Journal of Engineering Science and Technology
Vol. 12, No. 9 (2017) 2504 - 2512
© School of Engineering, Taylor's University

WORD SENSE DISAMBIGUATION FOR TAMIL LANGUAGE USING PART-OF-SPEECH AND CLUSTERING TECHNIQUE

P. ISWARYA*, V. RADHA

Department of Computer Science, Avinashilingam Institute
for Home Science and Higher Education for Women, Coimbatore, India
*Corresponding Author: iswaryacbe333@gmail.com

Abstract

Word sense disambiguation is an important task in Natural Language Processing (NLP), and this paper concentrates on the problem of target word selection in machine translation. The proposed method called enhanced Word Sense Disambiguation with Part-of-Speech and Clustering based Sense-collocation (WSDPCS) consists of two steps namely (i) Part-of-Speech (POS) tagger in disambiguating word senses and (ii) Enhanced with Clustering and Sense-collocation dictionary based disambiguation. In the first step an ambiguous Tamil words are disambiguated using Tamil and English POS Tagger. If it has same type of POS category labels, then it passes the word to the next step. In the second step ambiguity is resolved using sense-collocation dictionary. The experimental analysis shows that the accuracy of proposed WSDPCS method achieves 1.86% improvement over an existing method.

Keywords: Ambiguity, Clustering, Collocation, Dictionary, Disambiguation, Part-of-Speech.

1.Introduction

Word Sense Disambiguation (WSD) is a task of identifying correct sense of each ambiguous word occurrence in a sentence [1]. WSD is used in variety of applications namely Machine translation, Information Retrieval (IR), content & thematic analysis, grammatical analysis, speech processing etc. The paper focuses on developing WSD procedure for Tamil to English machine translation and intelligent IR system applications. For example consider a Tamil sentence:

- Sentence1: **என் கேள்விக்கு விடை சொல்**
- Sentence 2: ராமு வீட்டில் இருந்து **விடைப்** பெற்றான்.

Abbreviations

CWSD	Conventional word sense disambiguation
IR	Information Retrieval
WSD	Word Sense Disambiguation
WSDPCS	Word Sense Disambiguation with Part-of-speech and Clustering based Sense Collocation

In this example, an ambiguous word is “விடை” which has at least two possible senses, i.e., answer and relieved. The good quality of translation can only be achieved by choosing a right sense of an ambiguous word, and this process of identifying a correct sense for a word is done using WSD procedure. The WSD algorithm disambiguates the ambiguous word based on their context or collocations. Collocations are the words that are adjacent to a target word, strongly indicating the sense of an ambiguous word.

Basically there are three types of lexical ambiguities they are Polysemy, Homonymy and Categorical ambiguity [2].

- Polysemy ambiguity: It is a word or phrase with several senses, but they are related to one another. It occurs in the form of both noun and verb POS categories that can inflect for more than one sense. The word “ பிடி” is one of the most ambiguous polysemous word, and it has several senses such as capture, shape, catch and massage.
- Homonymy ambiguity: The word or phrase having multiple meaning, but they are of totally unrelated senses. Homonymous words have senses that are clearly distinct unlike the case of polysemous words. The Tamil word “மாலை” is homonymous word with two different senses such as evening and garland.
- Categorical ambiguity: The word or phrase having multiple meanings, and each meaning has different grammatical categories called categorical ambiguity. For example, the Tamil word “ஓடு” has distinct senses with different POS categories such as Tile (Noun) and run (verb).

The paper aims in disambiguating homonymous and categorical ambiguity Tamil words. The organization of the paper is as follows. Section 2 outlines the earlier works on WSD and Section 3 describes about the proposed POS and Clustering based WSD approach. The experimental results and their discussions are presented in Section 4. Finally Conclusion is stated in Section 5.

2. Related Works

Lesk [3] disambiguates multiple word senses by counting the overlaps between definitions of various senses. He carried out experiments with three dictionaries namely Oxford Advance Learner dictionary, Merriam-Webster 7th new collegiate and Collins English dictionary.

Dagan and Itai [4] resolve lexical ambiguities using statistical model. The parser initially identifies syntactic relation of an ambiguous word of the Source language, and corresponding target language. A bi-lingual lexicon is used to find all possible translations of ambiguous word. The procedure of using statistical mapping of syntactic relations is carried out to choose right alternative of target word. The experiments are evaluated with three sets of data in Hebrew and German language.

