International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Approaches To Word Sense Disambiguation

M. Trivedi (Author)
Dept. of Computer Engineering
D. J. Sanghvi College of Engineering
Mumbai, India

S. Sharma (Author)
Dept. of Computer Engineering
D. J. Sanghvi College of Engineering
Mumbai, India

K. Deulkar(Author)
Assistant Professor
Dept. of Computer Engineering
D. J. Sanghvi College of Engineering
Mumbai, India

***Abstract*—Removing ambiguous meaning of a word has been an extensive area of research in the field of computational linguistics. This paper presents a comprehensive study of the different dictionary approaches to Word Sense Disambiguation. Three approaches have been explored; graph based, ontology based and knowledge based. All the approaches use a corpus to evaluate words unlike in a machine learning approach. A machine learning approach trains a dataset of words and calculates the probability that a sense is correct. Knowledge based approaches are better but require world knowledge to be most efficient. Hence it is suggested to use hybrid methods which combine machine learning algorithms with corpus analysing algorithms.**

Keywords—Word Sense disambiguation; graph based approach; ontology; Lesk Algorithm; Conceptual Density; Random Walks.

I. INTRODUCTION

In computational linguistics, the sense of a word is its meaning in a part of speech. A word may have different senses and it is difficult for the machine to determine what sense to refer to in a sentence. Word sense disambiguation is a task of removing ambiguities and selecting the closest sense of the word in context.[1]Two main wide spectrum approaches are used for WSD. One approach involves referring to Knowledge based dictionaries to compare the word with its senses in the MRD's, Ontologies or Thesauri and determine the correct meaning[2]. The second approach is the machine learning approach which according to [3] is learning the classifier so that it can be applied to unseen senses.[4]Usually unstructured data is used for supervised and unsupervised approaches. This paper mainly focuses on providing a brief overview of the dictionary based approaches to WSD. Machine Readable Dictionaries are used to retrieve the senses of a word and are the primary requirements to implement dictionary based approaches.

The flow of this paper covers the following: Section 2 gives a brief overview of Word Sense Disambiguation and its approaches, section 3 summarizes the paper in conclusion and future work is proposed in section 4.

II. WORD SENSE DISAMBIGUATION

WSD is one of the central challenges in NLP and is mainly applied in Information Retrieval, and Machine Translation. A classification problem in WSD involves senses corresponding to classes and context is regarded as evidence. There are three broad approaches which are discussed in the following subsections.

A. Graph based approach

A traditional dictionary based approach to WSD entails a word's senses to be compared to those of the surrounding text. This method has drawbacks because the senses will be compared in a pair wise manner which exponentially increases computational complexity with increase in number of words. Agirre and Soroa [4] have proposed a graph based algorithm in which the graph is analyzed as a whole and given the relation between entities, globally optimal solution can be obtained.The graph is constructed with senses as nodes as given by [4] and edges represent the relation between nodes(senses) which may have some weight attached. Once the graph has been constructed, the PageRank algorithm is applied to its vertices according to its structural importance. The PageRank algorithm proposed by Brin and Page [5]andis given by

$$R'(u) = c\sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \quad (1)$$

Where E is a source of rank, c is the normalization factor, and R' is the assignment of PageRank.

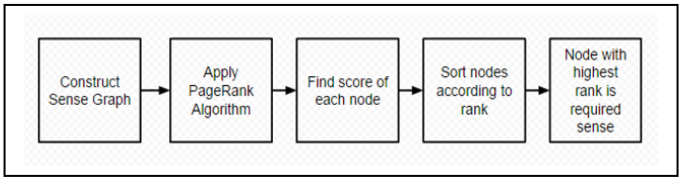

Fig. 1.Steps involved in graph based approach.

B. *Ontology based approach*

Ontology is a knowledge based model of concepts existing in the world and the relations between them. A taxonomic structure is used for organizing ontologies with sub classes inheriting properties of the base class [6]. In addition, new facts can be inferred from old facts using latest semantic web technologies.

