS tagging, chunking, etc.

Memory-Based Tagging (MBT) works on the principle of Memory-Based Learning and is used as a method to sequence tagging. It works on the idea that words occurring in similar contexts will have the same tag.

3. RELATED WORKS

Various research works have been adopted to extract the semantic roles from text documents.

The first attempt to label the semantic roles automatically has been carried out by Gildea and Jurafsky [3]. They used statistical classifiers in their work to extract the semantic roles. In their work, each sentence is labeled manually using the semantic roles mentioned in the Frame Net semantic labeling project. Different type of features such as head word, parse tree path, phrase type and position are considered in their work.

Radhika K.T et al proposed a system [5] for semantic representation of Malayalam text. In their work syntactic analysis and semantic analysis is done using semantic roles which works based on karaka theory. Semantic roles are expressed as conceptual structures.

Shabeena and Sandeep in their work [6], used semantic role labeling approach to perform shallow parsing. There are two types of parsing, Deep semantic parsing and Shallow parsing. In deep semantic parsing, the underlying semantics of the sentence is represented in the form of predicate logic or any formal representation. But in shallow parsing case role identification is being carried out. Semantic roles of nouns are identified using Karaka theory.

Jisha and Rajeev used the Karaka theory based on Paninian's grammar to identify the semantic roles [7]. Case markers are used to identify the relation between noun and verbs. A statistical model based on CRF is used in their work.

Sindhu L et. al [8] described a method for detecting plagiarism in Malayalam documents. Semantic role labels are extracted and computed the similarity to detect plagiarism and. These semantic roles identifies the relation between verb and its arguments.

Hacioglu and Ward in their work [9], considered semantic role labeling problem as a chunking task . In this semantic chunks are defined as a group of words that fills the semantic roles defined in semantic frame. Chunking task can be easily converted as a tagging task using IOB representations. They used support vector machines to solve semantic chunking problem.

Van den Bosch et al., in their work considered semantic role labeling as a classification problem. Memory-based learning is applied to semantic role labeling [10]. They have used feature selection and parameter optimization by iterative deepening in developing the system.

Hacioglu, in his paper, explained the development of a dependency tree based semantic role labeler. This is done by considering semantic role labeling as a classification problem. Here the dependency relations are classified into one of the labeled semantic roles.

Kouchnir implemented a semantic role labeler for the CoNLL 2004 shared task [11].The task was divided into two sub-tasks, recognition of arguments of a proposition and assignment of semantic role labels to the arguments extracted. A semantic databases like Wordnet is used to cluster the words.

Park et al. proposed a semantic role labeling model with two phases [12]. The two phases are the identification phase and the classification phase. Various machine learning techniques have been applied to solve the Semantic Role Labeling problem. The techniques include support vector machines, conditional random fields (CRFs), and memory-based learning.

Das et al. extracted semantic role labels of nouns in Bengali using the method of 5W distilling [13]. The 5W task summarizes the information contained in a natural language sentence by considering it as the answers to the 5W questions. These questions are Who, What, When, Where and Why. Maximum Entropy (MEMM) based statistical model accompanied by rule-based post processing is used for SRL. Semantic roles were extracted to build semantic structure of Malayalam text. The system used vibhakthi (case endings) to Karaka mapping for identifying the semantic roles.

Lakshmi et. al presented a clause boundary identification system in their paper which is useful for better semantic representations. CRF method is used for clause boundary identification and is a statistical learning approach.

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Maas Anwar et. al identified semantic role for two Indian languages, Hindi and Urdu. They used statistical classifiers. The classifiers are annotated with hand crafted semantic roles under Prop Bank project.

S. Lakshmana Pandian and T. V. Geetha, performed semantic role labeling for Tamil documents. A hybrid approach based on syntactic, semantic and statistical evidences is used in their work. Two phases are used – Learning phase and Evaluation phase. Two components are used in Learning phase, Maximum entropy model and a learning component. Four components are used in Evaluation phase, MEM evaluator, Verb Frame Invoker, Rule Based Probability Assigner and Expectation Maximizer component.

