International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Query Expansion using Semantic Network for Information Retrieval in Telugu Language

Neeraja Koppula, B. Padmaja Rani, Koppula Srinivas Rao

Abstract*: Now-a-days digital documents are playing a major role in all the areas /web, as such all the information is digitalised. Queries are used by the search engines to retrieve the information. Query plays a major role in information retrieval system, as a result relevant and non relevant documents are retrieved. Query expansion techniques will better the performance of the information retrieval system. Our proposed query expansion technique is Word Sense Disambiguation. This is to find the correct sense of the ambiguous word in regional Telugu language. In Query expansion, if the added query term is an ambiguous word, accuracy of relevant documents will be very less. So to avoid this, proposed method Word Sense Disambiguation (WSD) is used, which is related to NLP Natural Language Processing and Artificial Intelligence AI. WSD improves the accuracy of information retrieval system.

Index Terms*: Query Expansion (QE), Informational Retrieval System (IRS), ambiguous word, Word Sense Disambiguation (WSD), Natural Language Processing (NLP), Artificial Intelligence (AI), Telugu Language.

I. INTRODUCTION

In Recent years, all documents are digitalized. Retrieving of information from these digital documents is termed as information retrieval system. In Informational Retrieval System, accuracy purely depends on the query. Some techniques are used to expand the query term, replacing query term by synonyms, automatic spelling error correction, and morphological forms and reweighting the terms. This process is known as Query Expansion. Sense determination is a recent trend in information retrieval system. In query expansion using the query term ambiguous word, the sense of the ambiguous word (query term) is expanded so that only relevant documents are retrieved. This increases the accuracy of the information retrieval system. The process of sense determination is termed known as (WSD) Word Sense Disambiguation System. Our proposed method is query expansion, determining the sense of the query term for information retrieval on Telugu Text corpus. A query added term is a ambiguous word, a word having multiple senses in different context. A Query is reformulated by adding the appropriate meaning of the query term in the given context. Finding the sense of the query term in a given context is known as word sense disambiguation. It is an upcoming technology in Natural Language Processing. WSD is an open challenge in Artificial Intelligence, especially in regional Telugu Language.

- Query Expansion mechanism improves the precision values of the Information retrieval system.
- Given an input query, the top K documents are retrieved, and these documents are divided into relevant and irrelevant documents in regional Telugu language.
- Reformulate the query such that the query term senses are used to retrieve the relevant documents to increase the performance of the Information Retrieval system
- For this process of query expansion, our proposed system, query expansion methodology is to find the correct sense of added query term. This is known as word sense disambiguation (WSD).

II. LITERATURE SURVEY

A. *Query Expansion Techniques*

Telugu is the Indian language very rich in morphology compared with other south Indian languages. Due to this reason maximum no of words will be having more than one sense. Sense identification is very important for Indian languages to improve the information retrieval results where different query expansion methods are applied. In information retrieval system for Indian languages recent trend is Query expansion this improve the search results where different methods are being proposed to expand the root query. Query reformulation process is to find an appropriate query term which is added to the old query. One of them is given a query, searching for the senses of the ambiguous (disambiguated) words in the given query terms which expands the query automatically. Enriching a user's query with senses of ambiguous word can improve search performance in a text retrieval system. "A word can have more than one meaning "called as ambiguous words. In this context word and sense relationships must be taken into consideration to reformulate root query in regional Telugu language. Sense determination is a recent trend in information retrieval systems, information extraction and machine translation.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Mrs. K. Neeraja, CSE, ML RIT, Hyderabad, India
Dr. B. Padmaja Rani, CSE, JNTUH, Hyderabad, India.
Dr. K. Srinivas Rao, CSE, MLRIT, Hyderabad, India.

The tasks like language understanding, language parsing, machine translation, in addition to information retrieval, question answering and summarization requires disambiguating the multiple meanings of a word in different contexts. To identify the correct meaning of a word in a sentence or a paragraph is referred to as word sense disambiguation (WSD). It is the crucial step in natural language understanding which followed by parts-of-speech (POS) tagging as meaning exists only for nouns, verbs, adjectives, and adverbs.

