ISSN 2047-3338

A Novel Scheme to Identify the Word Sense in Question Answering Systems

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

[1]MCA Dept. Vivekanandha Institute of Engineering College for Women, Tiruchengode (T.N), India
[2]Department of CSE, Bannari Amman Institute of Engineering & Technology, Sathy (T.N), India
[3]Department of EEE, Annamalai University, Chidambaram (T.N), India
[1]meenasi.c@gmail.com; [2]ctptr@yahoo.co.in; [3]profmramasamy@gmail.com

***Abstract*— A novel approach to identify the sense of the word appearing in a sentence is proposed in this paper. The basic idea is to facilitate the natural language processing through the broad manifold of unsupervised observations. The target word is tokenized and a relationship with other words is constructed through mapping to constitute the training set using specific domain wordnet. It constitutes the corpus from where similar word senses are arrived at and there from be-hives its role into the question answering system. The scheme is evaluated through precise performance indices for two specific domains to illustrate its applicability in the present day context.**

***Index Terms*— Corpus, Identifier, Semantic Relation and Word Sense Disambiguation**

I. Introduction

WSD (Word Sense Disambiguation) appears to be the most interesting and long-standing problems in Natural Language Processing (NLP). The most obvious application of WSD is machine translation, which requires understanding the source language translation and generate there from the sentences in the target language. It involves two definite stages since a word in the source language may have more than one possible translation in the target language. For example, the English word *"drug"* can be translated for its sense of *"medicine"* or for its sense of *"dope"* depending on the context. It is significant to process the text and correctly identify the sense for which it is intended in the text.

The different meanings of polysemous words are known as senses and the process of deciding which is being to be used in a particular context Word Sense Disambiguation [6]. WSD, in its broadest sense, can be considered as determining the meaning of every word in context, which appears to be a largely unconscious process in people's minds. As a computational problem it is often described as "AI-complete", that is, a problem whose solution presupposes a solution to complete NLU or common-sense reasoning [1].

The ambiguous words in the queries seem to inherit difficulties and the retrieval engines therefore need WSD for filtering out documents with senses irrelevant to the query. In speech synthesis, it is important to determine the correct pronunciations of words in order to generate speech that sounds natural. This process is difficult since there exists some words which are pronounced in more than one way depending on their meaning. For example, as mentioned in [2], *"lead"* is pronounced in one way when it is used in the sense "be in front" and in another way when it is used in the sense *"a type of metal"*. WSD augurs to help speech synthesis by identifying the correct sense of the word and provide the correct pronunciation. The reverse problem may occur in speech recognition for homophones where the words are spelled differently but pronounced in the same way. WSD offers substance even if different spellings are treated as different senses.

Computationally WSD can be seen as a classification problem, where word senses are the classes and the context provides the evidence and each occurrence of a word is assigned to one or more possible classes based on the evidence. It means that words are assumed to have a finite and discrete set of senses from a dictionary, a lexical knowledge base, or application-specific sense inventories (commonly used in MT). The fixed inventory makes the problem tractable and reduces the complexity of the problem. Some authors focus on the limitations of the fixed setting [3], [4] and argue that a more dynamic approach is appropriate to represent the word meaning in a corpus. Irrespective of the procedure that is adopted to understand the language, a robust NLU interface is desired to be able to tell which sense, among a list of senses, is intended in a given context.

It is pragmatic that algorithm based approaches are complex and hence necessitate a new approach of Word Sense Identification (WSI) which assumes lot of significance in the emerging Natural Language Understanding scenario.