Yarowsky [5] stated that the large sense tagged corpus is not necessary to achieve higher WSD performance. The author presented the unsupervised learning method of WSD based on the property one sense per collocation and one sense per discourse. His experimental result exploits an iterative bootstrapping procedure, and it outperforms the supervised learning methods.

Ng and Lee [6] resolves ambiguity using exemplar-based learning algorithm. This algorithm integrates multiple knowledge resources, namely part-of-speech of nearby words, morphological form, an unordered set of surrounding words, local collocations and syntactic relation. The experiments were conducted on common dataset and large scale dataset. The whole approach is named as LEXAS which performs better than the previous existing works.

Yarowsky [7] proposed a word sense disambiguation approach that disambiguates English word senses using statistical models of Roget Thesaurus categories. The model overcomes the knowledge acquisition bottleneck problem by enabling training on unrestricted monolingual text without human intervention. The statistical model consists of three steps; the first step involves collection of set of context words that are representative of Roget categories. In the second step identify the words that occur frequently in collective context, and assign probabilistic weight for each of it using maximum likelihood estimator. The category having maximum weight is predicted, and that type of sense is assigned.

Brown et al. [8] introduced statistical machine translation system for disambiguating English and French word senses. The translation model incorporated the concept of Viterbi algorithm for the alignment of short French and English sentences obtained from Canadian Hansard Corpus. The right sense is assigned from the translation model by computing higher mutual information between English and French words. The experimental results show that the proposed model decreases the error rate by thirteen percent.

Sharma and Niranjan [9] integrated the clustering and classification technique for optimizing the performance of word sense disambiguation. Initially K-means clustering technique is applied on the dataset, which results in K clusters. After clustering, random forest classification technique is applied on cluster dataset. These experiments are carried out using data mining tool called WEKA, and it makes use of data file called poach.arff from WORDNET. Their experimental results show that the proposed K-means cluster with random forest achieves 1.3% improvement than using random classifier method alone.

Baskaran and Vaidehi [10] presented an unsupervised approach to word sense disambiguation that extracts collocations automatically from large corpus. The context space is created, by finding all context words for each ambiguous word. To group the occurrences of ambiguous word into different clusters, in such a way that has maximum intra-cluster similarity, and having minimum or zero inter-cluster similarity. To construct sense collocation dictionary, from different clusters top collocations are extracted, and human annotators with the help of collocations assign appropriate sense for each ambiguous word.

Through observation from related works of WSD, it is noted that the experiments were carried out using different methods such as dictionary, thesaurus, sense tagged corpus, supervised, unsupervised and knowledge based methods. Only limited amount of works have focused to resolve word sense disambiguation in Tamil language. In an existing Tamil WSD system [10], it focuses to disambiguate only the homonymous ambiguity Tamil words. In an existing framework of WSD, the user selects a K-value in the K-means clustering; when this value is inappropriately selected, it degrades the clustering performance. The human annotators constructed the sense collocation dictionary which is time consuming and laborious task. To overcome the above challenges the proposed WSDPCS method enhances the existing system, which handles categorical ambiguity words in addition to homonymous Tamil words. To find an optimum cluster, the selection of K-value is automated using ensemble mechanism. In WSDPCS approach, the sense-collocation dictionary construction is automated using bilingual dictionary and Word Net.

3.Proposed POS and Clustering based WSD

The proposed method called enhanced WSD with Part-of-Speech and Clustering based Sense-collocation (WSDPCS) procedure is designed to solve the challenges in the prior work. The process of collocation extraction is automated using clustering technique, and disambiguation of the word sense is carried out in two steps that are pointed below.

(i) Part-of-Speech (POS) tagger in disambiguating Word Senses.

(ii) Enhanced with Clustering and Sense-collocation dictionary based disambiguation.

3.1. POS tagger in disambiguating word senses

The first step is to disambiguate an ambiguous Tamil word directly by applying the Tamil POS tagger [11]. Also the corresponding senses of an ambiguous word are tagged by English POS tagger [12]. If the two different senses have same POS category labels, which are handled by passing an ambiguous word into the next step. Otherwise a categorical ambiguity word is disambiguated using Tamil and English POS tagger. The process of disambiguating word senses using POS tagging is shown in Fig. 1. The input Tamil word is tagged using Tamil POS tagger, and it is given for translation to identify possible English word senses. Then the word senses are tagged using English POS tagger, and labels (POS tags) of both the ambiguous word and corresponding sense words are compared. Finally if the label of sense tag matches with an ambiguous tag label that type of sense is assigned to an ambiguous word.

