International Journal of Engineering & Technology, 7 (2.6) (2018) 189-192

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper



Kannada word sense disambiguation by finding the overlaps between the concepts

B H Manjunatha Kumar[1*], Dr. M. Siddappa[2], Dr. J. Prakash[3]

[1]Reaserch scholar, Visvesvaraya Technological University, Belagavi, Karnataka, India
[2]Prof. and Head, Dept. of CSE, Sri Siddhartha Institute of Technology, Tumkur, Karnataka, India
[3]Prof. and Head, Dept. of ISE, Bangalore Institute of Technology, Bangalore, Karnataka, India
**Corresponding author E-mail: bhm.nlp@gmail.com*

Abstract

We propose three approaches for disambiguating the Kannada word based on an adaptation of dictionary-based Lesk's word sense disambiguation technique. Instead of making use of the regular dictionary as the repository of glosses, we used Indo – WordNet lexical database as the source of senses. Here we adopt a current method of measuring semantic relatedness between the concepts of the Kannada words taken from Indo – WordNet. This measure is dependent on identifying and counting the number of common words present between the glosses of a pair of concepts in accordance with Indo – WordNet.

Keywords: *Indo – WordNet, Kannada Word Sense Disambiguation, semantic relatedness, WordNet.*

1. Introduction

The existence of several meanings in a single word is deeply rooted in natural language. Each natural language possesses plenty of words having several senses. In the English language, the word bark has multiple meanings like the noise or harsh sound of dogs, the tough external layer of a tree and or boat. Humans are fairly good at deciding the proper meaning. However, this task is difficult for computers. The computational process of finding the proper meaning of a word having several meanings is called as WSD – Word Sense Disambiguation. In spite of this arduousness, however, we are fascinated with automating this task and this can play a paramount role in the area of machine translation. The technique proposed, implemented and evaluated in this paper uses Lesk [1] method to measure the semantic relatedness. In WordNet, every concept is described by gloss. Lesk algorithm utilizes the gloss textual content to express the actual concept. In Lesk [1] algorithm the degree of relatedness is calculated by finding the overlaps between the glosses of two concepts, along with concepts which are directly connected to them as per WordNet. In this paper, we have used Indo-WordNet [2] to extract the lexically expressed concepts. We modify the Lesk algorithm to Indo- WordNet and this technique does not require any training corpus.

A good amount of related work papers on utilizing measures of semantic similarity are referred. Budanitsky et al. [3] examined 5 approaches for semantic relatedness measure. They contrasted the efficiency of the five approaches in correcting the spelling mistakes. They recorded an accuracy of 65%. They also identified a measure based on information content proposed by Jiang is better than Hirst [4], Resnik [7], Lin [6], and Leacock et al. [5]. Reddy.S et al. [8] contrasted and evaluated 6 measures of semantic relatedness in semantic classification and labeling for the Hindi language. They identified that performance of modified Lesk is higher than other five measures. Sinha et al. [9] suggested a graph based unsupervised algorithm for WSD and also produced outputs on dataset like SENSVAL-2 and SENSEVAL-3 using 6 various measures. Torres et al. examined amalgamation of similarity measures and their experimental outcomes indicate that mixture of various combination measures performs more accurate as compared to every individual measure. Even though, these kinds of outcomes cannot be comprehended for the Kannada language. In our paper, we tried to find out the benefits of Lesk technique to measure the semantic similarity for the Kannada language. To the best of our insight, this kind of work isn't accounted previously for Kannada WSD. Sinha et al. [14] conducted context overlapping for Hindi WSD by utilizing extended Lesk algorithm. They used the sense definition of the polysemous Hindi target word taken from Hindi WordNet to perform the overlapping. The sense which got the highest sense score is assigned as an appropriate sense of the ambiguous word of Hindi language. Domain-specific word sense disambiguation was done by Kharpra et al. [12] in three languages like English, Marathi, and Hindi. They used main senses of the target word in a particular domain. Singh et al. reported the consequences of doing stemming, elimination of stop words and also the dimension of context window on overlap based Lesk algorithm in WSD of Hindi language and they recorded 9.24% of improvement in accuracy. R. Sawhney et al. [18] proposed a modified WSD algorithm for Hindi language. They used Lesk technique to disambiguate the ambiguous word. A. R. Pal et al. [20] presented a knowledge based method for WSD of Bengali language. They used Bengali wordnet as the knowledge base and achieved an accuracy of 75%. A survey on word sense disambiguation [19] is referred to understand the merits and demerits of supervised and unsupervised approaches for WSD.