From the literature review, it is clearly understood that semantic role labeling is widely used in various applications such as information retrieval, document summarization, text categorization etc. In Malayalam, only one work is reported towards abstractive summarization which is not using semantic role labeling technique. SRL techniques can be explored further to be utilized in Malayalam document summarization. Very few works have been carried out in Malayalam in the area of semantic role labeling. These works rely mainly on algorithms based on support vector machines and CRF. The proposed work is domain dependent and memory based learning approach is used for SRL. MBLP technique learns from experiences instead of extracting rules or make abstract representations and uses the concept of reusing memory for remembering that experiences directly. This concept works well in a particular domain.

4. PHASES OF SEMANTIC ROLE LABELING

The proposed system helps to identify the semantic roles of Malayalam web documents in Cricket domain. As part of this work the online Malayalam documents in the cricket domain are collected, The over all architecture depicted in Figure 2 consists of various phases like Tokenization, POS tagging, Chunking, Clause boundary identification and finally Semantic role labeling.

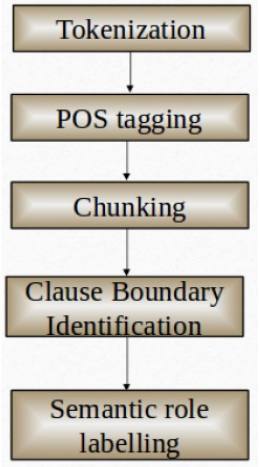

Figure 2: Architecture Diagram of Semantic Role Labeling

4.1 Tokenization

Tokenization is the process of splitting a text into fundamental units or tokens like words and punctuation. This is the first preprocessing step in natural language processing. Consider the example,

സച്ചിൻ ബാറ്റു ചെയ്തു. (Sachin battu cheythu.) / (*Sachin did batting*)

Here tokens are

സച്ചിൻ, ബാറ്റു, ചെയ്തു (Sachin, battu, cheythu)

Tokens are generated considering the space between words. This step is performed by stripping the text using space.Based on this space it split the sentences into individual tokens. The implementation of tokenization is done using a python program.

4.2 Compound Word Splitting

Malayalam is an inflected and agglutinative language. A root word in Malayalam can be inflected in many ways. A word can be inflected by adding suffixes to the word. For instance a word can inflected by pratyayas or case markers.

For example:

Some of the compound word formations in Malayalam are given below.

1. Compound words with noun-auxiliary verb combination.

   Eg: പുസ്തകമാണ് / (pusthakamaanu) // (it's book)

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This can be splitted as പുസ്തകം /pusthakam (noun) + ആണ് / aanu (auxiliary verb)

2. Compound words with verb-auxiliary verb combination

   Eg: ചിരിച്ചതാണ് / chirichathaanu //

   This can be splitted as ചിരിച്ചത് / chirichath (verb) + ആണ് / aanu (auxiliary verb)

3. Compound words with adjective-auxiliary verb combination

   Eg: വലുതാണ് / valuthaanu //

   This can be splitted as വലുത് / valuth (adjective) + ആണ് / aanu (auxiliary verb)

4. Compound words with verb-verb combination

   Eg: ഓടിക്കളിച്ചു / odikkalichu //

   This can be splitted as ഓടി / odi (verb) + കളിച്ചു / kalichu (verb)

Deciding the POS tag of compound words is a difficult task, especially when the words belongs to different parts of speech categories. Therefore it is be better to split the compound words and tag them accordingly [14]. So the tokenized sentences in the corpus are given as input to a compound word splitter. The compound word splitter identifies the compound words and split them into its constituents. Compound Word Splitter is implemented using a python program by incorporating a classifier.

For eg: സെഞ്ചുറിയടിച്ചു / (centuariyadichu) can be splitted as two words like സെഞ്ചുറി (centuary), അടിച്ചു (adichu).

4.3 POS Tagging

POS tagging is the process of assigning a label o tag to each and every word in the text based on the context. The tokens generated after compound word splitting are tagged using MBT. BIS tagset has been recommended to be used as a common tagset for the part of speech annotation of Indian languages [15] [16]. POS tagger is implemented using Memory Based Tagger (MBT) [17]. The training set consists of 14000 tokens pertaining to cricket domain.