3.1.1. Enhanced with clustering and sense-collocation dictionary based disambiguation

The section consists of two phases that are training phase and testing phase. The Architecture of WSD uses clustering and sense-collocation dictionary which is presented in Fig. 2. Initially in Tamil document corpus, stop words are removed and ambiguous word list is created. Collocations are context words that appear on either side of an ambiguous word till specified window size, and different window sizes are used by different researches. In this research work, size of the sliding window is 25, i.e. context words of 25 are extracted before and after of an ambiguous word. All the context words of an ambiguous word are collected from the document corpus to create a Context space which is denoted as S and it is shown in Fig. 3. Context Space S consists of context vectors, and these context words are morphological analyzed to return root words and these are included in context word list.

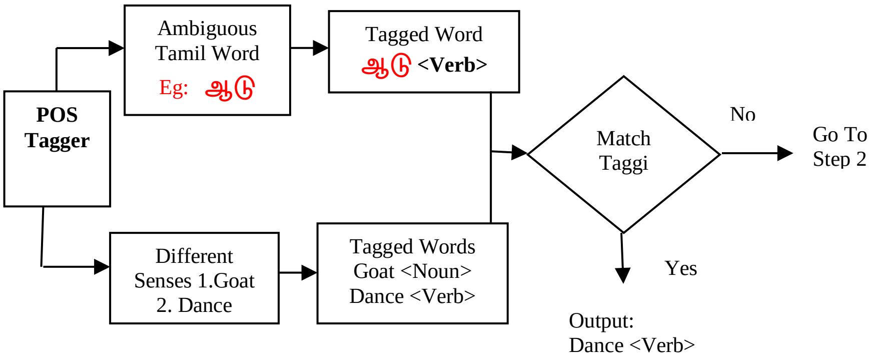

Fig. 1. Process of disambiguating word senses using POS tagging.

The next step in training phase is Enhanced K-means clustering that produces k-final clusters. The performance of conventional K-Means algorithm depends on the selection of K value, and if the user selects unsuitable K value that degrades clustering as well as disambiguation performance. To solve the above issue of selection of optimal K value, an ensemble technique is used. Enhanced K-means algorithm generates K centroid points, and different K cluster centers ranging between 2 to 50.

The squared Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster. The cluster centroids are recomputed iteratively until no object moves to the clusters. To find an optimal K value and optimal clustering set, majority voting algorithm is implemented. The selection of optimal K-value is embedded during the clustering process, and this automated process saves search time while considering large number of ambiguous words.

At the end of clustering, k-final clusters are created and the collocations present in these clusters are contextually similar. Before extracting the collocations from the clusters, identification of potential seed words in cluster is an important task. In each cluster, collocations are ranked according to the high priority based on the log-likelihood ratio using the following Eq. (1).

$$log\left(\frac{P(Sense_A/Collocation_i)}{P(Sense_B/Collocation_i)}\right) \qquad (1)$$

If log-likelihood score of the collocation word is higher than the threshold value, then it is selected for top seed collocation; otherwise it is rejected.

The top collocations are extracted from the clusters are assigned with right senses in an automatic manner. The top Tamil collocations are translated into English collocations using bilingual dictionary. Then the ambiguous Tamil word is associated with right English sense word with the help of Word Net. In English, Word Net [9] is frequently used for both its semantic and syntactic information to disambiguate words in general texts. Due to lack of parallel corpus and Tamil Word Net, the disambiguation is carried out using bilingual dictionary and Word Net. Finally sense-collocation dictionary is constructed having attributes such as ambiguous words, major sense and top Tamil collocations and their English translations. The words that cannot be disambiguated using step1, is handled using sense-collocation dictionary. The sample entries in sense-collocation dictionary are shown in Table 1.

