Kannada Word Sense Disambiguation for Machine Translation

S. Parameswarappa
Dept. of Computer Science & Engg.
Malnad College of Engg. Hassan,India
Visvesvaraya Technological University

V.N.Narayana
Dept of Computer Science & Engg.
Malnad College of Engg. Hassan,India
Visvesvaraya Technological University

ABSTRACT
Polysemous Words can have more than one distinct meaning. Word sense disambiguation (WSD) is the ability to identify the exact meaning of such polysemous words in context in a computational manner. WSD is considered as an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problem in Artificial Intelligence. In this paper, we propose an Integrated Kannada Word Sense Disambiguation system which includes a suite of high performance Natural Language Processing (NLP) modules implemented in Perl (Program Extraction and Reporting Language) to carry out word sense disambiguation task. The corpus builder module will construct the raw Kannada corpora using web. The proposed system uses randomly selected sentences from the corpora as a test bed for disambiguation. The electronic machine readable dictionary is built by Dictionary builder module using the corpora. The Target Word Sense Disambiguation module will disambiguate the potential ambiguous target words in a sentence. The polysemous verb in a sentence is disambiguated by Verb Sense Disambiguation module. The rule based disambiguator will disambiguate all ambiguous words with different lexical category. Experiments conducted and the results obtained have been described. The efficiency of the system proved to be reliable and extendable.

General Terms
Word Sense Disambiguation, Machine Translation, Natural Language Processing, Artificial Intelligence, Corpus Linguistics, Lexicography.

Keywords
Kannada Word Sense Disambiguation, Kannada Corpus, Kannada machine readable dictionary, Target Word, Verb Sense Disambiguation, Verbalizer.

1. INTRODUCTION
Indian languages come from four different language families - the Indo-Aryan, The Tibeto-Burman, The Austro-Asiatic and the Dravidian. Kannada language belongs to Dravidian family [1]. Kannada is one of the technologically least developed languages in India today. This is ironical since Kannada has a very old and rich literary tradition, it is currently spoken by 60 Million people and Karnataka is in the centre stage of IT (Information Technology) revolution in the country. As of today, the only corpus we have is, roughly 3 Million word corpus developed by CIIL (Central Institute of Indian Languages) Mysore long ago. Lack of basic resources such as corpora is one of the major reasons for our lagging behind in language technology. Several languages in India today have 30 to 50 Million word corpora. There are hardly any electronic dictionaries, morphological analyzers, POS taggers and Computational Grammars or Parsing systems for Kannada worth taking seriously. Naturally, we are lagging behind in many areas of linguistics as also in language technologies [2]. To address the issue, the proposed tool is a milestone in Kannada language technology development. As a contributory work, the tool builds Kannada raw corpora, machine readable dictionary and provides the solution for solving Kannada WSD problem during Machine Translation.

The world languages are classified in to two categories. Namely, fixed word order and free word order. In the former case, the words constituting a sentence can be positioned in a sentence according to grammatical rules in some standard ways. On the other hand, in the later case, no fixed ordering is imposed on the sequence of words in a sentence. An example for fixed word order language is English and that of pure free word order language is Sanskrit [3]. Kannada is a relatively free word order language. Because, it is an agglutinating language of the suffixing type. Nouns are marked for number and case and verbs are marked, in most cases, for agreement with the subject in number, gender and person. Kannada language exhibits a very rich system of morphology. Morphology includes inflection, derivation, conflation (sandhi) and compounding [4].

The study conducted on a Kannada dictionary with around 50000 entries developed by us, reveals the fact that irrespective of lexical categories many words have more than one meaning. As an example consider a word ನೆರೆ [nere], it has five meanings such as ಪ್ರಾಯಕ್ಕೆ ಬರು [praayakke baru] 'Biologically mature', ಪ್ರವಾಹ [pravaaha] 'flood', ಗುಂಪು ಸೇರು [gumpu seeru] 'gather', ಅಕ್ಕಪಕ್ಕ [akkapakka] 'neighbor', ಕೂದಲು ಬೆಳ್ಳಗಾಗು [kuudalu beLLagaagu] 'white hair'. The constructed corpora contain 2153 occurrences of the word ನೆರೆ [nere]. The target word sense disambiguation module will assign the correct meaning for potential ambiguous words like ನೆರೆ[nere]. Among lexical categories the verbs seems to exhibit high ratio of semantic ambiguity than other categories. We found 314 ambiguous verbs out off 2202 verbs in a dictionary. The high score of polysemy with verbs is an indication of how important verbs are in developing natural language applications. Frequently used verbs