The Traning format will be:

സച്ചിൻ (Sachin) - N-NNP

ബാറ്റ് (Bat) – N-NN

4.4 Chunking

Chunking is used to split the large sentence into small phrases [18]. This will be helpful in finding the semantic roles easily . The different types of chunks used in this system [19] are given in the Figure 3.

Sl. No.	Chunk Type	Tag Name
1	Noun Chunk	NP
2.1	Finite Verb Chunk	VGF
2.2	Non-finite Verb Chunk	VGNF
2.3	Infinitival Verb Chunk	VGNIF
3	Adjectival Chunk	JJP
4	Adverb Chunk	RBP
5	Chunk for Negatives	NEGP
6	Conjuncts	CCP

Figure 3: Chunk Tag Set

4.4.1 Noun chunk (NP)

A noun chunk is the important chunk and is made up of non-recursive noun phrases and post-positional phrases. The head of noun chunk will be a noun. Usually, specifiers will be the left boundary of the noun chunk and head noun acts as the right boundary. Consider the following example:

ഈ മഹത്തായ സെഞ്ചുറി / ee mahathaya centuary (This great centuary) will be represented as,

(ഈ / ee (DM-DMD) മഹത്തായ / mahathaya (JJ) സെഞ്ചുറി / centuary (N_NN))$_{NP}$

4.4.2 Verb chunk

There are different types of verb chunks. A verb group consists of the main verb and its auxiliaries, if any. The different types of verb chunks with their associated tags are explained below.

- Finite Verb Chunk (VGF): A verb group sequence (V VAUX VAUX . .) contains a main verb and its auxiliaries. The group itself can be finite or non- finite. The main verb will be finite. Consider one example:

  സച്ചിൻ ബാറ്റ് ചെയ്തു / Sachin bat cheythu. (Sachin did batting.)

  This will be represented as:

  സച്ചിൻ / Sachin (N-NNP) ബാറ്റ് / bat (N-NN) (ചെയ്തു / cheythu (V_VM_VF))$_{VGF}$.

- Non-Finite Verb Chunk (VGNF): A non-finite verb chunk will have a non-finite verb as its main verb. Consider one example:

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

കളി നടക്കുന്ന സമയത്ത് മഴ പെയ്തു. (kali nadakkunna samayath mazha peythu. / At the time of playing, it is rained.)

This will be represented as :

കളി / kali (N-NN) (നടക്കുന്ന / nadakkunna (V_VM_VNF))$_{VGNF}$ സമയത്ത് / samayath (N-NN) മഴ / mazha (N-NN) പെയ്തു / peythu(V_VM_VF).

- Infinitival Verb Chunk (VGINF): An infinitival verb chunk will have a infinitive verb as its main verb. Consider one example:

ദിവസവും നടക്കുന്നത് എനിക്ക് ഇഷ്ടമാണ്. / divasavum nadakkunnath enikku ishtamaanu (I like to walk daily)

This will be represented as :

ദിവസവും / divasavum (N-NN) (നടക്കുന്നത് /nadakkunnath(V_VM_VINF))$_{VGINF}$ എനിക്ക്/enikku (PR_PRP) ഇഷ്ടമാണ് /ishtamaanu (V_VM_VF).

4.4.3 Adjectival Chunk (JJP)

This chunk consists of all adjectival chunks including the predicative adjectives. Consider the example:
സച്ചിൻ നല്ല ബാറ്റ്സ്മാൻ ആണ് / Sachin nalla batsman aanu (Sachin is a good batsman.)

This will be represented as :

സച്ചിൻ / Sachin (N_NNP) (നല്ല / nalla (JJ))$_{JJP}$ ബാറ്റ്സ്മാൻ / batsman (N_NN) ആണ് / aanu (V_VM)

4.4.4 Adverb Chunk (RBP)

This chunk consists of pure adverbial phrases. Consider the example:

മത്സരം വേഗത്തിൽ അവസാനിച്ചു / malsarm vegathil avasaanichu (competition finished fast).