In this paper, first, we describe the original Lesk technique for WSD, followed by our WSD algorithm. Next, we present Dataset used in the experiment, followed by the experimental results and discussion on the same. In the end, we summarize with a discussion on the outcomes, and also on recommendations for future enhancements.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Lesk Algorithm for WSD

The original version of Lesk algorithm disambiguates the target word in small phrases. Here the target word is disambiguated by comparing gloss of each sense of a target word with the glosses of each and every remaining word in the sentence. The sense allotted to the target word whose definition has maximum words in common with the other remaining words definitions. For example, on the words pine cone, this technique finds out that these words have poly senses.

*Senses of the word **pine:***
1st sense: Kind of evergreen resinous coniferous shrub the family Pinus, with lengthy needle-shaped leaves.
2nd sense: Suffering an emotional and also physically weaken, particularly due to a broken heart.

*Senses of the word **cone:***
1st sense: A solid or hollow object which tapers from a circular or roughly circular base to a point.
2nd sense: One of two kinds of light-sensitive cell in the retina of the eye.
3rd sense: A fruit of some evergreen tree.

After comparing two senses of the word pine with three senses of the word cone separately it is identified that the words *evergreen tree* common in one sense of both the words. Then both of these senses are considered as the suitable senses of pine and cone whenever they used collectively.

3. WSD Algorithm

In this paper, we implemented the Lesk idea in three approaches. In the first approach, we find the common words present in the sentence which contains the target word and the gloss of the target word. The amount of common words present is used to decide the correct sense of the target word.
The second approach is identical to the first approach; here we added the example texts provided by the WordNet to the glosses. As in first approach amount of common words present is calculated to assign the correct sense to the target word.
In the third approach, we extended the concept of direct overlap technique for performing disambiguation. This approach uses an illustration or instance where only one target word occurs in the input sentence. A window containing a target word and few words surrounding the target word on the left and right side of the target word is defined. In this algorithm, we have considered N to the left side and another N words to the right side of the target word. These words are considered in the window only if they are present in the Indo – WordNet database. If any surrounding word is not present in the Indo – WordNet, then that word is ignored and next word is considered as the member of the context window. Before searching in Indo – WordNet, the word is converted to its equivalent root word using Stemmer or Lemmatizer, for this process we are using Kannada stemmer and lemmatizer designed by us in our earlier work [17]. Total size of the context window is 2N+1 including one target word. In our experiment, we have considered the window size of 3 Indo – WordNet words. In this approach, the comparison is done between the concepts of every pair of words present in the context window. During the comparison concept of each sense of the target, the word is considered. The amount of common words present in both the strings is considered as the score. Scores of each comparison are added with respect to the senses of the target word. The sense with the highest score is considered as the correct sense of the target word.

4. Dataset

All the three approaches presented in this paper are implemented and evaluated on the dataset containing 10 Kannada polysemous words shown in Table 1. These words and their concepts are taken from Indo – WordNet. Instances were collected from various sources like the Kannada corpus created by ILCI phase – II [16] project, from different Kannada websites and also from Google. While collecting these instances we have considered different domains like sports, literature, news, and so on. The transliteration and translation of the dataset are also shown in Table 1.

