Malayalam Word Sense Disambiguation using Maximum Entropy Model

Jisha P Jayan, Junaida M K, Elizabeth Sherly
Indian Institute of Information Technology and Management
Kerala

***Abstract*— Word Sense Disambiguation (WSD) is one of the pressing problems in Natural Language Processing (NLP), which determines the right sense of an ambiguous word in the specific context. WSD is considered as a harder problem as it depends on a set of classes, which vary depending on the context. There have been a number of researches performed in WSD in many languages, from dictionary-based methods to supervised learning methods and unsupervised learning approaches. In this paper, an attempt is made for the disambiguation of Malayalam words, which has a rich set of ambiguous words having different meanings. A semi supervised machine learning techniques combined with statistical approach mainly Maximum Entropy is experimented for Malayalam, which shows promising result for a set of corpus trained.**

Keywords — Machine Translation, Word Sense Disambiguation, Machine Learning, Maximum Entropy Model.

I. INTRODUCTION

Word Sense Disambiguation (WSD) or Lexical Ambiguity Resolution is a fundamental task, which processes to identify the sense of a word in a given sentence. A word can have multiple meanings and such words are called polysemy. In English, words like bass, line, hard etc. can be considered as an example of polysemous words. The task WSD is a historical one in the field of NLP. In 1940 itself, it was conceived as a fundamental task of Machine Translation (MT). So WSD is considered as one of the difficult problems in Artificial Intelligence, whereas for a human being it is much easier to identify the sense of a word. However, an automated WSD is essentially an important component in NLP applications and machine translation systems, including information retrieval (IR), text summarization etc.

Automatic WSD systems are available for many languages like English, Spanish, Chinese and some Indian languages. The work on automatic WSD for non-English languages is a severe problem, especially for unstructured language like Malayalam. Malayalam language is one of the major Dravidian languages spoken in India by around 36 million people. But research in Malayalam language processing is in a developing stage. The most probable reason is being lack of labeled text corpora and unavailability of Malayalam language tools. Word sense disambiguation is highly dependent on the sense inventory like a dictionary and thesaurus.

Ambiguity of natural language is a one of the major problems associated with the NLP applications. The WSD task is an important component of several NLP systems, such as machine translation, Question Answering (QA), information retrieval, information extraction and speech processing applications. Researchers in MT have concentrated efforts on WSD since the earliest NLP applications. MT researchers identified that their results would be considerably better by using WSD methods to disambiguate words in the automatic translation for various pairs of languages.

This paper is organized into different sections. First section dealt with the introduction part. The second section deals with the major works carried out in this area. The next section explains the proposed work. The fourth section includes the implementation and the result obtained. The fifth section concludes the paper with future works that can be done as an outcome of this work.

II. STATE OF THE ART OF WSD

There are many different approaches used for identifying WSD. The two main approaches are Dictionary based approaches and Corpus based approaches. The former approaches are mainly using external lexical resources such as dictionaries, thesaurus, WordNet, etc. These are easy to implement because they require a simple lookup of a knowledge resource like a machine readable dictionary. In later approaches large training corpus is involved with so many algorithms.

The corpus based methods use techniques from statistics and machine learning to induce models of language and large samples of text [1]. Learning can be done with supervised or unsupervised methods, which learns sense classifiers from annotated data with minimal or partial human supervision respectively. Semi supervised and unsupervised algorithms do not need large amount of annotated corpora as it uses word specific classifiers [2, 3]. The semi supervised method makes use of annotated corpora as seed data for bootstrapping process. An example for the semi supervised algorithm is Kannada WSD uses decision list [4].

Niladri Chatterjee and Rohit Misra [5] presented a trainable model for Word Sense Disambiguation (WSD) for resolving the ambiguity of English words. The proposed model applies concepts of information theory, in particular the maximum entropy model.

S. Parameswarappa [6] described the possible techniques for Kannada target word sense disambiguation using compound word clue and syntactic features in a local context. The ambiguous target word disambiguated using supervised machine learning with a naive Bayes classifier.

Lesk [7] was one of the first researchers who tried to disambiguate Machine Readable Dictionaries (MRD) using algorithms. His algorithm became well-known among WSD researchers. His algorithm was primarily an overlap based algorithm which suffers from overlaps scarcity. These methods, highly rely on lexical resources such as machine readable dictionaries, thesaurus, etc. In English, this method achieved 50-70% accuracy [5] in correctly disambiguating the words.