in Kannada ತಿನ್ನು [tinnu] 'to eat', ಹೋಗು [hoogu] 'to go', ಮಾಡು [maaDu] 'to do' are also most polysemous. Some of them function as verbalizers when used with nouns. The verb sense disambiguation module will disambiguate the polysemous verbs in a sentence. Likewise, consider a word ಮಧುರ [madhura], it has two meanings such as name of a person or melodious belongs to two lexical categories noun and adjective respectively. The rule based disambiguation module will disambiguate the words of this kind. In order to achieve high quality translation output in machine translation, word sense disambiguation is one of the most important problems to be solved. This is the motivation behind the present work. The following example illustrates the need of WSD in machine translation. The English translation for the Kannada sentence ನೆರೆಯಲ್ಲಿ ನೆರೆಹೊರೆಯವರೆಲ್ಲಾ ಕೊಚ್ಚಿ ಹೋದರು by online google translator is 'ನೆರೆಹೊರೆಯವರೆಲ್ಲಾ neighborhood went to Kochi'. It is wrong. The correct translation is, 'neighbors are washed away in the flood'. The error in the translation is due to incorrect assignment of meaning to polysemous words in a sentence. In the above sentence, the words ನೆರೆ [nere], ಕೊಚ್ಚಿ [kocci] are all polysemous. The ನೆರೆ [nere] has five meanings as depicted above. Likewise, the ಕೊಚ್ಚಿ [kocci] has two senses as 'name of the place' (noun) and as 'washed away' (verb). The translator assigns wrong senses for both the words. Hence, the output is wrong. Given polysemous words and their possible senses, as defined in a knowledge base, the WSD can be defined as the task of assigning the most appropriate sense to the word within a given context. The WSD is necessary not only in Machine Translation but also in almost every application of language technology including information retrieval or extraction, knowledge mining or acquisition, lexicography, semantic interpretation etc.

The rest of the paper is organized as follows. Section 2 explores the previous literature. Section 3 describes the proposed system architecture for Kannada word sense disambiguation. Section 4 discusses the evaluation of the system. Section 5 concludes the paper.

2. LITERATURE REVIEW

Based on how the disambiguation information is acquired by the WSD system, they are classified as knowledge-based, corpus-based and hybrid systems [5].

Knowledge-based approaches encompass systems that rely on information from an explicit lexicon such as Machine Readable Dictionaries, thesauri, computational lexicons such as Wordnet [6] or hand crafted knowledge bases. Knowledge based approaches to WSD such as Lesk's algorithm [7], Walker's algorithm [8], Conceptual density [9] and random walk algorithm [10] essentially do machine readable dictionary look up. However, these are fundamentally overlap based algorithms which suffer from overlap scarcity, dictionary definitions being generally small in length.

Corpus-based methods are further classified in to supervised, semi-supervised and unsupervised methods. Supervised and semi-supervised methods make use of annotated corpora to train from or as seed data in a bootstrapping process. They are mostly word specific classifiers. Some of the examples for supervised learning algorithms are WSD using SVM [11], Exemplar based WSD [12] and decision list based algorithm [13]. An example for semi-supervised decision list algorithm is [14]. Unsupervised algorithms work directly from un-annotated raw corpora. They have the potential to overcome the new knowledge acquisition bottleneck and they have achieved good result. [15, 16] are some of the examples of unsupervised approaches.

Hybrid approaches like WSD using Structural Semantic Interconnections [17] use combination of more than one knowledge sources such as wordnet as well as a small amount of tagged corpora. This allows them to capture important information encoded in wordnet as well as draw syntactic generalization from minimally tagged corpora. These methods, which combines evidence from several resources seem to be most suitable in building all words disambiguation tool and are the motivation for our work.

Many researchers focused on disambiguation of selected target words. [18, 19] are some of the examples for target word sense disambiguation. Instead of disambiguating all ambiguous words in sentence in a single go, they tried to disambiguate only selected target word using hierarchical information in a wordnet and syntactic features. The mechanism for representing Telugu complex predicates in wordnet was proposed by [20]. The compound words are kind of complex predicates. The present work uses the compound words clue to disambiguate the target word. The target word sense disambiguation module is proposed based on the work of [21].

The study conducted on Kannada dictionary reveals the fact that, the verbs exhibits high ratio of ambiguity than other lexical categories. It motivates us to introduce separate verb sense disambiguation module in the present tool. This module is built based on [22] - [29] works.

Earlier days WSD systems are all rule governed. They used set of hand crafted rules for disambiguation tasks. Even though, building rules manually is difficult, labor intensive and time consuming but the performance of the system is extremely good. The statistical analysis conducted by us on Kannada corpora reveals the fact that the degree and nature of word ambiguity in Kannada is systematic and only portion of words in corpora are ambiguous, it can be resolved at maximum extent by building disambiguation rules. This is the basic motivating factor for us to introduce Rule based disambiguation module in the present tool. The module is built based on the work of [3] and [30] - [32].