B. *Word Sense Disambiguation*

Approaches to WSD are classified as supervised, unsupervised, and knowledge-based methods. In Telugu Language word sense disambiguation is at infant level. The research work in Telugu Language is at nascent stage. Supervised approaches train a statistical model to assign a concept (sense) to a target term based on its context. There are three approaches: knowledge-based, supervised and unsupervised are used for word sense disambiguation. Knowledge-based approaches needs machine readable sources such as dictionaries, sense inventories, LKB Lexical Knowledge Base, thesauri and Word Net etc and uses techniques like maximum number of gloss overlaps, semantic similarity, selection preferences and heuristics. The supervised WSD uses machine learning techniques on manually created sense-annotated data divided into train and test data. These methods assign meaning to the unknown word. The machine learning algorithms include decision tress, neural networks, support vector machines, naive bayes, instance based learning and combining the methods (Ensemble) using voting technique, probability mixtures, ranks and Ada boost. The unsupervised WSD discriminate the word meanings without assigning meaning to the words and thus did not require annotated corpus. Context clustering, word clustering, co-occurrence graph and spanning tree based techniques fall under unsupervised WSD. The supervised approaches produce superior performance compared to knowledge-based which are better compared to unsupervised methods. So far reasonable works reported for English and European languages but countable works reported in Indic languages especially in Telugu. The main reasons are: the lack of publicly available Word Net resources and morphological inflections.