For an ontology based approach a domain specific ontology O is used. Let O = {C, R}, where C denotes concepts and R denotes relations. The set C of concepts is divided into two sets C_U and C_A. C_U consists of n-grams that occur only once in the ontology and C_A contains all the n-grams that occur more than once. Therefore, $C_U \cup C_A = C$, and $C_A \cap C_U$ = null. For word sense disambiguation we extract a list of concepts from each document based on n-gram matching.

C. *Knowledge based approaches*

C.1. Lesk Algorithm

The lesk algorithm was first proposed by Michael Lesk in 1986[7] which stated that the algorithm can choose the correct sense by checking for the overlap in features between the dictionary definition of the ambiguous word and its neighboring words in the context.A count is assigned to a sense of the word every time an overlap of features is found. The sense with the highest count can be interpreted as the required meaning. A modified version of Lesk Algorithm put forth by [8] explores relationships between meanings by extending those senses which are semantically related. Further, [9] applies Lesk algorithm using a Distributional Semantic Model and replaces overlap with similarity. The DSM is visualized to have points in space and the distance between these points indicate the semantic similarity. The points represent information about co-occurring context words which is essential to calculate the similarity.

C.2. Conceptual Density

Conceptual density is an elaboration of Conceptual Distance wherein [10] explains that the system has to know how words are clustered into semantic classes and how these classes are arranged in a hierarchy. In order to accomplish this, the Conceptual density of nouns needs to be maximized. To maximize conceptual density [10] proposes a search for combinations of senses from an array of nouns. Each sense of a word is contained in a sub-hierarchy. The sub-hierarchy having maximum density is chosen as the meaning of the word.
The formula for the conceptual density is given by:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c} \qquad (2)$$

Where c is a given concept, m is the marks of senses of the words to be disambiguated; nhyp is mean number of hyponyms per node. The 0.20 was observed to smooth the exponential i and the algorithm gave the best performance with 0.20.
The algorithm works on the nouns in the following way, as described by [10]:
Step 1: Represent the nouns along with their senses and hypernyms on a lattice. Non- noun words are not considered.
Step 2: Using the formula, Conceptual density *c* is calculated for each concept in WordNet with respect to the senses contained in the sub-hierarchy.
Step 3: The sense with the highest *c* is chosen. The words below it are chosen as correct senses for respective words.
Step 4: The above steps are repeated till no further disambiguation can be done.

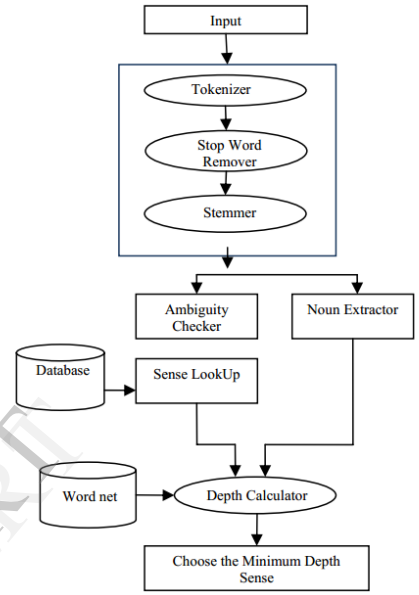

Fig. 2. Flowchart giving an overview of Conceptual density [10]

C.3. Random Walks

Random Walk algorithm is a graph-based sequence labeling algorithm proposed in [11] for linguistic annotation tasks. The basic idea of the algorithm is to annotate each word in a sequence by exploiting relations identified among them.
Consider the following example:

The **church bells** no longer **rung** on **Sundays**.

church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edification

Sunday
1: first day of the week; observed as a day of rest and worship by most Christians

Fig. 3.Senses of words obtained from WordNet from [11].

The algorithm to be followed is:
Step 1: Add a vertex corresponding to the senses of the words. For example, there will be three vertices for the word 'ring'.
Step 2: Connect the vertices by weighted edges using definition based semantic similarity measure(Lesk's method).
Step 3: Use a ranking algorithm to find score of each vertex. This is indicated as values between brackets next to each node.
Step 4: Select the vertex (i.e., sense) that has the highest score.