Table 1: Dataset

Words	Transliteration	Translation	Number of Senses
ಬೆಳೆ, ವರ್ಷ, ಯಂತ್ರ, ಗುರು, ಉತ್ತರ, ಚರಕ	Bele,Varsha,yantra,guru, uttara,charakha	Corp, year, machine, Teacher, answer, spinning wheel	2
ತೆರೆ, ಕಡಿತ	There, kaditha	Open or screen, itching or bite	3
ಮೂಲ	Moola	Basic reason or fundamental	4
ವರ್ಗ	Varga	Square, group, class, transfer	6

5. Results and Discussion

We have evaluated the proposed algorithm by conducting the test for target words listed in Table 1. The result of the proposed technique is shown below, where ವರ್ಗ (varga) is the target word. Various senses associated with the target word ವರ್ಗ are shown in Table 2. We have considered three instances which contains the target word ವರ್ಗ.

Sentence 1: ಏಳರ ವರ್ಗ ನಲವತ್ತೊಂಭತ್ತು

Sentence 2: ನನ್ನ ಬ್ಯಾಂಕ್ ಖಾತೆಯ ವರ್ಗಾವಣೆಗಾಗಿ ಅರ್ಜಿ ನೀಡಿದ್ದೇನೆ

Sentence 3: ವರ್ಗದ ವಿಸ್ತೀರ್ಣ ಕಂಡುಹಿಡಿಯುವ ಸೂತ್ರ ಉದ್ದ ಗುಣಿಸು ಅಗಲ

There are six senses associated with the target word ವರ್ಗ. Glosses and corresponding example sentences of these senses are taken from the Indo – WordNet are shown in Table 2. Scores of the three approaches are shown in Table 3.

6. Conclusion and Future work

Here we have used Lesk measure semantic relatedness in the three approaches to disambiguate the ambiguous word. In the first approach, we made the comparison between the input sentence which contains the target word and the gloss of the target word. In the second approach both gloss and the example statement of the target, the word is considered for the comparison with the input sentence. In the third approach to improving the accuracy, we have defined the context window and we have compared the gloss of all the words in the window with each gloss of the target word. In all the three approaches we have calculated the amount of common words present and correct sense is the one which is having a maximum score is assigned to the target word.
In the future, we can increase the accuracy of the disambiguation process by considering the glosses of all the synsets, and glosses

of the relations like hypernym, a hyponym of the target word for comparison.

Table 2: Senses gloss in Kannada and English language

Sense number	Sense in Kannada	Sense in English
1	ರಕ್ತಸಂಬಂಧದ ಏಕತೆಯನ್ನುಳ್ಳ ವರ್ಗ ಅಥವಾ ಸಮೂಹ	People descended from a common ancestor
2	ಅಧಿಕಾರಿ ಅಥವಾ ಕಾರ್ಯಕರ್ತರ ಒಂದು ಸ್ಥಾನ ಅಥವಾ ವಿಭಾಗದಿಂದ ಬೇರೆ ಸ್ಥಾನ ಅಥವಾ ವಿಭಾಗಕ್ಕೆ ವರ್ಗ ಮಾಡುವ ಕಳುಹಿಸುವ ಕ್ರಿಯೆ	The act of transferring something from one form to another.
3	ಯಾವುದೇ ಸಂಖ್ಯೆಯನ್ನು ಅದೇ ಸಂಖ್ಯೆಯಿಂದ ಗುಣಿಸಿದಾಗ ಬರುವ ಮೊತ್ತ	The product of two equal terms.
4	ಯಾವುದೋ ಒಂದು ಆಕೃತಿಯ ಉದ್ದ, ಅಗಲ ಮತ್ತು ನಾಲ್ಕು ಕಡೆಯೂ ಒಂದೇ ಸಮವಾಗಿ ಇರುವ	(Geometry) a plane rectangle with four equal sides and four right angles.
5	ಆಸ್ತಿ, ಹಕ್ಕು ಮೊದಲಾದವುಗಳ ವರ್ಗಾವಣೆ ಮಾಡುವ ಕ್ರಿಯೆ	The act of transferring something from one form to another.
6	ಒಂದೇ ಬಾರಿಗೆ ಡಿಗ್ರಿ ಪಡೆಯುವ ವಿದ್ಯಾರ್ಥಿಗಳ ಸಮೂಹ	A body of students who graduate together.