From [33] – [36] we deduced the fact that a large and representative corpus has many uses and applications. What all we can do with a corpus is limited by our imagination and creativity. A corpus forms the very basic of all language and linguistic studies. But, unavailability of such useful resource for Kannada in large scale is a bottleneck for language research. It motivates us to build Corpus building module in the present tool. The module is built based on the work of [37] - [39].

Like corpora, equally important resource for language research and development is electronic machine readable dictionary. Even till date, there is no such tool worth mentioning for Kannada Language. This fact motivates us to introduce Dictionary building module in the present tool, we constructed the electronic machine readable dictionary for Kannada language based on [40] work.

3. SYSTEM ARCHITECTURE

Figure 1 shows the proposed architecture for Kannada Word Sense Disambiguation. The system is modular. In a computational frame work, it is not enough to have a modular design, the interaction between the modules and the exact flow of data from module to module needs to be worked out too. Here the main data flow is in a pipe line, there is no need to go back and forth, data flows in a linear fashion and each module adds its contribution.

The architecture consists of following modules based on their functionality. Namely, Corpus Builder, Sentence Extractor, Dictionary Builder, Kannada Shallow Parser, Word Classifier, KTWSD (Kannada Target Word Sense Disambiguator), KVSD (Kannada Verb Sense Disambiguator), KRBWSD (Kannada Rule Based Word Sense Disambiguator).

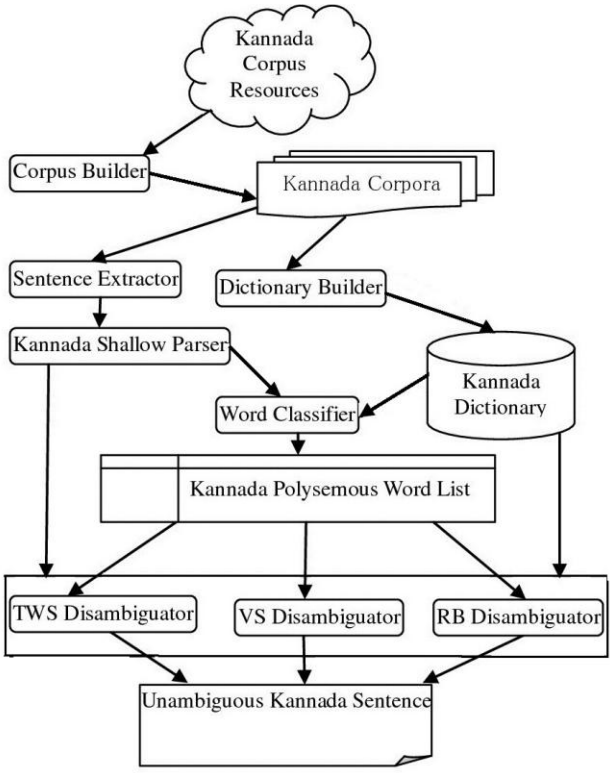

Fig 1: The system architecture.

The functionality of each of these modules is explained briefly as follows.

3.1 Corpus builder

A corpus is a collection of documents in electronic, computer processable form [41]. Open, freely and publicly available corpora can be used by all researchers as standard data sets to develop and test their systems.

Corpus builder module is a set of Perl programs implementing an iterative procedure to build Kannada corpora from the web.

The procedure requires is, first a set of "seed" words list is built and later a set of “seed” URLs (Uniform Resource Locator) containing documents in the Kannada language is collected by sending queries to commercial search engines (Google and Yahoo). The obtained seeds are then used to start a crawling job using the open-source, command-line based downloading tool "wget". The downloaded documents are then processed in various ways in order to build Kannada raw corpora such as HTML (Hyper Text Markup Language) code removal, boilerplate stripping, and language identification, duplicate and near duplicate detection. We conducted an evaluation of the module by using it for constructing Kannada corpora from the Kannada corpus resources.

3.1.1 Kannada corpus resources

Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. As of July 2011, it has 10,917 articles in Kannada, have been written collaboratively by volunteers around the world. Almost all of its articles can be edited by anyone with access to the site [42].

Wiki dictionary is a Wikipedia's sister project and is hosted by the Wikimedia Foundation, a non-profit organization. Kannada language Wiki dictionary is a collaborative project by them to produce a free-content multilingual dictionary. It aims to describe all words of all languages using definitions and descriptions. For Kannada-Kannada-English and Kannada-English, there are 115,120 words available in wiki dictionary itself as on July 2011 [43].

Sampada is a community of people passionate about literary activities in Kannada and is one of the largest Kannada communities on the Internet. Recent Discussions, Blogs, Articles, Recent Activities, Proverbs, Recent Feedback's, Poems and Fifteen Books and Novels are the different categories of the corpora available in Sampada Kannada Community [44].

