International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7829-7834
© Research India Publications. http://www.ripublication.com

WordNet in Malaylam language for Cricket domain

Sreedhi Deleep Kumar
PG Scholar
Department of Computer Science and Engineering
Vidya Academy of Science and Technology, Thrissur, India

Reshma E U
PG Scholar
Department of Computer Science and Engineering
Vidya Academy of Science and Technology, Thrissur, India

Sunitha C
Associate Professor
Department of Computer Science and Engineering,
Vidya Academy of Science and Technology, Thrissur, India

Amal Ganesh
Assistant Professor
Department of Computer Science and Engineering
Vidya Academy of Science and Technology, Thrissur, India

Abstract

WordNet is an information base which is arranged hierarchically in any language. Usually, WordNet is implemented using indexed file system. Good WordNets available in many languages. However, Malayalam is not having an efficient WordNet. WordNet differs from the dictionaries in their organization. WordNet does not give pronunciation, derivation morphology, etymology, usage notes, or pictorial illustrations. WordNet depicts the semantic relation between word senses more transparently and elegantly. In this work, the words or the terms are stored in the XML format. The relationships of the various word are also mentioned. Hence when the user needs to know about a particular term or a word, they can directly search for that word so that all the details regarding the word will be gathered and dispalyed from the WordNet. The informations includes its Synonyms, Meaning of the word, An example statement for the word and also its available relations such as hypernymy and holonymy. The ultimate objective of this project is to create a WordNet in Malayalam language pertaining to cricket domain.

Keywords: WordNet, Indexed file system, Malayalam, Cricket, Synonym, Hypernym, XML

INTRODUCTION

In the area of Natural Language Processing, WordNet plays an important role. WordNet is a semantic dictionary that was designed as a network following the idea that representing words and concepts as an interrelated system is consistent with evidence for the way speakers organize their own mental lexicons[1]. Nowadays , WordNets are available in many Languages. But in Malayalam there is not a good wordnet available yet.

India is a country with diverse culture, language and varied heritage. Due to this, it is very rich in languages and their dialects. Being a multilingual society, a dictionary in multiple languages becomes its need and one of the major resources to support a language. There are dictionaries for many Indian languages, but very few are available in multiple languages. WordNet is one of the most prominent lexical resources in the field of Natural Language Processing.

There are numerous languages in India which belong to different language families. These language families are Indo-Aryan, Dravidian, SinoTibetan, Tibeto-Burman and Austro-Asiatic. The major ones are the Indo-Aryan, spoken by the northern to western part of India and Dravidian, spoken by southern part of India. The Eighth Schedule of the Indian Constitution lists 22 languages, which have been referred to as scheduled languages and given recognition, status and official encouragement.

Malayalam has official language status in the state of Kerala and in the union territories of Lakshadweep and Puducherry. It belongs to the Dravidian family of languages and is spoken by approximately 33 million people, according to the 2001 census. Malayalam most likely originated from Middle Tamil (Sen-Tamil) in the 6th century. WordNets are available in many Languages. But in Malayalam there is not a good WordNet available yet. As a beginning, this project aims to create a WordNet in Malayalam language concentrating on a single domain. This project deals with the Cricket domain.

WordNet is a semantic dictionary that was designed as a network following the idea that representing words and concepts as an interrelated system. WordNet groups words into sets of synonyms and provides short definitions and usage examples, also records a number of relations among these synonym sets or their members.

WordNet can be seen as a combination of dictionary and thesaurus. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms,but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

Properties of WordNet A WordNet can provide the following information:

- **Synonymy:**

This one is easy and links words that have similar meanings, e.g. happy and glad.

- **Antonymy:**

The opposite of synonymy, e.g. happy and sad

- **Hypernymy**:

Hypnernymy refers to a hierarchical relationship between words. For example, furniture is a hypernym of chair since every chair is a piece of furniture (but not vice-versa).

- **Hyponymy**:

Hyponymy is the opposite of hypernymy. Dog is a hyponym of canine since every dog is a canine.

- **Meronymy**:

Meronymy refers to a part/whole relationship. For example, paper is a meronym of book, since paper is a part of a book.

- **Troponymy**:

Troponymy is the semantic relationship of doing something in the manner of something else. For example, walk is a troponym of move and limp is a troponym of walk.