II. Literature Review

Several semantic similarity measures have been proposed in the literature, all of them computing metrics on semantic nets. A few of them have been found to estimate the similarity as the minimum length between concepts [5], [6]. The notion of information content has been defined as the probability of occurrence in a large corpus and evaluated as a measure of

semantic relatedness between words by quantifying the information content of the lowest common subsumer of two concepts [8]. A formula to measure similarity between words for different PoS has been introduced by Mihalcea and Moldovan (1999) through a creation of connections through the glosses [9]. Introduce the notion of conceptual density defined as the overlap between the semantic hierarchy rooted by a given concept C, and the words in the context of C has been suggested by Agirre and Rigau [10]. Recently graph-based methods for knowledge based WSD have been found to attract the NLP community [11], [12], [13], [14], [19]. These methods have been found to exploit the structural properties of the graph underlying a particular knowledge base. The graph based WSD methods have been found to be particularly suited for disambiguating word sequences and used the interrelations among the senses in the given context.

There are other methods that have been found to rely on the explicit structure of knowledge bases. For instance, some algorithms for WSD have been tailored to orient selectional preferences as a way for extracting the possible meaning of a word in a given context [15], [16]. The preferences have been found to capture information about the possible relations between word categories and represented common sense knowledge about classes and concepts. Several other approaches have been proposed to acquire and determined the selectional preference between two concepts [17], [18]. There are also a number of methods that have been based on semantic similarity or relatedness. However new dimensions are to be explored in view of the ever growing complexities and evinces a need to arrive at a more comprehensive identification strategy.

III. PROBLEM STATEMENT

A relational type strategy, each particular to a preferential domain is formulated as a methodology to quickly identify the sense of the target word appearing in sentences that phase though a Question Answering system.

A. Proposed Strategy

The perception of distinguishing between multiple possible senses of a word is an important subtask in many NLP applications. However, despite its conceptual simplicity and it is obvious formulation as a standard classification problem, achieving high levels of performance on this task is remarkably an elusive goal. If the ambiguity is in a sentence or clause, it is called structural syntactic ambiguity where as if it is in a single word, it is called lexical semantic ambiguity.

WSD refers to the resolution of lexical semantic ambiguity and its goal is primarily to attribute the correct senses to words in a given context. Many words posses' more than one possible meaning, for example:

Bat can be a small nocturnal creature
or a piece of sports equipment
and a bank can mean the edge of a river
or a place where financial transactions are articulated

A word may have many meanings some of which are very different from each other. In fact in some cases a word can have two meanings which are the opposite of one another such as *"consult"* which can mean to ask for advice or give advice and *"clip"* which can mean to fasten or detach. It is important for an automated system to correctly determine the meaning in which a word is used, in such a circumstance.

WSD algorithms exist in corpus and knowledge based categories in accordance with the way they acquire information. In *corpus based approaches*, the information is gained through training on some corpus. A corpus provides a set of samples that enables the systems to develop some numerical models. It is further available in two subclasses such as Supervised and Unsupervised disambiguation, pivoted in the manner in which the corpus is trained. In supervised WSD the training data is sense-tagged whereas in unsupervised WSD the training data is a raw corpora and are not disambiguated.

The underlying assumption in unsupervised training is that similar senses occur in similar contexts, and thus senses can be induced from text by clustering word occurrences using some measure of similarity of context. Then, new occurrences of the word can be classified into the closest induced cluster/senses, called sense induction (Schutze, 1998).The infrequent senses and senses that imbibe few collocations are hard to isolate in unsupervised disambiguation. In general, the accuracy of unsupervised WSD systems are 5% to 10% lower than that of other algorithms since no lexical resources for training or defining senses are used. The methods correspond to clustering rather than sense tagging tasks. A completely unsupervised disambiguation may not be possible for word senses since sense tagging requires characterization of the senses.

B. Experimental Procedure

The construction of semantic relations is through an attempt in improving Lin's algorithm using semantic dependencies from the WordNet.

1) Database Creation

The corpus is constructed for two state of the art domains such as an Education system and a Sports fraternity. A relationship of the polysemous word with other words in any sentence is established using the format.

$$t_w\#\ \textbf{is part of}\ \#t_1, t_w\#\ \textbf{is a kind of}\ \#t_2, \ldots t_w\#\ \textbf{contains}\ \#t_n$$

The entries in Table given below include the wide variety of n such polysemous words in the two specified domains.

