Text mining approach for WSI in QA systems through relation construction

[1]C. Meenakshi, [2]P. Thangaraj, [3]M. Ramasamy

[1] Ph.d Scholar
Mother Terasa Women's University, Kodaikanal

[2] Dept. of CSE,
Bannari Amman Institute of Technology,
Sathyamangalam

[3] Dept. of EEE,
Annamalai University,
Chidambaram

Abstract

The paper attempts to identify the exact sense of the word through a process of establishing relations between similar words in the sample ontology. The philosophy relates to analysing the variance in sense among similar words with the help of a data mining tool to arrive at the exact sense of the target word as part of a question answering system. The focus spreads to derive new information from data and create patterns across datasets. The semantic relation among the words is calculated by assigning weights to investigate the variance and land at the exact sense. The distribution factor accrued from the manipulation of assigned weights reveals the significance of training to arrive at a better grading. The performance evaluated in terms of precision and accuracy exhibits the suitability of the proposed approach for use in the real world applications.
Keywords: *word sense identification, data mining, word relations, ontology*

1. Introduction

Word sense identification continues to remain as a key task in the fields of information extraction, machine translation and similar related areas. It espouses a direct need for machines to trace the correct sense through comprehending the context. The major portion of this problem relies on the surrounding context and examining the properties exhibited in that context. The widespread prevalence of ambiguity further demands refined approaches to address the issues.

The identification of word senses conceives an interesting area of concern in the computer treatment of language and endeavours to accomplish most natural language processing tasks. It reflects its help in instances where the aim does not enter into understanding the language. The inherent difficulty of sense identification exists especially in cases where the sense of the word does not seem implied in the sentence [1].

Homonymy and polysemy are concerned with lexical ambiguity. Homonyms are words that are written the same way, but are (historically or conceptually) really two different words with different meanings which seem unrelated [6][13]. The examples are *bill* ("a written statement of money owed" and "bird's beak") and *fair* ("attractive" and "market"). If a word's meanings are related, it is called a polyseme. The word *crane* is polysemous because its senses can be generalized as "strain out one's neck for fetching something", either a bird or construction equipment, that is they are related [5].

The identification of the word sense permeates as an open problem in Natural Language Processing and the inferencing mechanism among others include tasks related to discourse reference resolution and coherence [2]. It enforces its impact in most tasks that require a certain degree of semantic interpretation that extends as information extraction, intelligent human-computer interface, and question answering. The traditional methods may not subscribe to the needs of the present day world in the sense identification is conceptual and pertains to a large number of domains. It is in this perspective that there evokes a refined approach to improve the stream and arrive at a much closer identification of the word sense.

2. Literature Review

A statistical method has been suggested for assigning senses to words through christening procedure about the context and the error rate in machine translation system found to decrease using statistical methods [3]. The results of word sense scheme generated from a concept map has been found to enhance understanding of the concept and sort the meaning of the word [12]. A neural network based algorithm suitable for very large corpora has been developed to address the needs of real time speech recogniser [7].

A supervised learning method based on K-Nearest Neighbor algorithm has been developed for identifying the sense of a word [8]. It has been found to

involve the extraction of features from the set of words that occur frequently in the text and the set of words surrounding the ambiguous word [9][11]. The unsupervised methods have been found to induce word senses from input text by clustering word occurrences, and follow it up with classifying new occurrences into the induced clusters [5]. It has been relegated to be useful especially in the absence of neighbouring words with the same sense[10]. An input model of Neural Network that calculates the Mutual Information between contextual words and ambiguous word by using a statistical method has been suggested to the process of sense identification.

3. Problem Statement

The main theory echoes to predict the sense of a word through the use of a data mining tool in a question answering system. It bestows a procedure using which it forges a relationship between words and formulate a procedure to investigate the variance in sense within them. The strategy avails the data base from ontology and excerpts the use of appropriate weights to facilitate the process of training. The learning methodology incorporated in the scheme serves to yield the exact sense of the target word.