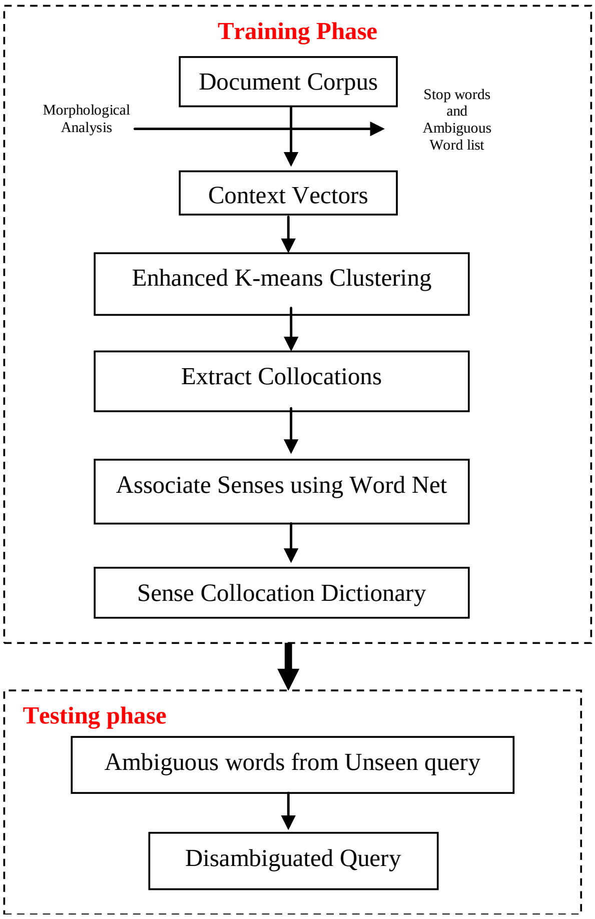
Training Phase
Document Corpus
Morphological Analysis
Stop words and Ambiguous Word list
Context Vectors
Enhanced K-means Clustering
Extract Collocations
Associate Senses using Word Net
Sense Collocation Dictionary
Testing phase
Ambiguous words from Unseen query
Disambiguated Query

Fig. 2. Architecture of WSD using clustering and sense-collocation dictionary.

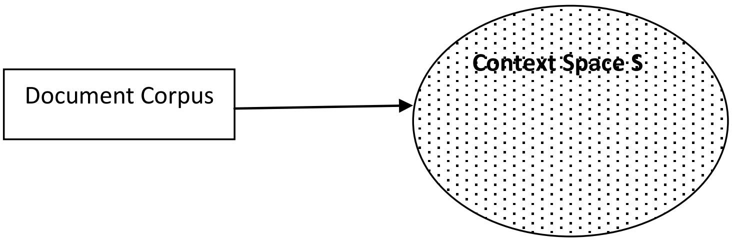

Fig. 3. Representation of context space.

Table 1. Sample entries in sense-collocation dictionary.

Ambiguous Words	Sense	Tamil collocations	English collocations
மாலை	Evening	காலை, பகல், சூரியன், நாள்,மாதம், இரவு மதியம், நேரம், விடியல்.	Morning, dawn, sun, day, month, night, afternoon, time, sunrise
மாலை	Garland	பூ, மணம், ரோஜா, மல்லிகை, பூஜை, நறுமணம் மகளிர், பூங்கொத்து	Flower, marriage, rose, jasmine, pooja, odour, women, bouquet
நூல்	Book	படி,ஆசிரியர், படிபகம், அச்சு, பக்கம் தாள், பரீட்சை	Read, teacher, library, print, page, paper, exam
நூல்	Thread	ஊசி,தறி, சாயம், இலை, கதர், நெசவு, உடுத்து, வண்ணம், துணி	Needle, Weave, dyeing, leaf, cotton, spinning, wear, colour, cloth

4. Results and Discussion

From FIRE dataset 2011, the Tamil corpus is used for evaluating the performance of the existing and proposed word sense disambiguation algorithms. Table 2 shows the average results with respect to the selected parameters produced by the existing conventional WSD (CWSD) and proposed (WSDPCS) algorithms. The average performance was estimated by considering the whole database.

From the average results, it is evident that the WSDPCS perform better than the conventional CWSD algorithm. The overall performance indicated by F-Measure and accuracy parameters show that the proposed WSDPCS algorithm has improved the process of word sense disambiguation by 4.11% and 1.94% when compared with CWSD algorithm.

