A Critical Approaches to Identification of Disambiguation Words in NLP: Current State of the Art

J. Sreedhar¹, Dr. S. Viswanadha Raju², Dr. A. Vinaya Babu³, Amzan shaik⁴, Dr. P. Pavan Kumar⁵

¹Associate Professor and HOD-CSE, ECET, Hyderabad, India.
²Professor and HOD-CSE, JNTUJ, Hyderabad, India.
³Professor in CSE and Principal, JNTUH, Hyderabad, India.
⁴Professor-CSE, ECET, Hyderabad, India.
⁵Professor in Telugu Department, HCU, Hyderabad, India.

Abstract—This paper presents the current state of art in word sense disambiguation in NLP. They offer several specific proposals to the community regarding improved evaluation criteria, common training and testing resources and definition of sense inventories. WSD is a open problem at lexical level of natural language processing. One requires parts of speech tagging and morphological analysis, software and lexical data are not hard to find.

Keywords—WSD, Telugu, Context-Base, Case-Base, Noun.

I. Introduction

In any language there are words which are ambiguous with more than one sense. For example the English word ‘bank’ has at least two senses, viz: Bank1 – as a financial organization and Bank2 – as the border of a water body. The task of word sense disambiguation (WSD) is to determine which of the senses of an ambiguous word is invoked in a particular use of the word. A standard approach to WSD is to consider the context of the word’s use. WSD is very important for applications like machine translation, information retrieval etc. For example, in a machine translation from English to Telugu, words like ‘bank’ has to be disambiguated so that they are translated correctly in the target language (i.e., in this case correct Telugu word is chosen). Also, in information retrieval system answering a query about ‘financial banks’ should return only documents that use bank in the first sense. In short, whenever a system’s actions depend on the meaning of the text being processed, disambiguation is beneficial or even necessary. Similarly a Question-Answering system should interpret the right meaning of an ambiguous word (in the query) to be able to answer the question correctly.

This paper is organized as follows: first, we formalize the Previous Mechanisms of WSD (Section 2), and Current Approaches of WSD (Sections 3). Next, we turn to the Conclusion (Section 4) followed by the References.

II. Previous Mechanism

Earlier attempts on WSD focused on Supervised learning, by assuming that resources like WordNet (Xiaobin Li et al. 1995) or sense-coded dictionary (Lesk 1986; Dagan et al.1991) or Thesauri (Yarowsky 1992) or hand-labeled training sets (Hearst 1991; Niwa and Nitta 1994; Bruce and Wiebe 1994) are available a priori. Development of these resources requires huge amount of human efforts and typically takes years for building. As these resources are not available for Telugu, such supervised techniques can not be immediately applied. Few attempts has been made on unsupervised WSD like (Yarowsky 1995), which seeks minimum human involvement, in the form of providing a few seed words that occur with each sense of the ambiguous word for bootstrapping the algorithm. The algorithm then classifies each occurrence of the ambiguous word in a corpus (training phase) into several clusters such that all the occurrences are in the same sense within a cluster. Additional co-occurrences are collected in this process, which are then used for disambiguating unseen texts from the held-out corpus (testing phase).

A. Surveillances

Observation 1:

Evaluation of word sense disambiguation systems is not yet standardized. Evaluation of NLP task including part of speech tagging and parsing are fairly standardized with most using training and testing corpus. There are as many test suites as there are researchers in this field. Some of them are Leacock et al. (1993) shared by Lehman (1994) and Mooney (1996) and evaluated on the system of Gale, Church and Yarowsky (1992).

Observation 2:

The potential for WSD varies by task. Disambiguating word senses is an intermediate capability that is believed to improve natural language processing. Sense information is potentially most relevant in the form of word equivalence classes for smoothing in language models. At the level of monolingual lexical information useful for high quality machine translation.
**Observation 3:**
Adequately large sense-tagged data sets are difficult to obtain. Availability of the data is the factor of contributing to part of speech tagging, parsing etc. The only broad coverage annotation of all the words in a sub corpus is the Word net semantic concordance providing large data for the study of distributional properties of polysemy in English. In addition, slowing annotation speed and lowering intra-and inter annotator agreement rates. The Word net is a central training resource for the disambiguation algorithms. Other source of sense tagged is from parallel aligned bilingual corpora. The disambiguation algorithms need tagged data for the evaluation statistics by annotators against annotators. 