The work by Sinha [8] mainly focused on Hindi. They used contextual overlap between sentential context and extended sense definitions from Hindi Word Net. Sense bag was created by extracting words from synonyms, glosses, example sentences, hyponyms, and glosses of hyponyms, example sentences of hyponyms, hypernyms, and glosses of hypernyms, example sentences of hypernyms, meronyms, glosses of meronyms, and example sentences of meronyms. A context bag was created by extracting words in the neighborhood, i.e. one sentence before and after, of the polysemous word to be disambiguated. The sense which maximized the overlap was assigned as winner sense.

Rosna P Haroon [9] proposed solution to Malayalam WSD uses a knowledge based approach. One approach based on a hand devised knowledge source and used the Lesk and Walker algorithm. The other is using the concept of conceptual density, by using Malayalam WordNet as the lexical resource. The knowledge base systems will result in poor accuracies, because of the accuracy of first algorithm depends on the stored tag words within the knowledge source. If more tag words are there, Algorithm work for so many sentences. The accuracy of the algorithm based on conceptual density depends on the implementation of semantic relations in WordNet.

III. PROPOSED METHOD

In the present work, we propose a machine learning approach using maximum entropy models for unrestricted Malayalam text WSD. This is an early attempt for automatic WSD using machine learning in Malayalam language.

A. Maximum Entropy Model

Maximum entropy (ME) principle states that the least biased model which considers all known information is the one which maximizes the entropy. We are further encouraged by the fact that information theory has also been applied to solve many problems in natural language processing. Ratnaparkhi [10] discusses the maximum entropy model and its application to some NLP tasks viz. Sentence Boundary Detection, Part of Speech Tagging and Parsing. The ME technique builds a model which assumes nothing other than the imposed constraints. To build such a model, we define feature functions. A feature function is a Boolean function which captures some aspect of the language which is relevant to the sequence labelling task.

Here we consider the sense of the target word as a random variable with various outcomes (which are precisely the various senses in which the given word can be used). In this model we estimate the probability of each sense of the target word given the context.

Feature function for WSD is,

$$F(o,c) = \begin{cases} 1, if o = x \text{ and } c = y, \\ 0, otherwise \end{cases}$$

where “o” stands for outcome and “c” stands for context. This function maps contexts and outcomes to a binary set. Depending upon the input values of o and c, the feature returns either 0 or 1. Thus a feature is 'active' (i.e. f = 1) only if the associated keyword is present in the context and also the sense being estimated is the same as the sense of the feature. Note that both the keyword and the sense are inbuilt within a feature.

For example, the word “രസം” [rasam] has five senses, example രുചി *[ruchi]* sense and അനുഭൂതി *[anubhoothi]* sense. We can construct features associated with both these senses. Suppose we want to notice the presence of the keyword "മധുരം"[madhuraM] and note its correlation with either of the senses. Then we can construct two features, in one feature ‘o’ is the രുചി *[ruchi]* sense, while in the other ‘o’ is the അനുഭൂതി *[anubhoothi]* sense. In both the features, c = y is true if " മധുരം " [madhuraM] is present in the context window, and false otherwise.

To find the maximum entropy distribution the improved iterative scaling (IIS) algorithm is used, which is a procedure for finding the maximum entropy distribution that conforms to the constraints imposed by the empirical distribution of the modelled properties in the training data. The algorithm should run for a fixed number of iterations or till the change in accuracy becomes negligible. In this work, we have taken 50 iterations as a rule of thumb, since any more iteration did not result in any significant change in the accuracy.

B. Complexity of Malayalam Language

This section introduces the linguistic preliminaries of Malayalam language and complexities involved in the Malayalam Word Sense Disambiguation. The world languages are classified into 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 latter case, no fixed ordering is imposed on the sequence of words in a sentence. An example of fixed word order language is English and that of pure free word order language is Sanskrit. Generally Malayalam is a free word order language.

Malayalam is an agglutinating language and exhibits very rich system of morphology. Morphology includes inflection, conflation (sandhi), and derivation. Here we will briefly describe the complexities involved in our work. In addition to the difficulties involved in Word Sense Disambiguation, the complexity level is even more in an unstructured language like Malayalam.