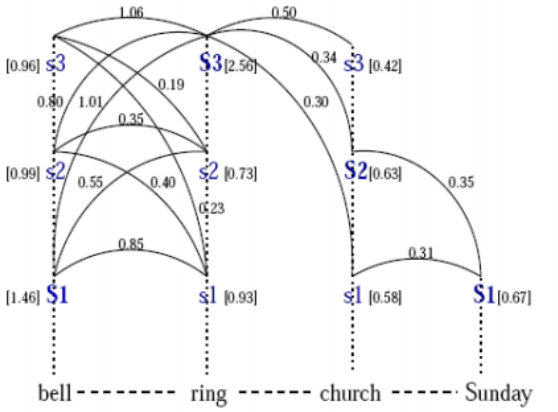

Fig. 4.The label graph constructed with scores for each sense from [11].

Thus, in the above example, the church bells no longer rung on Sundays can be disambiguated as The church#2 bells#1 no longer rung#3 on Sundays#1.

III. CONCLUSION

In a nutshell, there are umpteen approaches to remove ambiguity of word senses and this paper elucidates some of them. Graph based methods involve constructing graphs and senses with highest weights are selected. In Ontology based methods, one can infer new senses from old senses using sophisticated semantic web tools. Knowledge based approaches are widely researched because there is continuous scope of improvement. The possibility of using other approaches is discussed in the future work.

	Approach	***Type***
1.	Graph based	Spreading Activation
2.	Ontology based	-
3.	Lesk Algorithm	Dictionary
4.	Conceptual Density	Correlation
5.	Random Walks	Knowledge /Graph

Fig. 5.WSD Approaches and their type.

IV. FUTURE WORK

The approaches discussed in this paper are efficient up to a certain size of the corpus. As the users in the internet aregrowing, so is the communication which leads to creation of new words which also need to be added to the corpus for accurate results, instead of manually inserting handwritten synsets in the corpus, efforts are being made to modify the way data is stored in dictionaries such that they learn and create lexical graphs from user input. This can be an efficient way to analyze proper nouns. The process of learning should be automated and should be as accurate as possible.Hybrid approaches which combine dictionary and machine learning methods is a more recent topic of research and can significantly approve results.

REFERENCES

[1] R.Navigli,2009.[Online].Available:http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/ACM_Survey_2009_Navigli.pdf.
[2] R. Prokofyev, G. Demartini, A. Boyarsky, O. Ruchayskiy and P. Cudre-Mauroux,2014.[Online].Available:http://exascale.info/sites/default/files/OntologyBased%20Word%20Sense%20Disambiguation%20for%20Scientific%20Literature.pdf.
[3] E.Palta,2006.[Online].Available:http://www.it.iitb.ac.in/~esha/resources/firststage.pdf.
[4] E.Agirre and A. Soroa, 2009. [Online]. Available: http://www.aclweb.org/anthology/E/E09/E09-1005.pdf.
[5] S. Brin and L. Page, 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.
[6] F. Leao, K. Revoredo and F. Baiao, 'Learning Well-Founded Ontologies through Word Sense Disambiguation', ResearchGate, 2013. [Online].Available:http://www.researchgate.net/publication/258094329_Learning_Well-Founded_Ontologies_through_Word_Sense_Disambiguation.
[7] M.Lesk,1986.[Online].Available:http://promethee.philo.ulg.ac.be/engdep1/download/prolog/lexdis/docs/lexdis/otherpap/Lesk%20clean.pdf.
[8] S.Banerjee,2002.[Online].Available:http://www.d.umn.edu/~tpederse/Pubs/banerjee.pdf.
[9] P. Basile, A. Caputo and G. Semeraro, 2014. [Online]. Available: http://anthology.aclweb.org/C/C14/C14-1151.pdf.
[10] E. Agirre and G. Rigau, 'Word sense disambiguation using conceptual density', pp. 16--22, 1996.
[11] R. Mihalcea, 'Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling', pp. 411--418, 2005.