**Proposals:**
1. Adequately A better evaluation criterion: standard for evaluation of word sense disambiguation algorithm is the exact match criterion.

\[
\%\text{correct} = 100 \times \frac{\#\text{exactly matched sense tags}}{\#\text{assigned sense tags}} \tag{1}
\]

This criterion suffers some obvious drawbacks. Cross–Entropy is used to evaluate the model:

\[
\frac{1}{N} \sum_{i=1}^{N} \log_2 Pr_A(CS_i/W_i, \text{context}_i) \tag{2}
\]

Not all classification algorithms return probability values. This evaluation measure also does not give the exact match. A measure cross-entropy provide a fairer test. For the systems which yield poor estimated values a variant of cross entropy measure without the log term can be used to measure.

\[
\frac{1}{N} \sum_{i=1}^{N} Pr_A(cs_i/w_i, \text{context}_i) \tag{3}
\]

Latter the assigned tag is given probability 1 and all other senses probability 0.

2. Make evaluation sensitive to semantic or communicative distance between sub senses.

Evaluation metrics failed to take into account the senses while assigning penalties for incorrect labels. Classification between close siblings in sense hierarchy given little penalty while misclassification across homographs receive greater penalty.

Penalty distance \((\text{subsense}_1, \text{subsense}_2)\) captures the hierarchical distance derived from word net. Ex: Distance matrix for bank

These penalties are based on pair wise functional communicative distance. The simplest penalty weighting is to minimize the mean distance or cost between the assigned sense and correct sense to \(N\).

\[
\frac{1}{N} \sum_{i=1}^{N} \text{distance}(cs_i, as_i) \tag{4}
\]

The probabilities assigned to incorrect senses weighted by the communicative distance/cost. For the incorrect sense and the correct one.

\[
\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} \text{distance}(cs_i, s_j) \times Pr_A(s_j/w_i, \text{context}_i) \tag{5}
\]

They consider the \(S_i\) senses of word \(W_i\), the probability assigned by the classifier \(A\).

3. A frame work for common evaluation and test set generation.

Supervised and unsupervised need different systems and evaluations, unsupervised are evaluated by tagged corpus Ex: Word net and unsupervised method require larger data focused on polysemous. The proposed frame work is as follows;

- Collect a large un annotated corpus
- Select a sense inventory to which algorithm will be evaluated
- Select a subset of \(R<N\)
- Select a smaller subset of \(S<R<N\) and generate the test as:
  1. Select a set of \(M\) without knowing what those words would be.
  2. For each of \(M\) words, they’ll compute evaluation statistics by annotators against annotators.
  3. For each \(M\) words, where annotators disagree and make consensus choice.

In this state, there is are no changes made.
- Each algorithm do WSD on the test corpus.
- Evaluate the performance of the algorithm with \(M\) words annotated for the evaluation.
- Release the latest corpus that require supervised training.

Go back to step 3 for next evaluation.

4. A Multilingual sense inventory for evaluation.

A general criterion that can be applied to existing sources of word senses, make sense for both target application and for evaluation. The proposal is to restrict a word sense inventory to those distinctions that are lexicalized cross linguistically.
III. Current Approach

Here, the idea is to reduce the human effort needed for sense-tagging, when compared to (Yarowsky 1995). This approach is similar to and an extension of Context-group discrimination (Schutze 1998). In our present approach all the occurrences of the ambiguous words are classified into different clusters in such a way that all the occurrences are in the same sense within a cluster. Then co-occurrence words are collected for each cluster. These words are used for manually assigning the sense for each cluster. Also, it is planned to probe the applicability of the inflections of words in WSD for a rich inflectional languages like telugu. The hypothesis is that “Each sense of an ambiguous word will predominantly co-occur with words in a particular inflected form”. The preliminary investigations reveal that the hypothesis is indeed useful for some senses of an ambiguous word if not for all senses. So, it is proposed to use this information simultaneously with the co-occurrence information explained earlier.

A. Context Based Approach

The context, on which the sentence is appearing, provides valuable clues for Sense disambiguation. The ‘context’ means the nearby words that are present in the sentence containing the ambiguous word. These nearby words provide valuable clues in identifying the right sense for the word. However the notion nearby may not be really ‘nearby’ as these high informative words may also appear away for the ambiguous word. But it has been true in most of the cases that informative words occur near the ambiguous words and can be used reliably. Here the aim is to identify the context words for each sense of the ambiguous word that will uniquely represent one particular sense for that word in the given context.