4. Proposed Methodology

The system eschews a process of learning using a training set and acquires knowledge to resolve on its own the sense of a word in the process of answering. It draws a sequence of word relations with the aid of the ontology ware house and evolves the artefacts of data mining to extinct at the correct sense. The goal of data mining envisages retrieving relevant patterns to facilitate the process of identifying the sense of the word.

The emergence of intelligent methodologies foray their role through map based approaches and the training involved suits them to address reasoning related problems. There arises an exhaustive interest in the identification and automatic extraction of the attitudes, opinions and feelings expressed in texts. It evokes a need to provide tools for users of different domains which require for different reasons the automatic monitoring of information that expresses senses. The pertinent focus endeavours a system that can eliminate the effort to manually extract useful knowledge from the information available on the Word Net.

Most approaches to sense identification rely on lexicons of words to express the context and find it difficult to distinguish between different senses of a word. However, many keywords include both subjective and objective senses and even the purely subjective senses inherit a degree of positivity or negativity, depending on the context where the corresponding word appears.

The data mining techniques are built upon statistical ideas and there arise tools that emerge out of the basic ideas of probability, independence and causality. The popular tools that relate to identification of sense in a word are ancova and anova which intercepts from the analysis of variance and covariance respectively. The anova excerpts to reveal the main and interface effects of independent variables on dependent variables. The model elucidates the effect of interaction and the key statistic is the F-test of the difference of group means. It tests if the group means formed by values of independent variable are different by a large measure to eliminate its occurrence by chance.

The data in form of weights are organised, modelled and analysed with the statistical tool, ANOVA. The method assigns to each word the data, weight that bears the maximum estimated probability of occurring in the context.

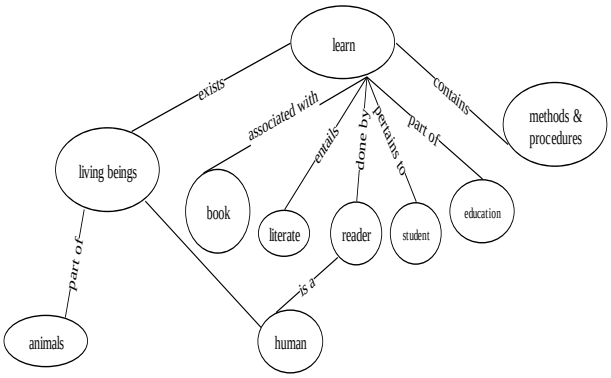

Fig. 1 Word sense relation set

The constructed relationship set expressed through Fig. 1based on wordnet after the stemming process is projected according to the target word for which the sense is to be identified and the correlation of the senses is mapped as exact sense as seen from the flow diagram in Fig. 2.

