IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-8 E-ISSN: 2347-2693

An Approach for Improving Accuracy of Machine Translation using WSD and GIZA

S.G. Rawat[1*], M.B. Chandak[2], N.A. Chavan[3]

[1*]Dept. of IT, G.H.Raisoni College of Engineering, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India
[2] Dept. of CSE, Shri Ramdeobaba College of Engg. and Management, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India
[3] Dept. of IT, G.H.Raisoni College of Engineering, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India

[*]*Corresponding Author: ssunitarawatt@gmail.com, Tel.: +91-7767960666*

Available online at: www.ijcseonline.org

Received: 28/Sep/2017, Revised: 07/Oct/2017, Accepted: 23/Oct/2017, Published: 30/Oct/2017

Abstract— Word Sense Disambiguation (WSD) is a challenging problem of Natural Language Processing (NLP). Though there are lots of algorithms for WSD available, still little work is carried out for choosing optimal algorithm for that. The job of word sense disambiguation is to decide the accurate meaning of an ambiguous term in a particular circumstance. When WSD is used in machine translation, an accurate translation in the resultant linguistic must be determined for an ambiguous term entry in the original language. Therefore Word Sense Disambiguation remains one of the most common real life problems that are associated to natural language processing which needs to be resolved efficiently.Unsupervised techniques use online dictionary for learning, and supervised techniques use manual learning sets. As there are some advantages and disadvantages of supervised learning and unsupervised learning, aim of this paper is to disambiguate the ambiguous word by using the hybrid approach for WSD. We have made use of parallel corpus and aligned the text by using GIZA.

Keywords—WSD, Machine Translation, Corpus, Supervised, Unsupervised.

I. INTRODUCTION

Around the world people use so many languages to talk whereas in most of the languages there are several words which represent meanings in different contexts. Therefore to determine the accurate meaning of an ambiguous term in a given circumstance we use WSD technique. For instance, a term 'light' (in English) can have various meanings like "physical weight", "relative darkness or lightness in terms of color", etc. Such types of terms with several meanings are denoted as ambiguous words. The method of determining the accurate meaning of an ambiguous word for a certain circumstance is known as Word Sense Disambiguation. As a person we have ability to distinguish the several meanings of an ambiguous word in a certain circumstance; however systems cannot take their decisions own until we fed them with proper instructions [1, 2].

Human being decides the meaning of a word based on their experiences related to that context. Whereas, machines do not have such ability to decide an ambiguous situation unless, some rules have been set into it. In supervised learning, system is trained by using a training set. In that training set lots of sentences are covered so that after learning system can predict the particular sense of the ambiguous words in the given circumstance. Precise learning set is generated as a result for every occurrence of different sense. Depend on the defined learning set, a system finds the possible sense of a vague term for the given circumstance. Training set is formed manually unable to make permanent rules for explicit system. Hence expected meaning of an ambiguous term in a precise circumstance may not be constantly recognized. Supervised learning is able to obtain fractional expected result, if the training set does not include adequate data for all feasible senses of the ambiguous word. It illustrates the consequence, only if there is data in the predefined database.

For avoiding the inadequacy of supervised learning, online dictionary is taken as training set in unsupervised learning, "WordNet" is the mainly extensively used online dictionary enclosing "words and their related meanings" along with "relations with dissimilar words" [3]. The WSD method is significant for various purposes like extraction of information, knowledge acquisition, and text classification and so on. WSD plays a significant part in machine translation.

The rest of the paper is arranged as follows: Section 2 comprises the related work in WSD. Section 3 discussed various approaches of word sense disambiguation. Section 4 describes about hybrid technique for WSD. Section 5 comprises of preparation of corpus, training of corpus and parallel corpus. Section 6 includes role of GIZA.

II. RELATED WORK

After knowing the problems associated with NLP, Researchers planned different techniques to override the difficulties, few techniques are depend on corpus evidence while others on dictionaries. Knowledge based methods lie on Wordnet. Knowledge based methods makes use of dictionaries to obtain exact meaning. Author used WSD with Conceptual Density in [4]. The main aim of this method is to find correct meaning depending on the conceptual distance method specifically in what manner ambiguous term and circumstance term are associated to each other. Author has extended the work in [5] with the little change in method to get the exact sense, and they name the method as Selectional preference which tries to find possible relations among word types.

In [6], author suggested a method Lesk and Extended lesk comes under Overlap based techniques that are entirely depend on similarity of term with context terms. Main problem of mentioned technique is, it greatly rely upon dictionaries that too comprise of a few limitations on obtaining the data. Machine learning techniques merely make use of corpus. Techniques come under supervised learning are SVM, baye's and so many. Naive baye's technique based on the conditional probability calculation, comprising of part of speech, cooccurrence, feature as collocation [7]. Decision list method is easily understandable if else then method. Exemplar-based technique is generally based on the examples.

After that the newest method is SVM. This method uses binary categories, depending on appropriate and inappropriate meanings, division of categories done. Creating manually tagged corpus is the most important problem with supervised learning. To conquer the difficulty of supervised learning, unsupervised methods came in existence. Author made use of feature selection technique by using corpus which comes under unsupervised approach in [10]. Drawback of unsupervised systems is that because of cluster issues the results are not enhanced than that of supervised systems.