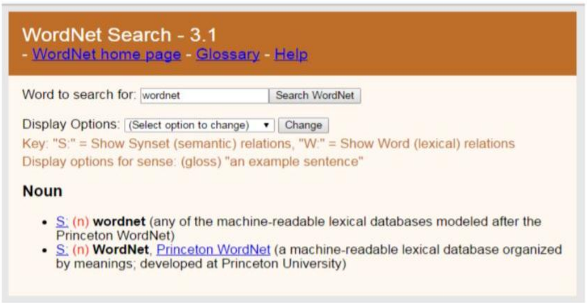
RELATED WORKS

This section enumerate the available WordNets and their properties

I. Princeton WordNet

This is the first WordNet to be developed. WordNet was created in the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George Armitage Miller starting in 1985 and has been directed in recent years by Christiane Fellbaum. WordNet is a lexical database for the English language. It groups English words in to sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. As of November 2012 WordNet's latest Online-version is 3.1.

The database contains 155,287 words organized in 117,659 synsets for a total of 206,941 wordsense pairs; in compressed form, it is about 12 megabytes in size. WordNet includes the lexical categories nouns, verbs, adjectives and adverbs but ignores prepositions, determiners and other function words. Words from the same lexical category that are roughly synonymous are grouped into synsets. Synsets include simplex words as well as collocations like "eat out" and "car pool." The different senses of a polysemous word form are assigned to different synsets. The meaning of a synset is further clarified with a short defining gloss and one or more usage examples. An example adjective synset is: good, right, ripe (most suitable or right for a particular purpose; "a good time to plant tomatoes"; "the right time to act"; "the time is ripe for great sociological changes") All synsets are connected to other synsets by means of semantic relations. These relations, which are not all shared by all lexical categories, include: Nouns. The figure of Princeton WordNet is as shown below.



Word to search for: wordnet

Display Options: (Select option to change)

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **wordnet** (any of the machine-readable lexical databases modeled after the Princeton WordNet)
- S: (n) **WordNet**, Princeton WordNet (a machine-readable lexical database organized by meanings; developed at Princeton University)

Figure 1: Princeton WordNet

II. EuroWordNet

EuroWordNet is a multilingual database with WordNets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The WordNets are structured in the same way as the American WordNet for English (Princeton WordNet, Miller et al 1990) in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each WordNet represents a unique language-internal system of lexicalizations. In addition, the WordNets are linked to an Inter-Lingual-Index, based on the Princeton WordNet. Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained in the individual WordNets. The database can be used, among others, for monolingual and cross-lingual information retrieval, which was demonstrated by the users in the project.

The EuroWordNet project was completed in the summer of 1999. The design of the database, the defined relations, the top-ontology and the Inter-Lingual-Index are now frozen. Nevertheless, many other institutes and research groups are developing similar WordNets in other languages (European and non-European) using the EuroWordNet specification. If compatible, these WordNets can be added to the above database and, via the index, connected to any other WordNet. The EuroWordNet format is defined by the EuroWordNet Database Editor Polaris.

III. ***IndoWordNet***

IndoWordNet is an integrated multilingual WordNet for Indian languages. These WordNet resources are used by researchers to experiment and resolve the issues in multilinguality through computation. However, there are few cases where WordNet is used by the non-researchers or general public. IndoWordNet is a linked lexical knowledge base of wordnets of 18 scheduled languages of India, viz., Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Meitei (Manipuri), Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu and Urdu. IndoWordNet Dictionary or IWN Dictionary is an online interface to render multilingual IndoWordNet information in the dictionary format. It allows user to view the results in multiple formats as per the need. Also, user can view the result in multiple languages simultaneously.

The look and feel of the IWN Dictionary is kept same as a traditional dictionary keeping in mind the user adaptability. So far, it renders WordNet information of 19 Indian languages. Dictionary words are included in the WordNet according to the frequency of their use. Transliteration, Short phrase, Coined word are typically needed in expanding from a culture or region specific concept. However, these options should be used with discretion, respecting the native speakers sensitivities. The Indo WordNet uses linked structure for storing the data. The IndoWordNet is as shown below.