Web blogs provide useful corpora in many domains. Some of the common blog sites in Kannada are Kannada Bloggers, Wordpress, BlogSpot, Ekanasu, Sampada and Indiblogger [45].

Prajavani (Kannada for Voice of the People) is a leading Kannada language newspaper in Karnataka. It is a sister publication of the Deccan Herald. As of 2011, it had a circulation of more than 600,000, making it the second-largest-circulation newspaper in Karnataka after The Hindu, and the largest-circulation Kannada language newspaper in the state [46].

Figure 2 shows front page of Prajavani, a Kannada daily news paper. It acts as an input for Corpus builder module.

Fig 2: Kannada Corpus Resource.

Figure 3 shows the partial Kannada raw corpus extracted from web resources using Corpus builder module. It is an output generated by a Corpus builder.

Kannada Raw Corpus
ಉಪ್ಪು ನೀರು... ಹಾರು ಬೂದಿ... ಕೈ ಕೋಡುವ ವಿದ್ಯುತ್
Kannada Transliteration
[uppuniiru... haaru buudi... kai koDuva vidyut]
English Translation
'Salt water... flying ash... fob off current'

Fig 3: Partial Kannada Raw Corpus by Corpus builder.

Figure 3 shows only partial corpus built by corpus builder just to illustrate the corpus building process. We constructed around five million word corpora using corpus builder module. In the subsequent sections we will use some of the example sentences not available in Figure 3 to illustrate the concepts.

3.2 Sentence extractor

The input for sentence extractor module is Kannada raw corpora. It extracts randomly selected sentences from the raw corpora required for disambiguation task. Some of the sentences extracted from corpora using sentence extractor are shown in Figure 4.

ನೆರೆಯಲ್ಲಿ ನೆರೆಹೊರೆಯವರೆಲ್ಲಾ ಕೊಚ್ಚೆ ಹೋದರು
ಪಾಕಿಸ್ತಾನ ನೆರೆ ಸಂತ್ರಸ್ಥರಿಗೆ ನೆರವು
ಬಟ್ಟೆ ಒಣಗಿದೆ
ಅವನ ಗಾಯ ಒಣಗಿದೆ
ಉದಯವಾಗಲಿ ನಮ್ಮ ಚೆಲುವ ಕನ್ನಡ ನಾಡು
ಹಾರು ಬೂದಿ

Fig 4: Test sentences extracted by sentence extractor.

Figure 5 shows the Kannada Transliteration and English translation of the sentences depicted in Figure 4.

Kannada Transliteration
[nereyalli nerehoreyavarella kocci hoodaru]
[paakistaana nere santhrastharige neravu]
[baTTegaLu oNagive]
[avana gaaya oNagide]
[udayavaagali namma celuva kannada naaDu]
[Haaru buudi]
[antaha halkaa kelasa maaDabeeDa]
[varSha nanna tangi]
[siite neredaru]
[raajabaag savaarana urusu khyaatavaagide]
English Translation
'Neigbors are washed away in a flood'
'Help for Pakistan flood victims’
'Cloth is dried'
'His wound has healed'
'May our beautiful Kannada state arise'
'Fly ash'
'Don’t do such cheap work'
'Varsha is my sister'
'Seethe matured biologically'
'Raajabaag savaara's fair is popular'

Fig 5: Transliteration & Translation for test sentences.

3.3 Dictionary builder

Knowledge of language is essential for meaningful communication through language. Words of a language and the phonological, morphological, syntactic and semantic information associated with them, forms a very important part of the knowledge of language. Knowing the words is an extremely important part of knowing a language. Dictionaries are storehouse of such information and therefore they have key role to play in Natural Language Processing (NLP).

We created a Kannada electronic dictionary containing around 50000 entries for our work using dictionary builder module. Each entry is on separate line. Each entry starts with the head word followed by tags separated by double vertical lines. Additional information fields, if any come at the end of each tag separated by double colons. Comments come at the end after hash. The dictionary is a single plain text file, amenable for manipulation through basic commands and tools such as grep, awk and sed. It is easy to write Perl scripts too. Internally, the dictionary will reside in an indexed m-way balanced tree structure.

3.4 Kannada shallow parser

The morphological and syntactic information of a given input sentence are obtained by using freely available Kannada Shallow Parser [47]. It is a shallow parser designed based on computational Paninian model [48]. It parses the given sentence at surface level and produces eight stages of intermediate outputs. The morphological information of each word in a sentence and useful surface level syntactic information of a sentence for disambiguation task are extracted using this module.

3.5 Word classifier

The word classifier identifies all monosemous and polysemous words in a given input sentence and creates a polysemous word list.