This will be represented as :

മത്സരം/malsarm(N_NN) (വേഗത്തിൽ / vegathil (RB))$_{RBP}$ അവസാനിച്ചു / avasaanichu (V_VM_VF)

The traning set of chunker consists of words/tokens, then corresponding POS tag, and finally the chunk tag. An instance of the training set is as follows :

സച്ചിൻ / Sachin N-NNP B-NP
ബാറ്റു / battu N-NN B-VGF
ചെയ്തു / cheythu V-VGF I-VGF
. RD-PUNC

For training the chunker 2500 chunk tag sets were used in our work. MBT is used for tagging. After training, the system automatically predicts the chunk tags based on the tags which were used for training.

4.5 Clause Boundary Identification

Sometimes the sentences are very large and complex. In such cases it would be better to split the complex sentences into clauses for representing the meaning efficiently [20] [21] [22]. In this work, the clause boundaries are identified considering it as a classification problem. The clause boundaries are identified from the clause start and clause end information. The features used to find the clause start and clause end are listed below:

- word
- POS tag
- chunk tag
- Number of verbs in the sentence
- Number of conjunctions and subordinations in the sentence
- Number of negations in the sentence

Consider one example for clause start identification training input:

സച്ചിൻ / Sachin N-NNP B-NP 1 0 0 -
ബാറ്റു / battu N-NN B-VGF 1 0 0 -
ചെയ്തു / cheythu V-VGF I-VGF 1 0 0 -
. RD-PUNC O 1 0 0 -

An instance of the clause start training set is :

സച്ചിൻ / Sachin N-NNP B-NP 1 0 0 S
ബാറ്റു / battu N-NN B-VGF 1 0 0X
ചെയ്തു / cheythu V-VGF I-VGF 1 0 0 X
. RD-PUNC O 1 0 0 - X

An instance of the clause end training set is :

സച്ചിൻ / Sachin N-NNP B-NP 1 0 0 X
ബാറ്റു / battu N-NN B-VGF 1 0 0 X
ചെയ്തു / cheythu V-VGF I-VGF 1 0 0 X
. RD-PUNC O 1 0 0 E

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The features used to find the clause boundaries are listed below:

- word
- POS tag
- chunk tag
- Number of verbs in the sentence
- Number of conjunctions and subordinations in the sentence
- Number of negations in the sentence
- clause start tag
- clause end tag

An instance of the clause boundary identification input is :

സച്ചിൻ / Sachin	N-NNP	B-NP	1 0 0 S X -
ബാറ്റ / battu	N-NN	B-VGF	1 0 0 X X -
ചെയ്തു / cheythu	V-VGF	I-VGF	1 0 0 -X X -
.	RD-PUNC	O	1 0 0 X E

An instance of the clause boundary identification training instance is :

സച്ചിൻ / Sachin	N-NNP	B-NP	1 0 0 S X S
ബാറ്റ / battu	N-NN	B-VGF	1 0 0 S X S
ചെയ്തു / cheythu	V-VGF	I-VGF	1 0 0 X X *
.	RD-PUNC	O	1 0 0 X E EE

The training corpus of the clause boundary identifier contains 9 columns. The first six columns uses the same features used for clause start and clause end identification. The seventh column contains the clause start tags. The eighth column contains the clause end tags. The last column indicates the clause boundary tag. This module is implemented using MBT.

4.6 Semantic Role Labeling

The grammatical relations in a sentence can be effectively used to represent the meaning of a sentence. These relations are of two types, karaka relations and non-karaka relations. Normally in a sentence, there will be an action and also the participants who took part in the action. The action is denoted by verb. The participants are represented by karakas. These karaka relations can be identified by pratyayas mentioned in the Table 1.

After the clause identification phase, identify the verb in each clause and also find the vibhakthis of nouns with the help of morphological parsing. The following rules are used to identify the karaka relations or semantic role labels in Malayalam.