2) Constructing Relations

The first step is to arrive with tokens after which a relationship of the target word among other words is determined with the help of training corpus as evinced through the format seen above. Then the constructed relation is mapped against WordNet to identify the conceptual sense which is coined as the correct sense. The steps involved in the experiment focus to disambiguate between the fine pair, then to include more pair of distinct senses and finally to identify the correct among the five senses.

TABLE 1
DATABASE CONTENTS WITH TEST WORDS

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	STRAIN	Stress, mental strain, nervous, breed, tenor, pain, strainer, tense, tense up [puree, deform, distort, filter]
3	PLAY	[role, skit, drama, maneuver, turn] [act, bring, work, run, take on]
4	BOARD	[plank, table, panel, circuit board, blackboard, control panel, game board, display board] [get on, room, lodge and take meals, provide food]
5	PEN	[enclosure for stock] [playpen for babies] [female swan] [write, compose]
6	RING	[closed chain, pack, mob, hoop, band] [echo, peal, sound, call, call up]
7	NET	[cyberspace, internet] [catching species] [network, mesh] playing net] [net profit, final, last]
8	NOTE	[brief note, annotate, promissory note, government note, musical note, tune] [observe, mark, take note, take down]
9	TOP	[spinning top, playing top, teetotum, whirling top,] [height, peak, crown, crest, tip, summit] [garment top, cover]
10	DEAL	[trade/business deal, bargain, agreement] [consider, look at, take, cope, manage, handle, care, conduct]

3) WSI Algorithm

The proposed methodology is coined using the steps detailed 1 through 6. The sequence of flow is detailed in Fig. 1.

1. *Given $T = \{t_1, t_2, t_3..t_n\}$ is set of terms from a user question or Q&A set, to disambiguate terms:*
2. *Find the stem words through stemming algorithm module*
3. *Remove the inflection words and other stop words such as conjunctions, prepositions, etc*
4. *Identify the polysemous word in the result set and frame the target word, t_w.*
5. *Pair the target word with other words in the set T and scan the database for relationship.*
6. *For each word pair,*
   If the word pair framed in the corpus is same as the observed word pair in set T, then it is selected as the correct sense.
7. *End for.*

IV. PERFORMANCE EVALUATION

The accuracy of the algorithm is brought out through a rigorous comparison of the designed scheme with that obtained using both Lin's algorithm and a heuristic approach is as shown in Fig. 2. The test kit has coverage of 5 words. The precision and recall are computed as shown below. The results reveal a higher precision recall rates for the proposed strategy. In addition to the scheme being adequately valuated through an effective comparison with the existing formulations, it

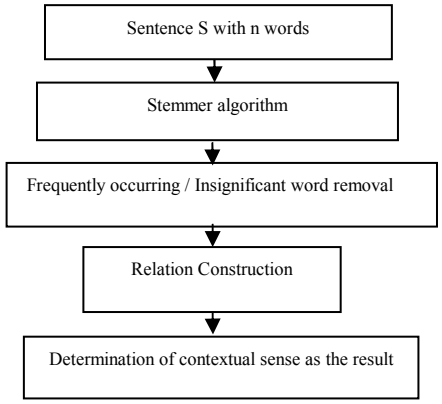

Fig. 1. Schematic representation of the proposed system

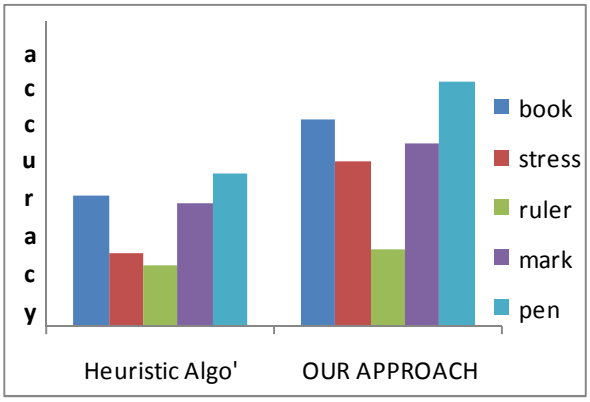

Fig. 2. Accuracy Percentage: WSI Algorithm Vs Heuristic Algorithm

TABLE 2
PRECISION AND RECALL

	Cover	Prec. %	Recall %
WSI Proposed work	W=10	72.1	64
WSD Heuristic algo		68.3	64