Table 1 shows the Kannada polysemous words list generated by word classifier for the example sentences shown in Figure 4.

Table 1. Kannada polysemous word list.

Words	Meanings
ನೆರೆ [nere]	Noun = {Flood, neighbor} Verb = {gather, biologically mature} Adjective = {white hair}
ಕೊಚ್ಚಿ [kocci]	Proper noun = { Name of a place} Verb/Adverb = { washed away, cut }
ಒಣಗಿದೆ [oNagide]	Verb = {dried, healed}
ಚೆಲುವ [celuva]	Proper noun = { Name of a person } Adjective = { beautiful }
ಹಾರು [haaru]	Verb = {jump} Verb/Adjective = {fly}
ಹಲ್ಕಾ [halkaa]	Proper Noun = { Name of a person } Adjective = { cheap }
ವರ್ಷ [varSha]	Proper noun = { Name of a person } Common noun = { Year }
ಸವಾರ [savaara]	Proper noun = { Name of a person } Common noun = { Rider }

3.6 KTWSD module

Disambiguating one target word in a sentence is called as Target Word Sense Disambiguation. In this case, the WSD is viewed as a typical classification problem. It uses machine learning techniques to train the system.

Consider an example ಪಾಕಿಸ್ತಾನ ನೆರೆ ಸಂತ್ರಸ್ಥರಿಗೆ ನೆರವು . It has ambiguous target word ನೆರೆ [nere]. The word ನೆರೆ [nere] has five distinct meanings as depicted in Table 1. In an example sentence, the correct meaning of ನೆರೆ [nere] is 'flood'. Assigning 'flood' meaning to a ನೆರೆ [nere] is called as Kannada Target Word Sense Disambiguation. The KTWSD module will execute the task.

This module uses the compound words clue and syntactic features in a local context for Kannada Target Word Sense Disambiguation. It is noticed that, the use of syntax will improve the performance of the WSD system. The module uses Kannada Shallow parser for syntactic analysis. The ambiguous target word is disambiguated using supervised learning techniques. The module works based on Naive Bayes classifier. The input for the module is ambiguous target word extracted from Kannada Polysemous Word List. The system uses the potential Kannada ambiguous target word list look up to select the target word. The module disambiguates the target word and returns the result. The following sentence is an output generated by the module for the illustrated example.

ಪಾಕಿಸ್ತಾನ ನೆರೆ ಸಂತ್ರಸ್ಥರಿಗೆ ನೆರವು. 'Help for Pakistan flood victims'

3.7 KVSD module

Kannada verbs exhibits high ratio of semantic ambiguity than other lexical categories of the language such as noun, pronoun, adjective etc. The study conducted by us on a dictionary reveal the fact that out of 2202 verbs in a dictionary 314 verbs exhibits ambiguity. The following examples illustrate the kind of ambiguity introduced by verbs.

ಬಟ್ಟೆ ಒಣಗಿದೆ. [baTTegaLu oNagive] 'Cloth is dried'

ಅವನ ಗಾಯ ಒಣಗಿದೆ . [avana gaaya oNagide] 'His wound has healed'.

In the above example sentences the verb ಒಣಗಿದೆ is ambiguous. It has two meanings such as 'dried' and 'healed'. In the sentence ಬಟ್ಟೆ ಒಣಗಿದೆ. The correct sense is 'dried' and in the second sentence ಅವನ ಗಾಯ ಒಣಗಿದೆ . The correct sense is 'healed'. Assigning correct sense to ambiguous verbs is called as Verb Sense Disambiguation. The KVSD module will address the issue.

The KVSD module uses argument structure for Kannada Verb Sense Disambiguation. The argument structure is the most significant component of the grammar that acts as an interface between syntax and semantics of the language. We argue that the argument structure of a verb will play a major role in disambiguation task. The context in which the ambiguous verb appears is the only means to resolve the ambiguity. Hence, the module considers the arguments and their relationship with verb

in a given sentence as a context to disambiguate the verb.

The concept of argument structure is borrowed from logic. It generally concern with relations between predicate and a set of arguments. The crucial element of a sentence in Kannada is predicate, which is usually a verb or noun. The present module considers verbal predicates only. The predicate determine the

presence or absence of other crucial elements in a sentence. The argument structure for the example sentences in focus are

ಬಟ್ಟೆ ಒಣಗಿದೆ = ಒಣಗು (ಬಟ್ಟೆ) [oNagu] ([baTTe]) 'dried(cloth)'

ಅವನ ಗಾಯ ಒಣಗಿದೆ = ಒಣಗು (ಅವನ, ಗಾಯ).

[oNagu] ([avana, gaaya]) 'healed(his, wound)'.

where ಒಣಗು is a predicate and ಬಟ್ಟೆ , ಅವನ,ಗಾಯ are arguments.