III. WSD APPROACHES

Word Sense Disambiguation methods are categorized into three main types- a) Knowledge based approach b) Supervised approach and c) Unsupervised approach.

A. Knowledge-based Approach

Knowledge-based methods rely on information that can be extracted or inferred from a knowledge source, such as a dictionary, thesaurus or lexical database. These methods learn based on information from curetted and structured data whereas supervised and clustering methods learn from example instances. The advantage of the knowledge-based methods over the supervised and the clustering methods is that training data is not required for each word that needs to be disambiguated. This allows the system to disambiguate words in running text, referred to as all-words disambiguation [11].

B. Supervised Approach

A supervised technique makes use of sense-tagged corpora to prepare the sense model, which helps to form link between contextual features and word sense. Hypothetically, it should do better than unsupervised methods as additional information is fed into the system. Since so many training corpora are available at the present time, the majority newly developed WSD algorithms are supervised. But, it does not mean unsupervised method is beyond mode [12]. Depending on whether features are directly related or not with the word sense in training corpora, supervised methods divide approximately into two classes, hidden models and explicit models.

C. Unsupervised Approach

Unsupervised learning learns how machines can be trained to signify specific input patterns in a means that imitate the statistical structure of the inclusive collection of input examples. In dissimilarity with reinforcement learning or supervised learning, there are no overt final outputs or ecological assessments related with every one input. Since unsupervised learning is probably to be greatly ordinary in the mind compare to supervised learning it is important.

IV. HYBRID APPROACH FOR WORD SENSE DISAMBIGUATION

In Hybrid approach we are going to merge supervised and unsupervised method to obtain the more correct output. In hybrid approach, stop words similar to 'a', 'an', 'the', and so on are being removed from input texts because these kinds of words are meaningless to obtain the "sense" of the given sentence. Afterward, the text excluding the stop words is gone through supervised and unsupervised algorithms in a corresponding manner [2]. "Module 1" is consists of supervised algorithm and, "Module 2" consists of unsupervised algorithm. These two modules are accountable to get the real sense of ambiguous words in the given context. The words which are unmatched in both the modules are being kept in a temporary database for additional usage. Afterward, outputs of "Module 1" and "Module 2" have been being examined to formulate the exact sense based on the context of the sentence in "Module 3". If either of the Module 1 or Module 2 by applying "OR" operation finds the sense then that exact sense is allocated to the unmatched words in the temporary database. Accuracy of results depending on the implemented methods is verified in "Module 4". If "Module 1" and "Module 2" obtains similar result obtained by applying "AND" operation, then the sense is assumed as disambiguated sense. Thus, matchless words (stored in a temporary database) have to be shifted to

associated sense bag according to the "Bag-of-Words" algorithm in "Module 1" to contribute in disambiguation technique here after. If not, obtained senses are believed as the possible senses and matchless words are being shifted to a probable database in "Module 5".

V. CORPUS

A. Preparation of Corpus

Data capture is the initial stage of corpus creation, which includes depiction the manuscript in online kind, by using OCR or hand, bring out software result, and so on. As Manual entry is costly and takes long time, not appropriate to construct enormous corpora. Similarly, for validating data if we use post-processing then OCR result may be costly.

B. Training of Corpus

Training corpus is a collection of texts, containing manually validated linguistic information, attributed to the original texts. Machine-learning programs use this information to make a statistical model and it can also be used in rule-based programs to find the accuracy. This kind of models can be used by Statistical programs for examining new, unidentified texts. For creating training corpus, given text which comprises of a series of characters has to be separated into words, sentences, punctuation and paragraphs. This process is known as segmentation and tokenization.

Along with every word two other data are credited: first is base word or a rule known as lemma (playing, playing -> play) whereas second is a morphosyntactic label. Further training corpus includes name entities (e.g Ram, NASA), syntactic data the relationship among pronouns with corresponding referents and so on.

C. Parallel Corpus

A parallel corpus is a corpus that contains a collection of original texts in one language and its translated texts into other languages. Generally, parallel corpora comprise texts from two languages only.

1) Types of parallel corpora

Parallel corpora can be consists of bilingual or multilingual, specifically they include texts of two or more than two languages. Parallel corpora can be either unidirectional (e.g. an English text translated into Hindi), bidirectional (e.g. an English text translated into Hindi and vice versa), or multidirectional (e.g. an English text such as an EU regulation translated into Hindi, Marathi, Gujarati, Tamil, Telgu, etc.).

2) Alignment of a parallel corpus

To make use of a parallel corpus appropriately it is essential to align the source text and translation(s) of source text. This means that we have to recognize the pairs of words, phrases and sentences in the original text and their associated text in the other languages. Parallel text alignment is important because during the translation process sentences may be fragmented, combined, removed, introduced or rearranged by the translator for converting it into destination language. In the process of alignment, anchor points such as proper names, numbers, quotation marks etc. are often used as points of orientation. The degree of correspondence between the texts of a parallel corpus varies depending on the text type.