III. PROPOSED MODEL

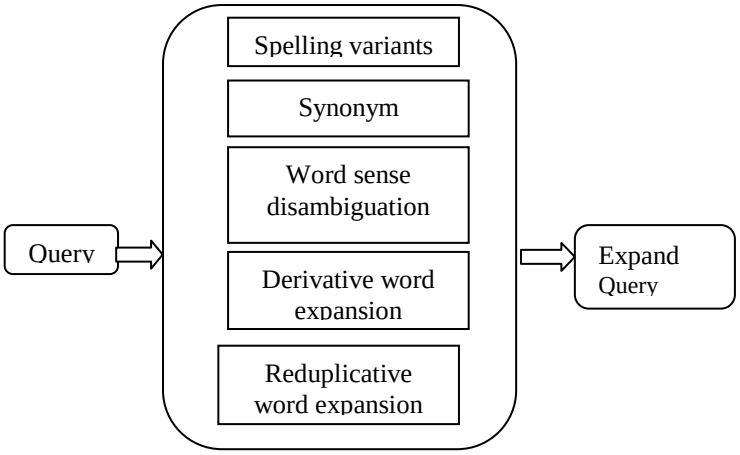

Figure 1: Query Expansion Techniques

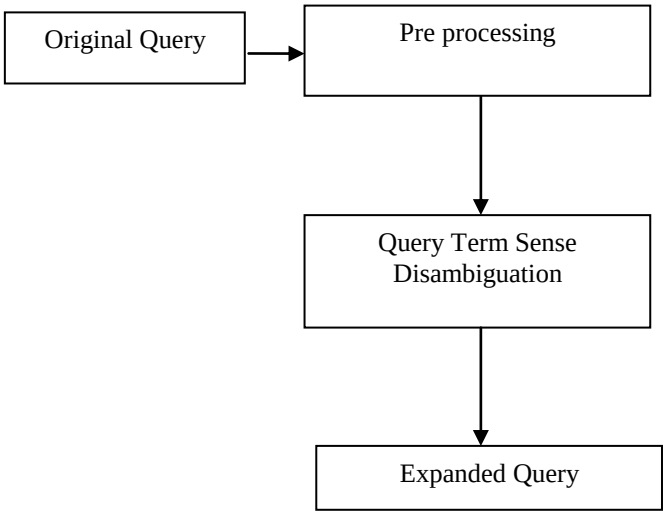

Figure 2: Proposed Model

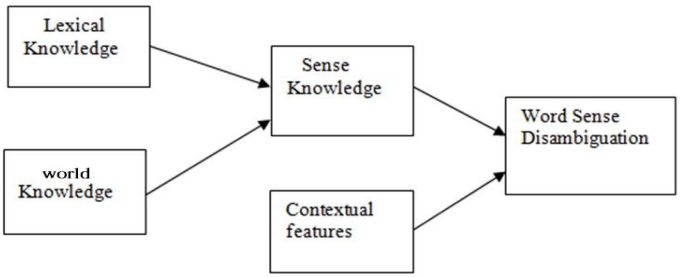

Figure 3: Query Term Sense Disambiguation Model

IV. PROPOSED METHODOLOGY

Our proposed method performance is better when compared with other query expansion techniques in information retrieval process. The query is expanded by replacing the sense of the query term. Proposed algorithm is to determine the sense of the query term in the given context, which improves precision of the system.

A. *Word Total Sense Score WTSS*

This method is Context Dependent. This is purely based on input context words, surrounding words of a target word are matched with the context words of a particular sense of a target word in LKB With proximity of 2 to 3 surrounding words. For each target word extract all the senses from LKB and for each sense extract all the context words, matching is done between the input context words and the LKB context words of each sense of the target word. For which sense maximum matching is obtained that sense is treated as an appropriate sense. The Methodology is, Input sentence is the test sentence with target ambiguous word and Output is the appropriate sense of an ambiguous word. Read input Query Sentence IS, Next is the Pre-processing stage, the removal of stop words and Stemming of the query sentence. All the words of the input query sentence are assigned with index value Wi, i ranges from 1 to n.

Now check for ambiguous word in the query. Extract each ambiguous word Wj from the LKB and compare it with the actual input context words. For each word Wj in PSW compare with input context word Wi, if both are equal, target ambiguous word is assigned with Wi, else add Wi to the context words. Now extract all the senses of a target ambiguous word of the query sentence from LKB Lexical Knowledge Base. For each sense Sk of a target ambiguous word assign Skscore ← 0. Extract all context words of the particular sense from LKB which is in PCSk list (with all senses and their context words). Compare each word Wi of input context words with context words of the sense stored in PCSk, if the matching is successful, increment Skscore with 1. Repeat the above process with until all input context words are compared with all context words of a particular sense in LKB. Now compare Skscore with score, if Skscore is greater than score, assign score with Skscore and Sense with Sk, then Output the sense.

B. *Algorithm*

Input: Query with target ambiguous word.
Output: Appropriate sense of ambiguous word
1. Read input sentence IS
Target word ← null, Context word ← null
Score ← 0 , Sense ← null
2. Pre-processing stage
a. Removal of stop words
b. Stemming
3. For each Wj in PSW
If (Wi==Wj)
Target word ← Wi
Else
Add Wi to context words
4. Repeat from step 4 to step 8 for all senses of query term which is an ambiguous word
Extract each sense Sk of a target word
Skscore ← 0
5. Extract context word list PCSk of sense k from LKB
6. Comparison of LKB context words and input context words
For each word Wi of input context words
If (Wi in PCSk)
Skscore ← Skscore + 1
7. If (Skscore > score)
Score ← Skscore
8. Sense ← Sk

V. EVALUATION

In our proposed system of query expansion in information retrieval system is precision and recall values. F-measure is the main evaluation in WSD systems. This is used to evaluate the performance of the Information retrieval system using WSD system. The accuracy of WSD system is measurable by the F measure, precision and recall, which values are calculated using the formulae. Precision is ratio of relevant items retrieved to the retrieved items. Recall is ratio of relevant items retrieved to the relevant items

Precision specifies the performance of the information retrieval system.

A. *LKB Lexical Knowledge Base*

For Telugu Language there is no Standard Word Net. Training Data is generated from తెలుగు సాహిత్యం, తెలుగు నవలలు. Training Datasets: A huge dataset with 1500 documents categorized as follow News papers, Medical field, క్రీడలు, కధలు, దేవాలయములు, రాజకీయలు, కవితలు, వేమన పద్యాలు, తెలుగు సాహిత్యం, తెలుగు నవలలు, on line e-paper. The above documents 70% are used for training phase and 30% are used for testing phase. The collected documents are with polysemous words. By using the training datasets, the context words for each sense of the polysemous word are generated and stored in LKB. For each polysemous word the average numbers of senses are three. For each sense generate the context words from training data. Context Words are generated for each polysemous word from training data depending on the senses. LKB Lexical Knowledge Base consists of Telugu dictionary, Telugu Training Data, Polysemy words with senses, Polysemy word senses with context words.