B. Case Relation Based Approach

A new approach is being tried here for WSD which uses the case markers of both context words and the ambiguous word. The hypothesis is, “Each sense of an ambiguous word with multiple senses is related to the one by words (in a specific window) in a particular fashion”. The hypothesis is supported by the fact that each sense of the ambiguous word occurs as an argument of a particular group of verb, and each group in turn takes arguments with different relations (expressed by the case markers). Thus, the case markers of the context words and that of the ambiguous word itself will act as an indicator in identifying the correct sense of the ambiguous word.

C. Integrating Context-Based and Case-Based Approaches

In the first phase (training phase) the analyzer uses the CIIL corpus to collect the collocations and the prominent cases of each sense of an ambiguous word. The results obtained from this phase are represented in a disambiguation dictionary. This dictionary typically contains the following information for each sense of the ambiguous word.

i) Collocational words

ii) Prominent case markers of the collocational words

iii) Prominent case markers of the ambiguous words in the next phase (testing phase) the analyzer uses these information to disambiguate a raw text.

Here the analyzer looks for matching context word or the case information inorder to choose the right sense of the target word. A weighting function is used to smooth out the differences while picking information these three resources.

IV. Conclusions

The most important of our observations about the current state of art in WSD is that it is still a hard, open problem, for which the field has not yet narrowed much. We proposed that the accepted standard for WSD evaluation include a cross-entropy like measure that tests the accuracy of the probabilities assigned to sense tags and offers a mechanism for assigning partial credits. We suggested a paradigm for common evaluation that combines the benefits of the “traditional interesting word”, evaluation with an emphasis on broad coverage and scalability. We outlined a criterion that should help on determining a suitable sense inventory to use for comparison of algorithm, compatible with both hierarchical sense partitions and multilingual motivated sense distinctions.

REFERENCES


ABOUT THE AUTHORS

J. Sreedhar is a Research Scholar, Department of Computer Science and Engineering, JNTUH, Hyderabad, India. He has received M.Tech.(Computer Science and Technology) from Andhra University. He has been published and presented good number of Research and Technical papers in International Journals, International Conferences and National Conferences. Presently Working as an Sr.Assoc Professor and HOD in the Department of CSE, ECET, Hyderabad. He published and presented good number of research papers in various conferences. His main research interests are NLP, IRS , Data Mining, and Compiler Design, FLAT, Computational Linguistics.

Dr. S. Viswanadha Raju obtained his Ph.D in Computer Science &Engineering from ANU. He obtained his M.Tech in CSE from JNTUniversity. He has a good academic background with a very sound academic and research experience. At present he is working as a professor and HOD-CSE in JNTUniversity, Jagityala,Hyderabad. He is guiding 10 research scholars for Ph.D and also conducted several conferences/workshops/seminars with sponsored agencies such as AICTE, DST, TCS, IEEE and CST. His research interests are Information Retrieval, Databases, Image Retrieval, Data Mining and related areas. He published 25 research papers in reputed International Journals/Conferences proceedings in his research area. He is active member in different professional bodies with life membership like IETE, ISTE and CSI.

Dr. A. Vinay Babu obtained his Ph.D in Computer Science & Engineering from JNTU. He obtained his M.Tech in CSE from JNTUniversity. He has a good academic background with a very sound and academic research experience. At present he is working as a professor in School of Information Technology in JNTUniversity, Hyderabad. He is guiding 10 research scholars for Ph.D and also conducted several conferences/workshops/seminars with sponsored agencies such as AICTE, DST, TCS, IEE and CST. His research includes Information Retrieval, Databases, Image Retrieval, Data Mining and related areas. He published 82 research papers in reputed International Journals/Conferences proceedings in his research area. He is active member in different professional bodies with life membership like IETE, ISTE and CSI.

Amjan.Shaik is a Research Scholar, Department of Computer Science and Engineering, JNTUH, Hyderabad, India. He has received M.Tech.(Computer Science and Technology) from Andhra University. He has been published and presented good number of Research and Technical papers in International Journals, International Conferences and National Conferences. His main research interests are Software Metrics, Software Engineering, Software Testing, Software Quality, Object Oriented Design and NLP.

Dr. P. Pavan Kumar, is working as Assistant Professor in the Department of Telugu at University of Hyderabad, India. He teaches Computational Linguistics in the department and guiding research in its functional aspects. He has published 19 books so far, as co-author, Editor and presented several papers in National and International conferences. He is a life member in the professional bodies like LSI, DLA.