1. If a clause contains only one noun and if its pratyaya is nirdeshika then it is called karta karaka (k1).

Eg: ദ്രാവിഡ് കളിച്ചില്ല / Dravid kalichilla (Dravid didn't play). Here ദ്രാവിഡ് / Dravid is having no pratyaya (ie Nirdeshika) and hence it's role is identified as k1.

2. If the clause contains two nouns both with nirdheshika, the first noun will be karta karaka and second one will be karma karaka.

Eg: സച്ചിൻ ക്രിക്കറ്റ് കളിച്ചു / Sachin cricket kalichu (Sachin played cricket). Here there are two nouns in nirdheshika. So the first noun സച്ചിൻ / Sachin is identified as k1 and ക്രിക്കറ്റ് / cricket as k2.

3. If the clause contains two nouns, a noun with nirdheshika (n1) and another noun is prathigrahika or samyojika (n2) then n1 is karta and n2 will be karma

Eg: ഇന്ത്യ ഇംഗ്ലണ്ടിനെ തോൽപിച്ചു / India Englandine tholpichu (India defeated England.). Here ഇന്ത്യ / India is identified as k1 and ഇംഗ്ലണ്ടിനെ / Englandine as k2

Eg: ഇന്ത്യ ഇംഗ്ലണ്ടിനോട് പരാജയപ്പെട്ടു / India Englandinodu paraajayappettu (India failed to England). Here ഇന്ത്യ / India is identified as k1 and ഇംഗ്ലണ്ടിനോട് / Englandinodu as k2

4. If the clause contains two nouns, a noun with nirdheshika (n1) and another noun (n2) in prayojika then n1 is karma and n2 will be karta.

Eg: ഇംഗ്ലണ്ട് ഇന്ത്യയാൽ പരാജയപ്പെട്ടു / England Indiayaal paraajayappettu (India defeated by England). Here ഇംഗ്ലണ്ട് / England is identified as k2 and ഇന്ത്യയാൽ / Indiayaal as k1.

5. If the clause contains two nouns, a noun with udeheshika (n1) and another noun in nirdheshika (n2), then n1 is karta and n2 will be karma

Eg: സച്ചിന് പരിക്ക് പറ്റി / Sachinu parikku patti (Sachin had injury). Here സച്ചിന് / Sachinu is identified as k1 and പരിക്ക് / parikku is identified as k2.

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Semantic Role Labeling is implemented by considering it as a tagging problem. The solution to this problem has been developed using MBT. The features used for training MBT are listed below.

• word

• POS tag

• chunk tag

• Number of verbs in the sentence

• Number of conjunctions and subordinations in the sentence

• Number of negations in the sentence

• clause boundary tag

Example for Semantic Role labeller input

സച്ചിൻ / Sachin N-NNP B-NP 1 0 0 S -

ബാറ്റു / battu N-NN B-VGF 1 0 0 * -

ചെയ്തു / cheythu V-VGF I-VGF 1 0 0 * -

. RD-PUNC O 1 0 0 E -

Example for Semantic Role labeller output

സച്ചിൻ / Sachin N-NNP B-NP 1 0 0 S k1

ബാറ്റു / battu N-NN B-VGF 1 0 0 * k2

ചെയ്തു / cheythu V-VGF I-VGF 1 0 0 * r

. RD-PUNC O 1 0 0 E -

The token സച്ചിൻ has no vibhakthi prathyayam. That is the default vibhakthi is Nirdesika because the token ends without any prathyayam. So the token will be Agent or Subject. If a word without any prathyayam and it is a human or non-human, then it is defined as Agent/Subject. If the token is സച്ചിനെ then the vibhakthi prathyayam will be Udesika. Then the corresponding Karaka role will be Beneficiary [23] [24].

If the given token is a verb, then it will be idenitified using the tag corresponding to the token. There may be different types of verbs exist. But the system will take V-VM-VF as root verb. because the finite verb have a finite clause with a termination.

Memory-based language processing (MBLP) technique is used to implement the approach.