B. *Results*

Our proposed method performs better particularly for Nouns. Sample test data

Table 1 Precision and Recall Values

Domain	Precision	Recall
Sports	69%	65%
Novels	71%	68%
Medical	60%	54%
Temples	65%	61%

The Proposed method is evaluated using precision and revall parameters. For Sports domain, our proposed model achieved 69 % of precision and 65% of recall, 71% of precision, 68% of recall obtained for novels domain. For Medical domain our WSD method generated 60% precision and 54% of recall and for temples domain our proposed method achieved 65% of precision and 61% of recall.

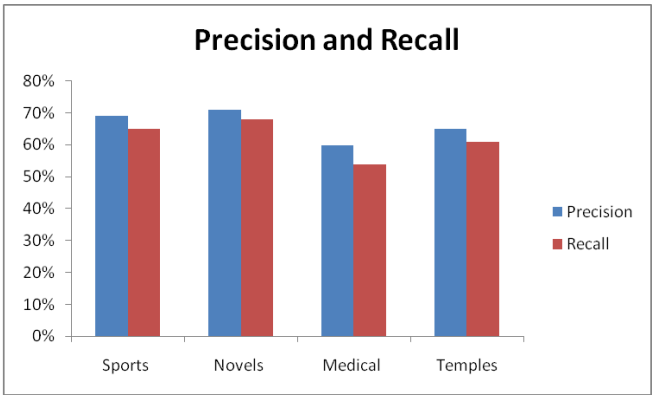

Figure 4 Performance Of The Proposed Method

VI. CONCLUSION AND FUTURE WORK

In this work, query expansion for informational retrieval process in regional Telugu Language.

In our proposed method query reformulation is performed by finding the correct meaning of the added query term in a given context, the system is tuned for noun disambiguation of the query term; the proposed algorithms for word sense disambiguation using knowledge-based approach are categorized into context independent and context dependent algorithms. The accuracy of the context dependent algorithms is more when compared with context independent algorithms. In this work we compared all the four proposed approaches with precision and recall values which are used to calculate accuracy. Among all the proposed approaches graph based word sense disambiguation accuracy is more. In this approach the performance factors are context size, number of iterations and damping factors. Word sense disambiguation in Telugu language has more scope than compared to any other regional Indian language. Future work, word sense disambiguation system for Telugu language can be developed using supervised and unsupervised approaches.

VII. ACKNOWLEDGEMENT

We would like to thank the reviewers, who greatly helped to make this article in better shape. We extend our thanks to the management of MLR Institute of Technology for providing excellent infrastructure to complete this research work. We would further more like to extend our thanks to the research and development team for continuous support