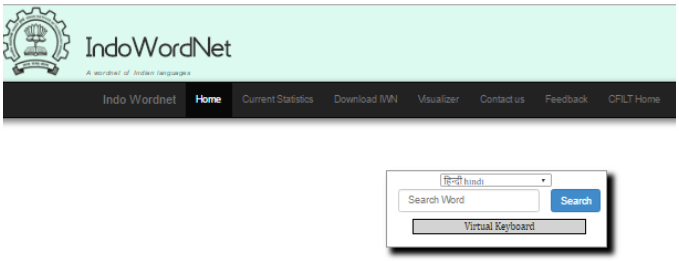


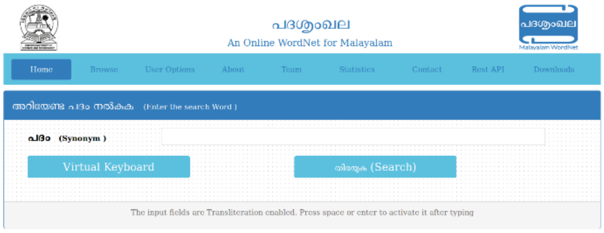
Figure: IndoWordNet

IV. Padasringala (Malayalam WordNet)

Malayalam WordNet is a component of Dravidian WordNet which in turn is the component of IndoWordNet. Malayalam WordNet is an online lexical database. Malayalam WordNet aims to capture the net work of lexical or semantic relations between lexical items or words in Malayalam. Malayalam WordNet is an crowd sourced project. IndoWordNet is publicly browsable, but it is not available to edit.

Malayalam WordNet allows users to add data to the WordNet in an controlled crowd sourcing manner. Either a set of experts or users itself could review the entries added by other members which helps in maintaining consistent data throughout. It also has a JSON and XML interfaces which helps the programmers to interact with the WordNet. It would be highly useful for the researchers, language experts as well as application developers.

The figure is as shown below:



The input fields are Transliteration enabled. Press space or enter to activate it after typing

Figure: Padasringala

COMPARISON TABLE

The comparison table of different available WordNets are given below

Table 1. Comparison of Methods

Sl.No	**Review of the existing WordNets**		
	WordNets	**Language**	
1	Princeton WordNet	English	The database contains 155,287 words organized in 117,659 synsets for a total of 206,941 word sense pairs; in compressed form, it is about 12 megabytes in size.
2.	EuroWordNet	European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian)	The wordnets are linked to an Inter-Lingual-Index, based on the Princeton wordnet. The languages are interconnected.
3.	IndoWordNet	Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Meitei (Manipuri), Marathi, Nepali, Odia, Punjabi, Sanskrit, Tamil, Telugu, Urdu	Dictionary words are included in the wordnet according to the frequency of their use.
4.	Malayalam WordNet	Malayalam	It has a JSON and XML interfaces which helps the programmers to interact with the WordNet.