VI. GIZA

A. Role of GIZA

Role of GIZA program is to align lexis and series of lexis in sentence allied corpora. GIZA can use to form bilingual dictionaries if parallel corpus is available or it can also use for lexical selection rules.

B. Introduction to GIZA

GIZA is a program that trains the IBM Models in addition to a HMM, and uses these models to compute Viterbi alignments for statistical machine translation [16]. Though GIZA++ can be used on its own, it typically the initial state for other machine translation systems, like syntactic and phrase-based. For example, running GIZA++ is initial phase in training the popular phrase-based translation scheme Moses [17]. The hierarchical phrase-based translation system Hiero [15] also uses GIZA++ to generate word alignments. [14] Used word alignments from GIZA++ to learn rules for syntax-based machine translation.

C. Role of System Training in Finding Sense Disability

In this project we have trained the system by using GIZA++. We have used two corpus related to tourism in English and Hindi language. GIZA++ trains the system by aligning the words of same sense from both the corpus. To each word in training-set some number has to assign. System gets trained by understanding which numbers are forming pair while alignment. After forming the pairs of the words in English and Hindi language, subsequent phase is to discover the meaning of ambiguous words in given context.

VII. CONCLUSION

Following conclusions are drawn depending on our learning of WSD approaches:

1. Keeping in mind drawbacks of present techniques such as knowledge based methods needs knowledge resources and complete list search, technique called as supervised method has drawback of insufficiency of data, as well as requirement of training enormous parameters and the last technique i.e. unsupervised method flops to differentiate among better meaning of an ambiguous term therefore this is an attempt to sort out the problem by signifying the hybrid method.

2. We made use of parallel corpus i.e. English and Hindi corpus. To get the accuracy in machine translation we have

used GIZA to align lexis as well as series of lexis in sentence allied corpora. GIZA trains the system and gives more accurate results.

REFERENCES

[1] A.R. Pal and D. Saha, "*Word Sense Disambiguation: A Survey*", International Journal of Control Theory and Computer Modeling (IJCTCM), Vol.5, No.3, pp. 1-16, 2015.

[2] A. Kundu, A. Singh, R. Shekhar, "*A Hybrid Approach to Word Sense Disambiguation Combining Supervised and Unsupervised Learning*", International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 4, No. 4, pp. 89-101, 2013.

[3] A.R. Pal, A. Munshi, D. Saha, "*An Approach To Speed-Up The Word Sense Disambiguation Procedure Through Sense Filtering*", International Journal of Instrumentation and Control Systems (IJICS) Vol.3, No.4, pp. 29-41, 2013.

[4] E. Agirre & G. Rigau, "*Word sense disambiguation using conceptual density*", In the Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark, pp. 16-22, 1996.

[5] E. Agirre, & D. Martínez, "*Learning class-to-class selectional preferences*", In the Proceedings of the Conference on Natural Language Learning, Toulouse, France, pp. 15–22, 2001.

[6] S. Banerjee & T. Pedersen, "*An adaptive Lesk Algorithm for Word Sense Disambiguation Using WordNet*", In the Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, London, UK, pp. 136-145, 2002, ISBN: 3-540-43219-1..

[7] G. Escudero, L.M`arquez and G. Rigau, "*Naïve Bayes and Exemplar-based approaches to Word Sense Disambiguation Revisited*", In the Proceedings of the 14th European Conference on Artificial Intelligence, pp. 421-425, 2000.

[8] R. Navigli, "*word sense disambiguation: a survey*", ACM computing surveys, 41(2), ACM press, pp. 1-69, 2009.

[9] E. Agirre and A. Soroa, "*Personalizing PageRank for Word Sense Disambiguation*," In the Proceedings of the 12th Conference European Chapter of the Association for Computational Linguistics, Greece, pp. 33–41, 2009.

[10] R. Mihalcea and D.I. Moldovan, "*Pattern Learning and Automatic Feature Selection for Word Sense Disambiguation*", In the Proceedings of the Second international Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Texas, pp. 127-130, 2001.

[11] R. Navigli and P. Velardi, "*Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation*", Ieee Transactions On Pattern Analysis And Machine Intelligence, Washington, DC, USA, Vol. 27, No. 7, 2005.

[12] X. Zhou and H. Han, "*Survey of Word Sense Disambiguation Approaches*", In the Proceedings of the 18th International FLAIR Conference, American Association for Artificial Intelligence, Philadelphia, pp. 307-313, 2005.

[13] D. Chiang, "*A hierarchical phrase-based model for statistical machine translation*", In Proceedings of the ACL-05, USA, pp. 263–270, 2005.

[14] M. Galley, M. Hopkins, K. Knight, and D. Marcu. "*What's in a translation rule?*", In the Proceedings of the NAACL-04, pp. 273–280, 2004.

[15] K. Philipp, et.al, "*Moses: Open source toolkit for statistical machine translation*", In the Proceedings of the ACL, Demonstration Session, pp. USA, 177–180. 2007.

[16] F. Och and H. Ney, "*A systematic comparison of various statistical alignment models*", International Journal of Computational Linguistics, USA, Vol. 29, Issue 1, pp. 19–51, 2003.