The module disambiguates the verb by extracting the verb and its arguments with their semantic features from the given sentence. A match with a relevant cluster of arguments and the argument structure frame of the verb results in the identification of the correct sense. The following sentences are the output generated by the KVSD module for the illustrated examples.

ಬಟ್ಟೆ ಒಣಗಿದೆ 'Cloth dried'.

ಅವನ ಗಾಯ ಒಣಗಿದೆ 'his wound has healed'.

3.8 KRBWSD module

Compare to English, the number of ambiguous words in Kannada are less. Many instances of ambiguities can be resolved at dictionary and morphological analysis level itself. But morphological analyzer itself will introduce some systematic ambiguities. The ambiguities introduced by morphological analyzer are rule governed. All these ambiguities are resolved by formulating set of syntactic and semantic rules. It is the motivation behind introducing KRBWSD module.

Consider an example ಉದಯವಾಗಲಿ ನಮ್ಮ ಚೆಲುವ ಕನ್ನಡ ನಾಡು . The word ಚೆಲುವ [celuva] is ambiguous in the above example. It has two distinct meanings such as name of a person (proper noun) or beautiful (adjective). In an example sentence, the correct meaning of ಚೆಲುವ [celuva] is 'beautiful'. Assigning 'beautiful' meaning to ಚೆಲುವ [celuva] is the responsibility of KRBWSD module.

The KRBWSD module implements syntactic and semantic constraints of the ambiguous words to be disambiguated. Syntactic constraints are framed by defining the follow set on lexical categories. The follow set puts the restriction on what all the lexical categories of the succeeding words in a sentence can follow lexical category of a word in focus. The disambiguation being done using syntactic constraints is purely in syntactic nature. We may also perform disambiguation based on semantic information and type of ambiguity. In a sentence, the ambiguous words can be disambiguated, using the ambiguous word neighboring words semantic and lexical information. By making use of neighboring words cues, semantic constraints are framed, then the semantic constraints are treated as a binary relation for word sense disambiguation. The following sentence is the output generated by the KRBWSD module for the illustrated examples.

ಉದಯವಾಗಲಿ ನಮ್ಮ ಚೆಲುವ ಕನ್ನಡ ನಾಡು.

[udayavaagali namma celuva kannada naaDu]
'May our beautiful Kannada state arise'

4. EVALUATION

The system is tested on randomly selected 500 sentences from Kannada raw corpora. These sentences are selected by sentence extractor module. The partial list of extracted sentences used during the system evaluation is shown in Figure 4. All major grammatical categories of words have been covered. Ten fold cross validation has been performed in all test cases.

4.1 Result

Table 2 describes the results obtained by different disambiguation modules in a proposed Kannada Word Sense Disambiguator tool. Entire system is implemented in Perl under Linux environment.

Table 2. Program result.

Test Sentences	Comments
ನೆರೆಯಲ್ಲಿ ನೆರೆಹೊರೆಯವರೆಲ್ಲಾ ಕೊಚ್ಚೆ ಹೋದರು.	Correct
ಪಾಕಿಸ್ತಾನ ನೆರೆ ಸಂತ್ರಸ್ಥರಿಗೆ ನೆರವು.	Correct
ಬಟ್ಟೆ ಒಣಗಿದೆ.	Correct
ಅವನ ಗಾಯ ಒಣಗಿದೆ.	Correct
ಉದಯವಾಗಲಿ ನಮ್ಮ ಚೆಲುವ ಕನ್ನಡ ನಾಡು.	Correct
ಹಾರು ಬೂದಿ.	Incorrect
ಅಂತಹ ಹಲ್ಕಾ ಕೆಲಸ ಮಾಡಬೇಡ.	No Output
ವಷ೯ ನನ್ನ ತಂಗಿ.	Partially Correct
ಸೀತೆ ನೆರೆದರು.	Incorrect
ರಾಜಾಬಾಗ್ ಸವಾರನ ಉರುಸು ಖ್ಯಾತವಾಗಿದೆ.	Incorrect

4.2 Discussions

During the process of building and testing the proposed systems, the following observations are made.

- The word sense disambiguation task highly depends on lexical and syntactic information along with semantic information. Hence, good parser will play a major role at syntax level during disambiguation process.
- The creation of Verb Argument Structure Frame file and Verb Argument Semantic Feature file will play a critical role in the verb sense disambiguation process. If these two files provide the exhaustive information then the performance of the proposed system is guaranteed to be high.
- Due to wrong analysis by the morphological analyzer the word ಹಾರು [haaru] assigned incorrect sense.
- Due to missing entry in the dictionary (Kannada Shallow Parser), the word ಹಲ್ಕಾ [halkaa] is not assigned with any sense.