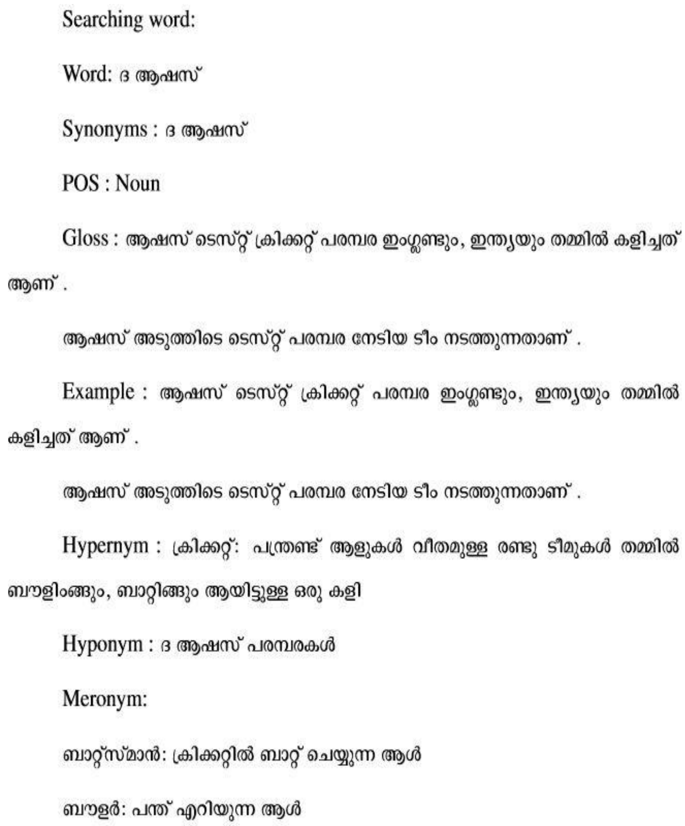
IMPLEMENTATION

To develop a WordNet which contains the words and the terms that are related to the Cricket domain. The implementation of WordNet in Malyalam is a complex task. As the first part, this project aimed to build WordNet for a particular domain. In this project, the Wordnet in Malayalam language for Cricket domain is implemented succesfully. Including more words in the dataset and also integrating with the other domains increase the accuracy and the relevance of the WordNet.

Gives the following information:

- Meaning of the word.
- An example sentence for the word.
- Relationships of each words like synonyms antonyms holonyms, hyponyms.

The example is as follows:



Searching word:

Word: ദ ആഷസ്

Synonyms : ദ ആഷസ്

POS : Noun

Gloss : ആഷസ് ടെസ്റ്റ് ക്രിക്കറ്റ് പരമ്പര ഇംഗ്ലണ്ടും, ഇന്ത്യയും തമ്മിൽ കളിച്ചത് ആണ് .

ആഷസ് അടുത്തിടെ ടെസ്റ്റ് പരമ്പര നേടിയ ടീം നടത്തുന്നതാണ് .

Example : ആഷസ് ടെസ്റ്റ് ക്രിക്കറ്റ് പരമ്പര ഇംഗ്ലണ്ടും, ഇന്ത്യയും തമ്മിൽ കളിച്ചത് ആണ് .

ആഷസ് അടുത്തിടെ ടെസ്റ്റ് പരമ്പര നേടിയ ടീം നടത്തുന്നതാണ് .

Hypernym : ക്രിക്കറ്റ്: പന്ത്രണ്ട് ആളുകൾ വീതമുള്ള രണ്ടു ടീമുകൾ തമ്മിൽ ബൗളിംങ്ങും, ബാറ്റിങ്ങും ആയിട്ടുള്ള ഒരു കളി

Hyponym : ദ ആഷസ് പരമ്പരകൾ

Meronym:

ബാറ്റ്സ്മാൻ: ക്രിക്കറ്റിൽ ബാറ്റ് ചെയ്യുന്ന ആൾ

ബൗളർ: പന്ത് എറിയുന്ന ആൾ

DESIGN

The design for the implementation is given below. It shows the available relations of the WordNet

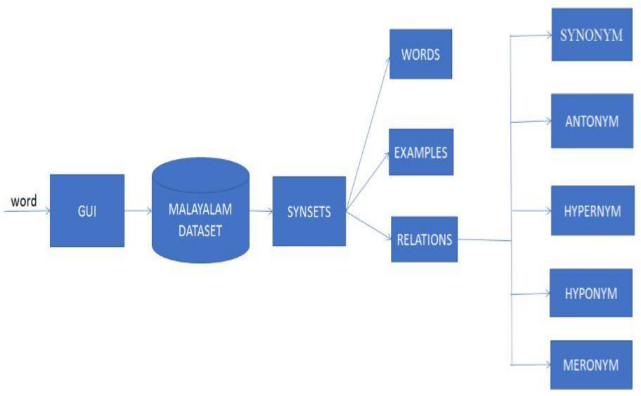


Figure: Design

WORKING

At first, the user will enter the desired word that needs to be searched. We have created a Malayalam WordNet for cricket domain and it consists of all the words that are related to the cricket field. So the user will be searching for a cricket related word. As the WordNet reveals all the details of the words as mentioned earlier, it displays every bit of information it contains.

Each word in the WordNet is saved as a synset. Each word will have its own Id. The synonym set of the word will have a separate id and they are having many other applications like word comparison etc. The synonym set id is called the synset id. So, using the Id's, we can compare the words. Eg: IF in two sentence in one of the document, there are synonyms of a particular word, we can match the two words.