Table 3: Sense Scores

	Sense1score	Sense2score	Sense3score	Sense4score	Sense5score	Sense6score	Remarks
1st approach							
Sentence1	1	0	0	0	0	0	Correct sense not assigned
Sentence2	0	0	0	0	2	0	Correct sense is assigned
Sentence3	1	1	1	2	0	0	Correct sense is assigned
2nd approach							
Sentence1	1	0	3	1	0	0	Correct sense is assigned
Sentence2	0	1	0	0	4	0	Correct sense is assigned
Sentence3	0	0	1	3	0	0	Correct sense is assigned
3rd approach							
Sentence1	0	1	1	1	0	0	Correct sense not assigned
Sentence2	0	0	0	0	0	0	Correct sense not assigned
Sentence3	0	1	1	2	0	1	Correct sense is assigned

References

[1] M.Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In proceedings of SIGDOC '86, 1986.

[2] HindiWord-Net,http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

[3] Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics 32(1), 13–47 (2006)

[4] Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 305–332 (1998)

[5] Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 265–283 (1998)

[6] Lin, D.: Using syntactic dependency as a local context to resolve word sense ambiguity. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 64–71 (1997)

[7] Resnik, P.: Using information content to evaluate semantic similarity in taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, pp. 448–453 (1995)

[8] Reddy, S., Inumella, A., Singh, N., Sangal, R.: Hindi Semantic Category Labeling using Semantic Relatedness Measures. In: Proceedings of Global WordNet Conference (2010)

[9] Sinha, R., Mihalcea, R.: Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. In: Proceedings of the International Conference on Semantic Computing, ICSC 2007, pp. 363–369 (2007)

[10] Torres, S., Gelbukh, A.: Comparing Similarity Measures for Original WSD Lesk Algorithm Advances in Computer Science and Applications. Research in Computing Science 43, 155–166 (2009)

[11] Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings on International Conference on Research in Computational Linguistics, Taiwan, (1997)

[12] Khapra, M., Bhattacharyya, P., Chauhan, S., Nair, S., Sharma, A.: Domain Specific Iterative Word Sense Disambiguation in a Multilingual Set-ting. In: Proceedings of International Conference on NLP (ICON 2008), Pune, India (2008)

[13] Singh, S., Siddiqui, T.J.: Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation. In: Proceedings of the International Conference on Information Retrieval & Knowledge Management, CAMP 2012, Malaysia, pp. 1–5 (2012)

[14] Sinha, M., Kumar, M., Pande, P., Kashyap, L., Bhattacharyya, P.: Hindi Word Sense Disambiguation. In: International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India (2004)

[15] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using WordNet. InInternational Conference on Intelligent Text Processing and Computational Linguistics 2002 Feb 17 (pp. 136-145). Springer, Berlin, Heidelberg.

[16] http://www.tdil-dc.in/index.php?option=com_download&task=showDetails&toolid=1884&lang=en

[17] B.H.ManjunathaKumar,Dr.M.Siddappa,Dr.J.Prakash "Unsupervised Kannada Stemmer using Partial Lemmatizer and Indo – Word Net" Vol. 7 - Issue 12 (December - 2017), International Journal of Engineering Research and Applications (IJERA) , ISSN: 2248-9622 , www.ijera.com

[18] R. Sawhney and A. Kaur, "A modified technique for Word Sense Disambiguation using Lesk algorithm in Hindi language," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014, pp. 2745-2749. doi: 10.

[19] B. H. Manjunatha Kumar, "A Survey on Word Sense Disambiguation", Language in India www.languageinindia.com ISSN 1930-2940 18:2 February 2018

[20] A. R. Pal, D. Saha and S. K. Naskar, "Word sense disambiguation in Bengali: A knowledge based approach using Bengali WordNet," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2017, pp. 1-5. doi: 10.1109/ICECCT.2017.8117900