Table 3 shows the results of CWSD and WSDPCS for 12 randomly selected ambiguous words for the selected metrics. For each ambiguous word, only two major senses are considered. In Table 3, the first and second columns show the ambiguous word and its corresponding two major senses. Both CWSD and WSDPCS were trained using 80% of the corpus data, and the remaining 20% were used as test data. The rest of the columns in the table present the performance of the algorithms in disambiguating the unseen raw occurrences from the 20% test corpus, with respect to the selected performance metrics.

It can be observed from the table, that the algorithm works well for the words such as kal, Kadai and ariah, but words such as tanti and vilangu produced lower results. This is probably due to the data sparseness in the corpus for these words. In all the cases, the proposed algorithm produces improved results than the existing algorithm, indicating that the optimization procedures included are successful in improving the task of word sense disambiguation.

Table 2. Average Performance of WSD algorithm.

Performance Metrics	CWSD	WSDPCS
Precision (%)	84.45	87.76
Recall (%)	80.87	84.62
F-Measure (%)	82.62	86.16
Accuracy (%)	93.80	95.66
Speed (Seconds)	1.28	0.97

Table 3. Performance of WSD algorithms.

Sample Tamil Words	Senses	F-Measure (%)		Accuracy (%)	
		CWSD	WSDPCS	CWSD	WSDPCS
வெள்ளி	Venus / Friday	83.40	85.43	91.87	92.86
விலங்கு	Animal / hand-cuff	76.16	78.29	89.04	89.64
கால்	Leg / measurement	79.82	80.90	93.14	94.15
வாதம்	paralysis / argument	83.23	85.93	91.40	93.12
தந்தி	Telegram / string	75.27	77.04	89.16	90.22
கதை	Story / ancient weapon	86.89	89.42	93.77	95.27
துண்டு	Towel / piece	77.99	80.05	90.56	91.84
ஆடு	Goat / dance	79.90	82.02	92.32	92.97
அரிய	Cut/ Rare	82.51	85.25	93.23	94.53
போலி	Sweet/piracy	84.06	86.59	92.15	93.00
மாலை	Evening/Garland	80.54	82.47	92.89	94.17

5. Conclusions

The paper describes the WSDPCS approach that handles homonymous and categorical type of ambiguity words. This method automates the selection of optimal K-value in the K-means clustering, and also construction of sense-collocation dictionary without human intervention. The approach reduces the processing time and achieves better performance than existing WSD method. The categorical ambiguity words are disambiguated using Tamil POS tagger, and homonymous words were handled with the help of clustering and sense-collocation dictionary. In future, the performance of WSD is improved by tagging the context words and assigning weights for those collocations.

References

1. Panagiotopoulou, V.; Varlamis, I.; Androutsopoulos, I.; and Tsatsaronis, G. (2012). Word sense disambiguation as an integer linear programming problem. *Proceedings of the 7th Hellenic conference on Artificial Intelligence: theories and applications*, 33-40.
2. Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Cambridge: Cambridge University Press.
3. Lesk, M. (1986). Automatic sense disambiguation: how to tell a pine cone from an ice cream cone. *Proceedings of the 1986 SIGDOC Conference*, New York, Association of Computing Machinery, 24-26.
4. Dagan, I.; and Itai, A. (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4), 563-596.
5. Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (Cambridge, MA), 189-196.
6. Ng, H.T.; and Lee, H.B. (1996). Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (Santa Cruz, CA), 40-47.
7. Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the 14th International Conference on Computational Linguistics* (COLING, Nantes, France), 454-460.
8. Brown, P.F.; Pietra, S.A.D.; Pietra, V.J.D.; and Mercer, R.L. (1991). Word-sense disambiguation using statistical methods. *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics* (Berkeley, CA), 264-270.
9. Sharma, N.; and Niranjan, S. (2012). Optimization of word sense disambiguation using clustering in WEKA. *International Journal of Computer Technology & Applications*, 3(4), 1598-1604.
10. Baskaran, S.; and Vaidehi, V. (2004). Collocation based Word Sense Disambiguation using Clustering for Tamil. *International journal of Dravidian linguistics*, 33(1), 13-28.
11. Iswarya, P.; and Radha, V. (2015). Improved tagging approach for part-of-speech in Tamil language using an ensemble. *International Journal of Applied Engineering Research*, 10(6), 14015-14028.
12. Stanford Log-linear Part-Of-Speech Tagger. Retrieved November 26, 2015, from http://nlp.stanford.edu/software/tagger.shtml