Software Requirements

- Python

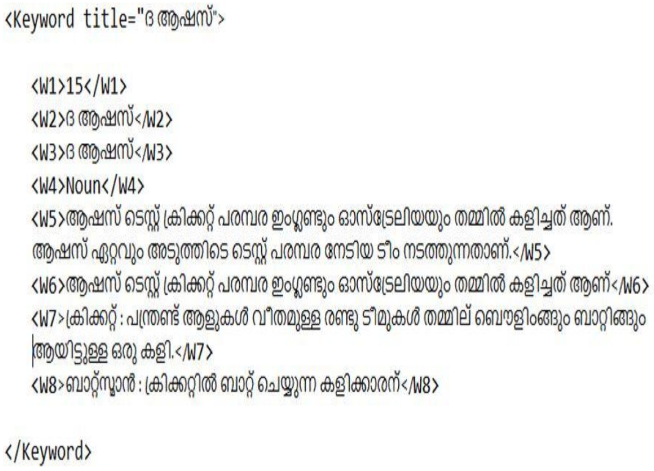
Python is a multi-paradigm programming language: object-oriented programming and structured programming are fully supported, and many language features support functional programming and aspect-oriented programming. Many other paradigms are supported via extensions, including design by contract and logic programming. Python uses dynamic typing and a mix of reference counting and a cycle detecting garbage collector for memory management. An important feature of Python is dynamic name resolution(late binding), which binds method and variable names during program execution. The design of Python offers some support for functional programming in the Lisp tradition. The language has map(), reduce() and filter() functions; list comprehensions, dictionaries, and sets; and generator expressions. The standard library has two modules (itertools and functools) that implement functional tools borrowed from Haskell and Standard ML.

The coding is done in Python language.

- XML

In computing, Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The design goals of XML emphasize simplicity, generality, and usability across the Internet. It is a textual data format with strong support via Unicode for different human languages. XML stores data in plain text format. This provides a software-and hardware independent way of storing, transporting, and sharing data. XML also makes it easier to expand or upgrade to new operating systems, new applications, or new browsers, without losing data.

With XML, data can be available to all kinds of ”reading machines” like people, computers, voice machines, news feeds, etc.In this project, XML is used for storing the data. It is stored in the format shown in the figure.



```
<Keyword title="ദ ആഷസ്">

    <W1>15</W1>
    <W2>ദ ആഷസ്</W2>
    <W3>ദ ആഷസ്</W3>
    <W4>Noun</W4>
    <W5>ആഷസ് ടെസ്റ്റ് ക്രിക്കറ്റ് പരമ്പര ഇംഗ്ലണ്ടും ഓസ്ട്രേലിയയും തമ്മിൽ കളിച്ചത് ആണ്.
    ആഷസ് ഏറ്റവും അടുത്തിടെ ടെസ്റ്റ് പരമ്പര നേടിയ ടീം നടത്തുന്നതാണ്.</W5>
    <W6>ആഷസ് ടെസ്റ്റ് ക്രിക്കറ്റ് പരമ്പര ഇംഗ്ലണ്ടും ഓസ്ട്രേലിയയും തമ്മിൽ കളിച്ചത് ആണ്</W6>
    <W7>ക്രിക്കറ്റ് : പന്ത്രണ്ട് ആളുകൾ വീതമുള്ള രണ്ടു ടീമുകൾ തമ്മില് ബൌളിംങ്ങും ബാറ്റിങ്ങും
    ആയിട്ടുള്ള ഒരു കളി.</W7>
    <W8>ബാറ്റ്സ്മാൻ : ക്രിക്കറ്റിൽ ബാറ്റ് ചെയ്യുന്ന കളിക്കാരന്</W8>

</Keyword>
```

Figure : Dataset stored in XML format

Here, in the dataset stored, W1 tag gives the Synset ID or the word ID. W2 gives the word itself. W3 gives the set of synonyms that are available for the word. W4 gives the tag set or the part of speech tag for the word. W5 gives the gloss or the meaning the word trying to portray. W6 gives an example statement related to the word so that the word will get more familiarized. W7 gives the hyponym of the word if any. W8 gives the hypernym of the word if any.