- The word ವರ್ಷ [varSha] assigned with Common noun meaning but it is a proper noun. Lexical sub categorization of words in a dictionary will solve the problem.
- The system assigns incorrect sense 'gather' for a sentence ಸೀತೆ ನೆರೆದರು [siite neredaru] instead of 'biologically matured' sense. This is because of the insufficient context information. This kind of problems can be easily addressed at discourse level analysis but it is behind the scope of the present work.
- In a sentence ರಾಜಾಬಾಗ್ ಸವಾರನ ಉರುಸು ಖ್ಯಾತವಾಗಿದೆ [raajaabaag savaarana urusu khyaatavaagide], ರಾಜಾಬಾಗ್ ಸವಾರನ [raajaabaag savaarana] is a proper noun and also, it is a multiword expression. But, during the disambiguation process, the system interpret it, as a two separate words and assigns the senses separately, it leads to incorrect disambiguation. Hence, handling multiword expression is a critical issue in the disambiguation task.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a Kannada Word Sense Disambiguator, a suite of Perl programs used to disambiguate the polysemous Kannada words. It is a need based valuable resource for executing NLP tasks. As an experimental setup, we constructed five Million Kannada web corpora using a Corpus builder module. Using this corpus, the Dictionary builder module built around 50000 words electronic machine readable dictionary. The correct sense of the polysemous word is assigned by three different modules namely KTWSD, KVSD and KRBWSD modules depending on the target word list and Lexical categories of the polysemous word. Experiments are conducted and the results obtained are described. The efficiency of the proposed system is proved to be reliable and extendable. Due to unavailability of the earlier systems for the same tasks, we are not able to do the performance comparison of the proposed system. The performance achieved by our system can be used as a baseline for further research in this direction.

Even though, the present research is a contributory work to computer processing of Kannada language in its own way, what we achieved so far is meager compare to what we actually required and what we want to become. It is a long journey, we must be ready to face many challenges to bring robust Kannada Word Sense Disambiguator, and hence the future work in this direction can be done in two fold. Firstly, fix the errors introduced by the dictionary and morphological analyzer in the present system with necessary treatment. Secondly, handle the ambiguities introduced at semantic and discourse level by incorporating the necessary modules in the existing system.

6. REFERENCES

[1] Kavi Narayana Murthy and Bharadwaja Kumar, G. 2006. Language Identification from Small Text Samples. Journal of Quantitative Linguistics. Vol 13, No 1. pp. 57-80.

[2] Kavi Narayana Murthy. 2001. Computer processing of Kannada Language. In Proceeding of the KUWH.

[3] Ray, P. R., Harsha, V., Sudesna Sarkar and Anupam Basu. 2003. Parts of Speech tagging and local word grouping techniques for Natutal language parsing in Hindi. In Proceedings of the ICON – 2003.

[4] Sridhar, S. N. 2007. Modern Kannada Grammar. Manohar Publications & Distributors. New Delhi.

[5] Eneko Agirre and Philip Edmonds. 2007. Word Sense Disambiguation: Algorithms and Applications. Text, Speech and Language Technology. Vol 33. Springer.

[6] Fellbaum Christiane. 1998. WordNet: An electronic Lexical database. MIT Press.

[7] Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceeding of the SIGDOC ’86.

[8] Walker, D. and Amsler, R. 1986. The Use of Machine Readable Dictionaries in Sublanguage Analysis. In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pp. 69- 83.

[9] Aggire, E. and Rigau, G. 1996. Word Sense Disambiguation using Conceptual density. In Proceeding of International Conference on Computational Linguistics.

[10] Mihalcea Rada. 2005. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP), Vancouver, Canada, pp. 411-418.

[11] Lee Yoong, K., Ng Hwee, T. and Tee chia, K. 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. Barcelona. Spain. pp. 137-140.

[12] Ng Hwee, T. and Hian, B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL). Santa Cruz. U.S.A. pp. 40-47.

[13] Yarowsky David. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the association for Computational Linguistics (ACL). Las Cruces. U.S.A. pp. 88-95.

[14] Yarowsky David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL). Cambridge. MA. pp. 189-196.

[15] Véronis Jean. 2004. HyperLex: Lexical cartography for information retrieval. Computer Speech & Language, Vol 18 No 3. pp. 223-252.

[16] Schütze, Hinrich. 1998. Automatic word sense discrimination. Computational Linguistics, Vol 24 No 1. pp. 97–123.

[17] Roberto Navigli, Paolo Velardi. 2005. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. IEEE Transactions On Pattern Analysis and Machine Intelligence.

[18] Banerjee, S and Pedersen, T. 2002. Lesk algorithm for word sense disambiguation using Wordnet. Computational Linguistics and Intelligent Text Processing. pp. 117-171.