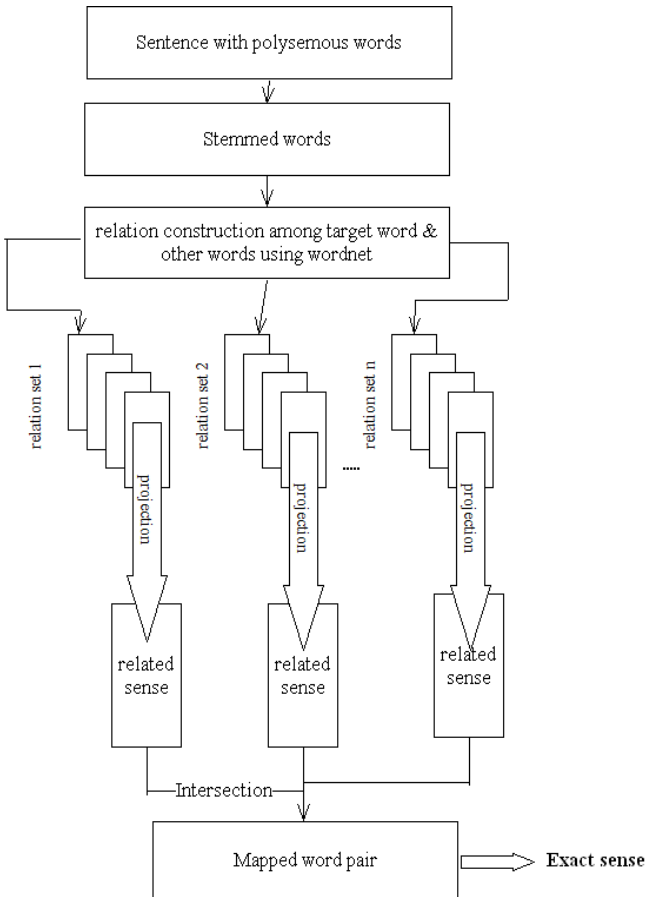

Fig. 2 Flow diagram

4. Simulation Results

The strategy allures to investigate to the performance for five words namely book, mark, pen, stress and ruler from a sample set of eight words seen in Table.1. through the use of the data mining procedure and two other similar methods degan method and yarowsky to bring out the relative merits of the proposed approach. The performance plot in Fig.3 explains the artefacts of the three mechanisms in the process of sense identification. The results measured through the precision and recall graph for the chosen test five words exhibit a higher degree for the data mining formulation as observed from Figs 4 and 5.

Table. 1. Database with word relations

S.No	Word	Sense Equivalents
1	BOOK	[a written work, composition, publish, print, page, bound, script, sheets, ledger, record, registry Holy book, Bible, Quran, Gita] [reserve, hold, register]
2	MARK	[grade, score] [distinguishing symbol, a reference point, impression created] [brand, stain] [punctuation, sign] [Apostle and companion of Saint Peter]
3	PLAY	[role, skit, drama, maneuver, turn] [act, bring, work, run, take on]
4	READER	[person who enjoys reading] [proof reader] [person reading lessons in church service] [designation, referee]
5	NOTEBOOK	[book with blank pages,etc] [pc, computer]
6	STICK	[element consisting of wood, thin branch of tree] [control stick, joystick] [threat of a penalty] [wedge, lodge, fix, bind, bond, adhere, cling, cleave]
7	SYSTEM	[arrangement, organization] [procedure/process for obtain an objective] [system of rules] [scheme, a group of related organs or parts or elements]
8	STRESS	[emphasis] [accent, emphasis, punctuate] [focus] [stain, force]