REFERENCES

1. E. Agirre, O. L. De Lacalle, And A. Soroa, “Random walks for knowledge-based word sense disambiguation, “Computational Linguistics, vol. 40, no. 1, pp. 57–84, 2014.
2. R. NAVIGLI, “Word sense disambiguation: A survey,” ACM Computing Surveys (CSUR), vol. 41, no. 2, p. 10, 2009
3. Agirre, E. And Soroa, A. (2009). Categorized Page Rank for Word Sense Disambiguation, in the proceedings of EACL-09, Athens, Greece.
4. Arindam Chatterjee, Salil Joshii, Pushpak Bhattacharyya, Diptesh Kanojia And Akhlesh Meena, “A Study of the Sense Annotation Process: Man v/s Machine”, International Conference on Global Wordnets , Matsue, Japan,, Jan, 2012.
5. BOSHRA F. ZOPONAL_BAYATY AND DR.SHASHANK JOSHI “Word Sense Disambiguation (WSD) and Information Retrieval (IR): Literature Re-view” ijarcsse, Volume 4, Issue 2, ISSN: 2277 128X, February 2014.
6. S. P. Ponzetto, R. Navigli, "Knowledge-rich Word Sense Disambiguation rivaling supervised systems,” Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1522-1531, 2010.
7. Rada Mihalcea, "Knowledge Based Methods for WSD", e-ISBN 978-1-4020-4809-2, Springer 2007.
8. ROBERTO NAVIGLI AND PAOLA VELARDI, "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 27, No. 7, pp. 1075-1086, July 2005.
9. Neeraja, Dr. B. Padmaja Rani, Dr. Koppula Srinivas Rao, Hybrid approaches for Word Sense Disambiguation: A Survey, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 23 (2015) pp 43891-43895.
10. Alok Ranjan Pal, Diganta Saha, Antara Pal, A Knowledge based Methodology for Word Sense Disambiguation for Low Resource Language. Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 2 (2017) pp. 267-283 © Research India Publications http://www.ripublication.com
11. NEERAJA KOPPULA AND DR. B.PADMAJA RANI , Word Sense Disambiguation using Knowledge Based approach in Regional Language journal of advanced research in dynamical and control system Issue :05-special Issue Year 2018 Pages :109-111.
12. A Knowledge-Based Approach to Word Sense Disambiguation by distributional selection and semantic features . Mokhtar Billami (LIF)
13. M. S. Nameh, M. Fakhrahmad And M.Z. Jahromi, “A New Approach to Word Sense Disambiguation Based on Context Similarity,” Proceedings of the World Congress on Engineering, Vol. I, 2011.
14. Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, B. Slator, Providing machine tractable dictionary tools, Machine Translation 5, 99–154, 1990.
15. Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya, , Hindi Word Sense Disambiguation, In-ternational Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, Novem-ber, 2004 http://www.cse.iitb.ac.in/~pb/papers/HindiWSD.pdf
16. M. Quillian, Semantic memory, in: M. Minsky (Ed.), Semantic Infor-mation Processing, the MIT Press, Cambridge, MA, pp. 227–270, 1968.
17. J. Cowie, J. Guthrie, L. Guthrie, Lexical disambiguation using simu-lated annealing, in: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, pp. 359–365,1992.
18. H. Kozima, T. Furugori, Similarity between words computed by spreading activation on an english dictionary, in: Proceedings of the 6th Conference of the European Chapter of the Association for Com-putational Linguistics, Utrecht, pp. 232–239,1993.
19. J. Veronis, N. Ide, Word sense disambiguation with very large neural networks extracted from machine readable dictionaries, in: Proceed-ings of the 13th International Conference on Computational Linguis-tics, Helsinki, pp. 389–394,1990.

AUTHORS PROFILE

Mrs. K. Neeraja obtained B.tech(CSE) from JNTU, Hyderabad, M.Tech(CSE) from, JNTU, Hyderabad. She is currently pursuing Ph.D in JNTUH. Research area is information retrieval system using word sense disambiguation.

Dr. B. Padmaja Rani obtained B.Tech. (Electronics and Communication Engg.) From Osmania University, Hyderabad, M.Tech. (Computer Science) from Jawarharlal Nehru Technological University, Hyderabad and Ph.D (Computer Science and Engineering) from Jawarharlal Nehru Technological University, Hyderabad.. Her research interests include Information Retrieval, Embedded Systems, Natural Language processing, Data Leakage Prevention, Big Data, Information Security, Cloud Computing. She has almost 60 publications in national and international journals or conferences. At present she is the Coordinator, TEQIP-III of JNTUH University.

Dr. K Srinivas Rao has received B.E from osmania **University**, M. Tech from JNTU, Hyderabad and PhD from Anna University, Tamilnadu. He is currently working as a principal of MLRIT, Hyderabad. He is having 20+ years of professional experience and 10+ years of research experience. His area of interest includes Information Retrieval, Natural Language Processing, Information Security , Big data and Cloud Computing, Dataware house and data mining.