ISSN: 2278-0181

In addition to, the work on the most probable reason is being lack of labeled text corpora and unavailability of Malayalam language tools like a dictionary and thesaurus. For example, consider a sentence ‘അവൻ നടന്നു.’ With the meaning ‘He walked’ and ‘യോഗം നടന്നു’ with the meaning ‘Meeting executed’. Here the distinction of the sense of the word ‘നട’ is very complex due to the lacks of capitalization information. Table I shows the some of the ambiguous words in Malayalam with some senses.

Word	Senses (classes)
രസം	അനുഭൂതി, കറി, ഇഷ്ടം, രുചി, മെർക്കുറി
നട	ക്രിയ, നടക്കുക , പടി
അടി	ചുവടളവ്, പാദം, തല്ല്
വാനം	മാനം, അഭിമാനം
ഉത്തരം	മറുപടി, ചോദ്യോത്തരം, താങ്ങ്

TABLE I. AMBIGUOUS WORDS AND SENSES

To begin with, this experiment requires a sense tagged corpus in-order to achieve considerable accuracy for disambiguation. Developing corpus is a tedious and very time consuming task. The next issue involved in this work is the unavailability of sense inventory which will decide appropriate senses to the specific word in a context. The most appropriate meaning of a word is selected from a predefined set of possibilities, usually known as Sense inventories. The other important issue of this language is the unavailability of an efficient POS tagger. In all other works on Word Sense Disambiguation uses POS features to improve the accuracy. The tagger for Malayalam we found needed a huge corpus for the training phase itself.

A lot of work is being done in the fields of corpus building, creating an efficient POS tagger, subject identification, etc. in Malayalam language, which will support further development in this field when Malayalam language is concerned.

IV. IMPLEMENTATION AND RESULT

This section describes the system architecture and parts of the system. The result obtained from the experiments is described here.

A. *Implementation*

The figure below shows the system architecture of the proposed system. The system consists of the following parts.

Since Maximum Entropy model is a machine learning technique it needs a training corpus. In these experiments, the system is trained using manually collected sentence from various Malayalam newspapers, Wikipedia articles, blogs, books, novels etc.

Before applying the maximum entropy algorithm, for both training and testing data, some pre-processing in the text have been performed. In order to improve accuracy, stop words are removed. Stop words are terms that are too frequently in the text. These terms are insignificant. So, removing them reduces the space of the items significantly in the training and testing text. After pre-processing, the maximum entropy algorithm is applied to the training data. Here we have implemented improved iterative scaling algorithm with 50 iterations. The output is a training model which will be used to classify new instances.

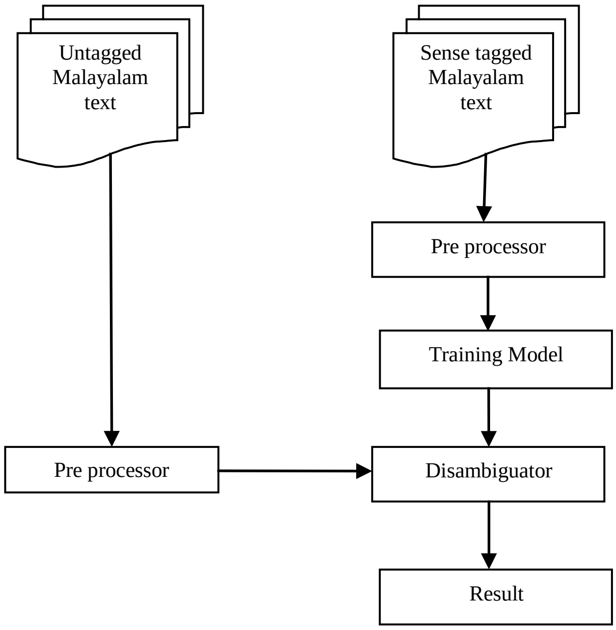

Fig1: Proposed System Architecture

Then input given to the disambiguator is an unlabeled Malayalam text after applying the same pre-processing technique used for training phase. The output obtained is a classifier that classifies each new instance into one of the senses of the target word. We have chosen the word രസം [rasam] as the target word. The each instance of the word രസം [rasam] in the corpora is tagged with one of the five possible Basahamithram Malayalam dictionary senses. Each instance is a single sentence.

B. *Result*

We used randomly selected sentence from the Malayalam web corpora to test the accuracy of the system. The result obtained for some of the instances used to test the program are shown in Table II.