[19] Sasi Kanth Ala and Kavi Narayana Murthy. 2004. Significance of syntactic features for word sense disambiguation. Advances in Natural Language Processing. In the proceedings of the fourth Int. Conf. EsTAL 2004. Alicante. Spain.

[20] Umamaheshwar Rao, G. and Rajyarama, K. 2010. Representation of Complex Predicates in Wordnet. In Proceedings of the 5th Global Wordnet Int. Conf. IIT, Bombay. India. pp. 271-276.

[21] Parameswarappa, S and Narayana, V.N. 2011. Target word sense disambiguation system for Kannada language. In Proceedings of Int. Conf. on Advances in Recent Technologies in Communication and Computing. ARTCom -2011. Bangalore. India.

[22] Baker, M. 1988. A Theory of Grammatical function changing. Chicago, The University of Chicago Press.

[23] Grimshow, J. A. and Mester. 1990. Argument Structure. Cambridge Massachusetts, MIT Press.

[24] Taegoo Chung 2000. Argument structure and English Grammar. Korea University.

[25] Hoa Trang Dang and Martha Palmer. 2005. The Role of Semantic Roles in Disambiguating Verb Senses. In Proceedings of the 43rd Annual Meeting of the ACL. Ann Arbor. pp. 42–49.

[26] Rafiya Begum, Samar Husain, Lakshmi Bai and Dipti Misra Sharma. 2008. Developing Verb Frames for Hindi. In Proceedings of LREC - 08.

[27] Rafiya Begum, Dipti Misra Sharma. 2010. A Preliminary Work on Causative Verbs in Hindi. Eighth Workshop on Asian Language Resources (ALR8) held in conjunction with The 23rd Conference on Computational Linguistics. COLING – 2010.

[28] Matthew Brook O'Donnell, Nick Ellis. 2010. Towards an Inventory of English Verb Argument Constructions. In Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics. pp. 9–16.

[29] Parameswarappa, S and Narayana, V.N. 2012. Kannada Verbs and their Automatic Sense Disambiguation. To appear in the Proceedings of Int. Conf. on Global wordnet. GWC-2012. Kunibiki messe, Japan.

[30] Warmter, S. 1989. Integration of syntactic and semantic constraints for Structural noun phrase disambiguation. In Proceedings of the IJCAI – 1989.

[31] Bharathi, A., Chaitanya, V., and Sangal, R. 1995. Natural Language Processing: A paninian Perspective, PHI – 1995.

[32] Parameswarappa, S and Narayana, V.N. 2011. Rule Based Kannada Word Sense Disambiguator. To appear in the Proceedings of Int. Conf. on Data Engineering and Communication System. ICDECS-2011. Bangalore, India.

[33] Barlow, M. 1996. Corpora for theory and Practice. In Journal of Corpus Linguistics. Vol. 1, No. 1. pp. 1-38.

[34] Lancashire, I., Percy, C and Mayer, C. 1996. Synchronic Corpus Linguistics. Rodopi, Amsterdam, Atlanta.

[35] Oostdijk, N and Hann, P. 1994. Corpus based research into Language, Rodopi, Amsterdam, Atlanta.

[36] Teubert, W. 2000. Corpus Linguistics: A Partian View. In Journal of Corpus Linguistics. Vol 4, No. 1. pp. 1-16.

[37] Adam Kilgarriff, Siva Reddy, Jan Pomikalek and Avinesh, P.V.S. 2010. A corpus factory for many languages. In Proceedings of the LREC – 2010.

[38] Adam Kilgarriff and Girish Duvuru. 2011. Large web corpora for Indian languages. In Proceedings. of International Conference on Information Systems for Indian Languages.

[39] Parameswarappa, S., Narayana, V.N. and Bharathi, G.N. 2012. A Novel Approach to build Kannada Web Corpora. To appear in the Proceedings of Int. Conf. on Computer Communication and Informatics. ICCCI-2012. Coimbatore, India.

[40] Kavi Narayana Murthy. 1997. Electronic Dictionaries and Comp tools. Linguistics Today. Vol. 1, No. 1. pp. 34-50.

[41] Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford University Press. Oxford.

[42] Wikipedia. (online) http://kn.wikipedia.org.

[43] Wiki Dictionary. (online) http://kn.wiktionary.org.

[44] Sampada. (online) http://sampada.net

[45] Kannada web blog. (online) http://kannadablogs.ning.com/

[46] Prajavani. (online) http://prajavani.net

[47] Parser. (online). http://ltrc.iiit.ac.in/analyzer/kannada/

[48] Bharathi, A and Sangal, R. 1993. Parsing free word order languages in Paninian framework. In Proceedings of the ACL-1993. pp. 105-111.