• PyCharm

PyCharm is an Integrated Development Environment (IDE) used in computer programming, specifically for the Python language. It is developed by the Czech company JetBrains. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django. Features includes the following:

- Project and code navigation: specialized project views, file structure views and quick jumping between files, classes, methods and usages
- Coding assistance and analysis, code completion, syntax and error highlighting, linter integration, and quick fixes
- Python refactoring
- Support for web frameworks: Django, web2py and Flask
- Integrated Python debugger
- Integrated unit testing, with line-by-line code coverage
- Google App Engine Python development
- Version control integration: unified user interface for Mercurial, Git, Subversion, Perforce and CVS with changelists and merge

RESULTS

This work includes the words that are related to the cricket domain and through this WordNet, the users will be able to be more familiarized with the terms that are related to Cricket.

This gives not only the meanings, but also the relationships of them. Providing the examples for the different terms will enable the users to understand the terms in a better manner.

The XML format is used to store the dataset.

There are two sections in the output results.

• **Searching** :

In searching, the user can search for the word by entering the word in the space provided.

• **Displaying** :

In the displaying, the details of the word that the user searched will be displayed. This includes the Synonyms, Meanings, Example statement, Hypernym and Hyponym of the word that is searched.

Screenshots

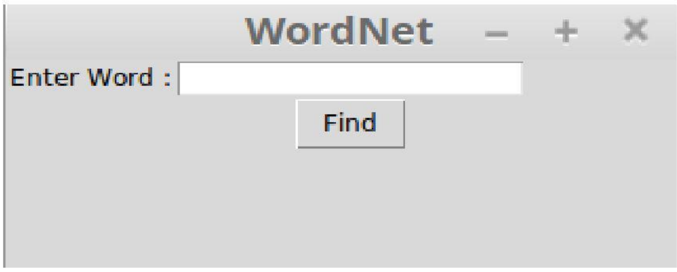
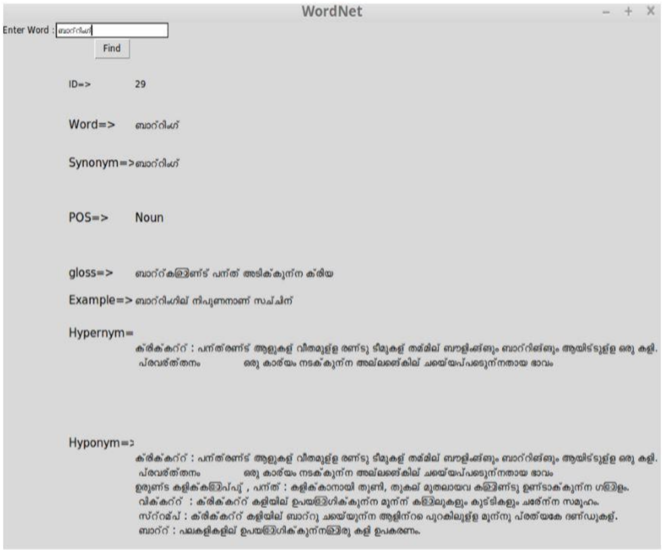


Figure : Search module



Enter Word : ബാറ്റിംഗ്

ID=> 29

Word=> ബാറ്റിംഗ്

Synonym=>ബാറ്റിംഗ്

POS=> Noun

gloss=> ബാറ്റ്കൊണ്ട് പന്ത് അടിക്കുന്ന ക്രിയ

Example=> ബാറ്റിംഗില് നിപുണനാണ് സച്ചിന്

Hypernym=

ക്രിക്കറ്റ് : പന്ത്രണ്ട് ആളുകള് വീതമുള്ള രണ്ടു ടീമുകള് തമ്മില് ബൌളിങ്ങും ബാറ്റിങ്ങും ആയിട്ടുള്ള ഒരു കളി.
പ്രവര്ത്തനം   ഒരു കാര്യം നടക്കുന്ന അല്ലെങ്കില് ചെയ്യപ്പെടുന്നതായ ഭാവം