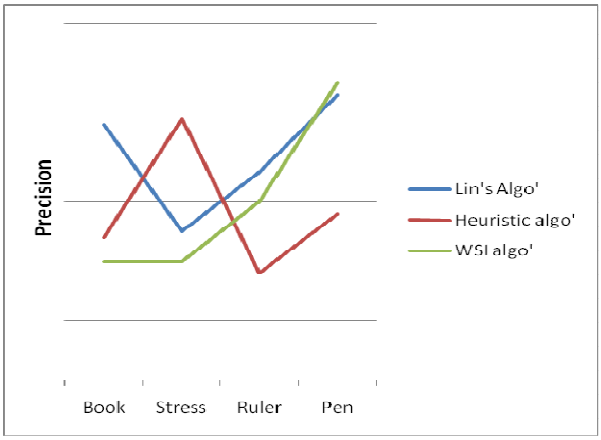

Fig. 3. Graphical representation of precision%

highlights its suitability for being applied in real time situations. The results are checked for Lin's algorithm which states that "two different words are likely to have similar meanings if they occur in identical local contexts", this means that the same knowledge sources are used for all words and that instead of building separate classifiers for each word, past usages of other words are used to disambiguate the current word. With the test kit words and found to give an irregular precision path in the graph, when compared to the WSI algorithm that improves gradually in its performance.

V. CONCLUSION

A novel strategy has been formulated for identifying the correct sense of a word in a sentence. The proposed disambiguation methodology has been constructed through the use of an appropriate database and the relevant semantic relations created using WordNet. The algorithm has been designed to cater to a wide variety of words in the chosen domain. The performance has been found to excel over the traditional methodologies and therefore owes its promising nature in this challenging arena. The exercise incarnates a far reaching implication and will go a long way in reaching out to the disabled citizens of the society.

REFERENCES

[1] 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.

[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, 1994. "Decision lists 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), Las Cruces, U.S.A., 88-95, 1994.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Rigau G. 1994. An experiment on Automatic Semantic Tagging of Dictionary Senses, WorkShop "The Future of Dictionary", Aix-les-Bains, France. published as Research Report LSI-95-31-R. Computer Science Department. UPC. Barcelona.

[11] Agirre E., Rigau G. 1996. Linking Bilingual Dictionaries to WordNet, in proceedings of the 7th Euralex International Congress on Lexcography (Euralex'96), Gothenburg, Sweden, 1996.

[12] 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.

[13] Using WordNet to Disambiguate Word Senses for Text Retrival, in proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 171-180, PA.

[14] Fass D., Guo C., McDonal J., Plate T. and Slator B. Wilks Y., 1993. Providing Machine Tractablle Dictionary Tools, in Semantics and the Lexicon (Pustejovsky J. ed.), 341-401. Yarowsky, D. 1992.

[15] 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.

[16] Ando R. and Zhang T. A framework for learning predictive structure From multiple tasks and unlabeled data. Journal of Machine Learning Research,volume 6, pp. 1817|1853, 2005.

[17] Ando R.K. Applying alternating structure optimization to word sense disambiguation.In Proceedings of the 10th Conference on Computational NaturaLanguage Learning (CoNLL), pp. 77{84. New York City, 2006.

[18] Agirre E. and Soroa A. Using the multilingual central repository for graphbasedword sense disambiguation. In Proceedings of LREC '08 . Marrakesh,Morocco, 2008.

[19] Agirre E. and Soroa A. Personalizing pagem rank for word sense disambiguation.In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09). Athens, Greece, 2009.