5. RESULTS AND DISCUSSIONS

In this paper, the semantic roles of nouns in relation with verbs is identified. Semantic roles are identified using vibhakthis in Malayalam. Eventhough there are many karaka relations mentioned in Panini's karaka theory, we considered the important karaka relations karta and karma only. Using karta, karma and verb relationships, a sentence can be semantically represented.

For implementing the system, Malayalam web documents in cricket domain were taken into consideration. The documents are collected from online Malayala Manorma news paper. Semantic Role Labeling is implemented by considering it as a tagging problem. The solution to this problem has been developed using MBT. POS tagger and chunker is also developed with domain based using MBT. This approach is found to be efficient since it reuses the solved examples. Also it is domain based which further improves its efficiency.

Very few works were done in Malayalam in the area of abstractive summarization due to the lack of resources for semantic representation. Semantic role labels especially karta and karma can be used for representing the sentences semantically which can be further used for various applications in Malayalam like Text categorization, Document Summarization, Sentiment analysis etc. The accuracy can be improved by enriching the training set.

The POS tagger, chunker and semantic role labeler are trained with the tokens from Malayalam news documents. The test sentences are also created from online cricket news. The system is tested with these test sentences. Table 1 shows the performance measures.

Table 1: Performance Measures

Test Cases	Precision	Recall	F-Measure
12	0.80	0.89	0.84
24	0.83	0.82	0.82
36	0.89	0.80	0.84

The analysis is done based on the primary performance attributes such as Precision, Recall and F-measure. Precision is the ratio of relevant items retrieved in relation to the total number of items retrieved which has both relevant and non-relevant items. Whereas Recall is the number of items retrieved in correlation to the number of relevant items retrieved from the dataset. F-measure is a composite score combining both Precision and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Recall. It is the weighted harmony score of the Precision and Recall.

$$\text{F- Measure} = \frac{2 \cdot P \cdot R}{P + R}$$

Here we tried to implement semantic role labeling for Malayalam sentences to represent them semantically. Compared to the other works in Malayalam pertaining to SRL, we used domain based memory learning approach which promised better results. Since it is domain based, this technique can be effectively used for domain based works in Malayalam NLP. POS tagger, Chunker and semantic role labeler were trained using domain based data. By increasing the size of data corpus, the accuracy can be further increased. Also in our work, we considered only karta and karma karaka relatios for SRL extraction. By including more karaka relations, better semantic representation can be achieved.

6. CONCLUSION

In this paper semantic role labeling for Malayalam web documents in cricket domain is implemented. It is found that domain based works produce better results. Semantic Role labeling is implemented sing MBLP approach [25]. It is a machine learning approach. Training is done by using memory based tagger (MBT). This method is used because of the two principles- simple storage representation in memory & solving new problems by reusing solutions from previously similarly solved problems. To find out semantic role labels tokenization, POS tagging, Chunking, Clause boundary identification have been carried out. A compound word splitter is used initially. Since all the tasks have been considered as a classification problem, the system gave a good performance. The main limitation of this method is that it relies on trained POS tagger and chunker as no standard pre processing tools are available in Malayalam. Also we considered karta and karma karaka relations only in this work. The accuracy can be increased more by enriching the training set and also by including more karaka relations. Since the semantic role labels represent the meaningful information in a sentence, they can be effectively used for NLP applications like Information retrieval, Document Summarization, Question Answering etc.

REFERENCES:

[1] Dipanjan Das, Andre F.T.Martins, “A survey on automatic text summarization”, Language Technologies Institute, Carnegie Mellon University, 2007.

[2] Atif Khan , Naomie Salim, “A Survey on Abstractive Summarization Methods”, Journal of Theoretical and Applied Information Technology, Vol. 59, 2014.

[3] D. Gildea and D. Jurafsky, “Automatic labeling of semantic roles,” Computational linguistics, vol. 28, no. 3, pp. 245–288, 2002.

[4] R. S. Bharati Akshar, Vineet Chaitanya, Natural Language Processing: A Paninian Perspective. Prentice Hall of India, Delhi, 1995.