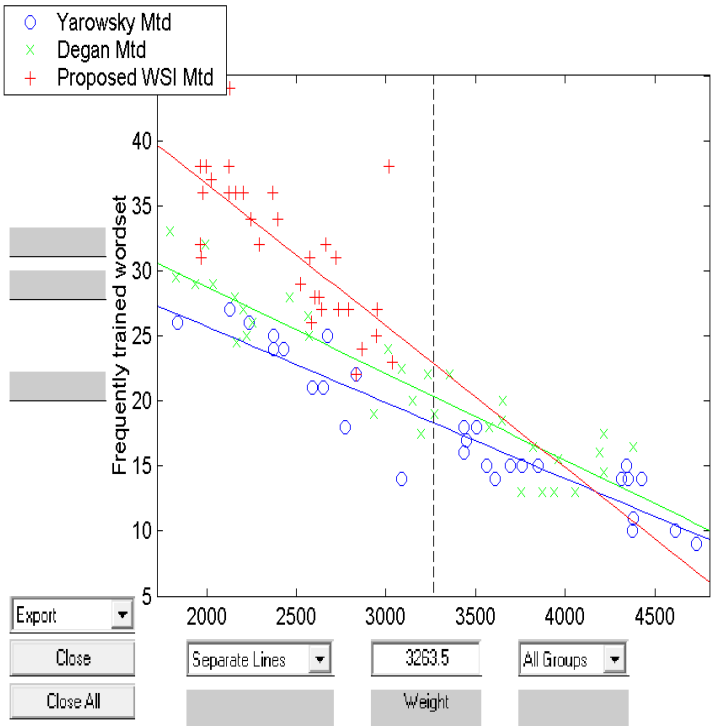

Fig. 3 Performance plot

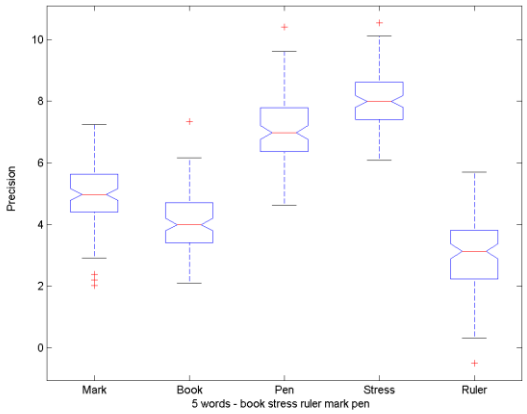

Fig. 4 Precision plot

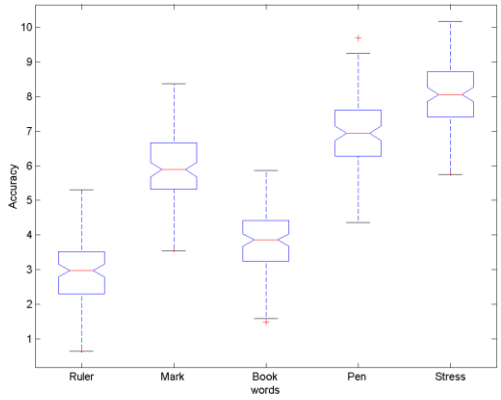

Fig. 5 Accuracy plot

5. Conclusion

A data mining algorithm has been articulated to extricate the sense of a word for use in a question answering system. The methodology has been framed for different class of words from a spectrum of domains. The data base has been obtained using the ontology and a process of weight based learning incorporated to analyse the variance in classes among the target words. The performance measures have been found to yield tall values for the data mining approach. The prediction chart has been found to foray the procedure and highlight the superior behaviour of the data mining scheme in this domain. The flexible nature of the proposed system increases its scope and will go a long way in enhancing its use in emerging fields.

References

1. Agirre, Eneko & German Rigau. 1996. "Word sense disambiguation using conceptual density", in Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark, 1996.
2. Amsler R. and Walker D. 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, 1986.
3. David. Yarowsky, 1992. "Word sense disambiguation using statistical models of Roget's categories trained on large corpora", in Proceedings of the 14th International Conference on Computational Linguistics (COLING), Nantes, France, 454- 460, 1992.
4. David. Yarowsky, 1995. "Unsupervised word sense disambiguation rivaling supervised methods", in Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, MA, 189-196,1995.
5. C. Meenakshi, P. Thangaraj and M. Ramasamy, "A Novel Scheme to Identify the Word Sense in Question Answering Systems",International Journal of Computer Science and Telecommunications [Volume 2, Issue 9, December 2011]
6. Michael Lesk. 1986. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone", in Proceedings of the 5th annual international conference on Systems documentation, Toronto, Ontario, Canada, 1986.
7. Norifumi Watanabe and Shun Ishizaki "Neural Network Model forWord Sense Disambiguation Using Up/Down State and Morphoelectrotonic Transform", Journal of Advanced Computational Intelligence and Intelligent Informatics Vol.11 No.7, 2007
8. Philip Resnik, 1999."Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research, 1999.
9. Resnik P. 1995. *Disambiguating Noun Groupings with Respect to WordNet Senses*, in Proceedings of the Third Workshop on Very Large Corpora, MIT. Richardson R., Smeaton A.F. and Murphy J. 1994.
10. Richardson R., Smeaton A.F. and Murphy J. 1994." Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.
11. Sussna M. 1993. *Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network*, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia. Voorhees E. 1993.
12. Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, Chikara Hashimoto, "Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information" Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 477–485, Prague, June 2007. c 2007 Association for Computational Linguistics
13. Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. 1993. Providing Machine Tractablle Dictionary Tools, in Semantics and the Lexicon (Pustejovsky J. ed.), 341-401. Yarowsky, D. 1992. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, in proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France."

C. Meenakshi is a research scholar in the Computer Science Department. Her current research interests include Natural Language Processing, Information Retrieval and Machine translation.

Dr. P. Thangaraj is Professor & Head, Computer Science and Engineering Department in Bannari Amman Institute in Technology, Sathyamangalam. His current research focuses on Natural languages, Fuzzy theory and computational problems.