Hyponym=>

ക്രിക്കറ്റ് : പന്ത്രണ്ട് ആളുകള് വീതമുള്ള രണ്ടു ടീമുകള് തമ്മില് ബൌളിങ്ങും ബാറ്റിങ്ങും ആയിട്ടുള്ള ഒരു കളി.
പ്രവര്ത്തനം   ഒരു കാര്യം നടക്കുന്ന അല്ലെങ്കില് ചെയ്യപ്പെടുന്നതായ ഭാവം
ഉരുണ്ട കളിക്കോപ്പ് , പന്ത് : കളിക്കാനായി തുണി, തുകല് മുതലായവ കൊണ്ടു ഉണ്ടാക്കുന്ന ഗോളം.
വിക്കറ്റ് : ക്രിക്കറ്റ് കളിയില് ഉപയോഗിക്കുന്ന മൂന്ന് കോലുകളും കുട്ടികളും ചേര്ന്ന സമൂഹം.
സ്റ്റമ്പ് : ക്രിക്കറ്റ് കളിയില് ബാറ്റു ചെയ്യുന്ന ആളിന്റെ പുറകിലുള്ള മൂന്നു പ്രത്യേക ദണ്ഡുകള്.
ബാറ്റ് : പലകളികളില് ഉപയോഗിക്കുന്നൊരു കളി ഉപകരണം.

Figure : Display module

APPLICATIONS

Some of the applications are as follows:

- Information retrieval and extraction
- Document categorization
- Language teaching and translation applications
- Machine Translation
- Automatic text summarization
- Word Sense Disambiguation
- Semantic Parsing
- Opinion Mining

FUTUREWORK

The system needed more word adding and voluntary work from those who is having deep knowledge in Malayalam vocabulary. The system can grow with the contributions from learned enthusiasts of Malayalam. It is a step towards an efficient WordNet for Malayalam language. As a begining, this work focus on a particular domain, so that the maximum possibilities of the domain can be included in the WordNet. As the future work, more words can be included in the dataset and also by taking the different other domains.

CONCLUSION

Malayalam is a complex Dravidian language mostly spoken in South India. The implementation of WordNet in Malyalam is a complex task. As the first part, this project aimed to build WordNet for a particular domain. In this project, the Wordnet in Malayalam language for Cricket domain is implemented successfully. Including more words in the dataset and also integrating with the other domains increase the accuracy and the relevance of the WordNet.

REFERENCES

[1] Banu M, Karthika C, Sudarmani P and Geethu T V"*Tamil document summarization using semantic graph method"* International Conference on Computational Intelligence and Multimedia applications. IEEE -2007

[2] Manjula Subramaniam, Prof. Vipul Dalal" *Test Model for Rich Semantic Graph Representation for Hindi Text using Abstractive Method"* (IRJET) Volume: 02 Issue: 02 | May-2015

[3] Jayashree. R , Srikanta Murthy .K and Sunny .K, "*Keyword Extraction Based Summarization of Categorized Kannaad Text Documents"*, International Journal on Soft Computing (IJSC) Vol.2, No.4, November 2011.

[4] Dr.M.Humera Khanam1 , S.Sravani" *Text Summarization for Telugu Document*" IOSR Journal of Computer Engineering (IOSR-JCE) Volume 18, Issue 6, Ver. 2016, PP 25-28

[5] Vishal Gupta1 and Gurpreet Singh Lehal" *Preprocessing Phase of Punjabi Language Text Summarizatio*n"Springer-Verlag Berlin Heidelberg 2011

[6] R. Kabeer, Sumam M I, *Text Summarization of Malayalam Documents- an Experience*, International Conference on Data Science and Engineering(ICDSE), 2014.

[7] Dr. A Jaya, Sunitha C, Amal Ganesh "*Abstractive Summarization Techniques in Indian Languages*", Peer review under responsibility of the Organizing Committee of ICRTCSE 2016 doi: 10.1016/j.procs.2016.05.121, International Conference of recent trends in computer science

[8] Mostafa Aref, Ibrahim Moawad, Soha Ibrahim., "*Rich Semantic Graph Generation System Prototype,* The tenth Conference on Language Engineering, Egypt, 2010

[9] C. Thaokar and L. Malik, "*Test model for summarize hindi text by extraction method*," in Information & Communication Technologies (ICT), 2013 IEEE Conference on, pp. 1138–1143, IEEE, 2013.

[10] M. Aref, I. F. Moawad "s*emantic graph reduction approach for abstractive text summarization*," In Computer Engineering and Systems (ICCES), 2012 Seventh International Conference on, pp. 132–138, IEEE, 2012