[5] Radhika K.T. , Dr. P.C. Reghu Raj (2013) Semantic Role Extraction and General Concept Understanding in Malayalam using Paninian Grammer, International Journal of Engineering Research and Development, e-ISSN: 2278-067X, p-ISSN: 2278-800X, V:9, Issue 3, PP. 28-33.

[6] Shabina Bhaskar, Sandeep Chandran (july 2015) Semantic Parsing Approach in Malayalam for Machine Translation, International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-018, vol. 4, Issue – 07.

[7] Jisha P Jayan, J Satheesh Kumar, Semantic Role Labelling for Malayalam, IJCTA, 9 (10), 2016, pp. 4725-4731.

[8] Sindhu L, Suman Mary Idicula (November 2015)SRL based Plagiarism Detection for Malayalam Documents IJCSI International Journal of Computer Science, Issues, volume 12, Issue 6, ISSN (print): 1694-0814/ ISSN (online): 1694-0784.

[9] 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, 2003, pp. 25–27.

[10] A. van den Bosch, S. Canisius, W. Daelemans, I. Hendrickx, and E. T. K. Sang, “Memory-based semantic role labeling: Optimizing features, algorithm, and output,” in Proceedings of the CoNLL-2004, 2004.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[11] B. Kouchnir, "A memory-based approach for semantic role labeling," in Proceedings of CoNLL2004 Shared Task, 2004.

[12] K.-M. Park, Y.-S. Hwang, and H.-C. Rim, "Two-phase semantic role labeling based on support vector machines," in Proceedings of CoNLL, 2004, pp. 126–129.

[13] A. Das, A. Ghosh, and S. Bandyopadhyay, "Semantic role labeling for bengali using 5ws," in International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), 2010. IEEE, 2010, pp. 1–8.

[14] Devadath V V, Litton J Kurisinkel, Dipti Misra Sharma, Vasudeva Varma, "A Sandhi Splitter for Malayalam" in 11th international conference on language processing, Goa University, GOA, 2014.

[15] Nitish Chandra, Sudhakar Kumawat, Vinayak Srivastra (23rd March 2014) Various tagsets for Indian languages and their performance in part 38 of speech tagging, Proceedings of 5th IRF International conference, Chennai,ISBN: 978-93-82702-67-2

[16] Jisha P Jayan, Rajeev R R Part of speech Tagger and Chunker for Malayalam Statistical Approach, Computer Engineering and Intelligent Systems, ISSN 2222-1719 (paper) ISSN 2222-2863 (Online) vol 2,No.3.

[17] R. Jesuraj and P. C. Reghu Raj, "MBLP approach applied to pos tagging in malayalam language," NCILC, 2013.

[18] D. Antony, "Malayalam pos tagger and chunker," CEN Amrita Viswa Vidyapeetham, August 2010.

[19] Rekha Raj C.T, Reghu Raj P.C (july 2015), Text chunker for Malayalam using memory-based learning in IEEE International conference on Control, Communication and Computing India, 2015.

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

[21] L. D. Sobha and S. Lakshmi, "Malayalam clause boundary identifier: Annotation and evaluation," WSSANLP-2013, pp. 83–90.

[22] A. Ghosh, A. Das, and S. Bandyopadhyay, "Clause identification and classification in bengali," in Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP, 23rd International Conference on Computational Linguistics (COLING), 2010, pp. 17–25.

[23] Maaz Anwar Nomani, Dipti Misra Sharma, Towards Building Semantic Role Labeller for Indian Languages, FC Kochi center on Intelligent Systems (KCIS), IIIT - Hrdrabad, India.

[24] S. Lakshmana Pandian1, T.V. Geetha1 (May 2009) Semantic Role Labeling for Tamil Documents, International Journal of Recent Trends in Engineering Vol. 1, No.1.

[25] Walter Daelemans, Jakub Zavrel, Antal van den Bosch, Ko van der Sloot, MBT: Memory-Based Tagger, Induction of Linguistic Knowledge Research Group Department of Communication and Information Sciences Tilburg University , ILK Technical Report ILK 10-04) - June 2, 2010 1988.