For the selected target word രസം [rasam] the overall accuracy obtained from the experiment is 65% . 40.20% baseline accuracy is obtained for the most frequent sense of the target word, in our corpus, the most frequently occurring sense for രസം [rasam] is അനുഭൂതി [anubhoothi].

TABLE II. SAMPLE EXECUTION RESULT FROM TESTING CORPORA

Test Sentence	Result
ദ്രാവക രൂപത്തിലുള്ള ഒരു മൂലകമാണ് രസം അഥവാ മെർക്കുറി.	Correct
വളരെ പെട്ടന്ന് ഉണ്ടാക്കാൻ പറ്റുന്ന ലളിതമായ ചേരുവകൾ ചേർത്തുള്ള ദക്ഷിണഭരതത്തിൽ ഉടനീളം ഉപയോഗിക്കുന്ന ഒരു കറിയാണ് രസം .	Correct
എന്തൊരു രസം ആ കാഴ്ച്ച കാണാൻ.	Correct
രസം പ്രധാനഘടകമായുള്ള മിശ്രലോഹത്തെയാണ് അമാൽഗം എന്നു പറയുന്നത്.	Wrong
ലളിതമായ ചേരുവകൾ ചേർത്തുള്ള കുരുമുളക് രസം കഴിക്കുന്നത് ദഹനത്തെ സഹായിക്കും.	Correct
പ്രധാനമായും ആറ് രസം ആണുള്ളത്, മധുരം, പുളിരസം, ലവണം, തിക്തം, കടു, കഷായം.	Wrong

The precision and recall obtained from the word രസം [rasam] for each sense is shown in Table III. The average precision and Recall for the program is 84.1% and 61.32% respectively.

TABLE III. PRECISION AND RECALL OF EACH SENSE OF THE WORD രസം [RASAM]

Label	Precision	Recall
രസം%അനുഭൂതി	53.8	100
രസം%രുചി	100	33.3
രസം%നവരസം	100	33.3
രസം%മെർക്കുറി	100	40.0
രസം%കറി	66.7	100

V. CONCLUSION AND FUTURE WORK

In this paper, we have exposed the research carried out on applying statistical and machine learning based algorithm for the Word Sense Disambiguation (WSD) problem. An attempt has been made for the disambiguating Malayalam words, which has a rich set of ambiguous words with different meanings. This is an early attempt towards the Word Sense Disambiguation for Malayalam language using Maximum entropy based machine learning approach. However, no tagged corpus was available to us for use in this task. Manual sense tagging is quite a time consuming and difficult process. The accuracy of the WSD and performance of the system depends on the size of the corpus. As a future work Corpus creation and sense tagging can be automated. Future work can also include the improving of the performance of this system by using large training corpus and handling morphology exhaustively.

REFERENCES

[1] S.Banerjee and T. Pederson, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet", In Proceedings of the Third International Conference on Computational Linguistics And Intelligent Text Processing, 2002.

[2] Yorowsky David, "Unsupervised word sense disambiguation revealing supervised methods", In Proc. of the 33rd Annual Meeting of the association for Computational Linguistics (ACL), 1995.

[3] Yorowsky David, "Decision list 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),1994.

[4] S.Parameswarappa1 and V.N. Narayana, "Kannada Word Sense Disambiguation Using Decision List", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), June, 2013.

[5] Niladri Chatterjee and Rohit Misra, "Word-Sense Disambiguation using Maximum Entropy Model", International Conference on Methods and Models in Computer Science, 2009.

[6] S.Parameswarappa1 and V.N. Narayana, "Target word sense disambiguation System for Kannada Language", Proceedings of Int. Conf on Advances in Recent Technologies in Communication and Computing, 2011.

[7] M. E. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proc. SIGDOC Conference, Toronto, Ontario, June, 1986.

[8] M. Sinha, M. Kumar, P. Pande, L. Kashyap and P. Bhattacharyya, "Hindi Word Sense Disambiguation", International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, 2004 sylvania, 1998.

[9] Rosna P Haroon, "Malayalam Word Sense Disambiguation", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), December, 2010.

[10] Adwait Ratnaparkhi, "Maximum Entropy Models For Natural Language Ambiguity Resolution", A Dissertation in Computer and Information Science, Presented to the Faculties of the University of Penn