International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020
ISSN: 1475-7192

A KNOWLEDGE BASED APPROACHFOR WORD SENSE DISAMBIGUATION OF TELUGU LANGUAGE

[1]Pasupuleti Ranjith Kumar, [2]K Vinith Reddy,[3]M SenthiRaja

***ABSTRACT**--Words in the Natural language often correspond to different meanings in different contexts. Such words are referred to as polysemous words i.e. words having more than one sense. A knowledge based algorithm is proposed for disambiguating Telugu polysemous words using computational linguistics tool, Word Net. The task of word sense disambiguation requires finding out the similarity between the target word and the nearby words. In this algorithm similarity is calculated either by finding out the number of common words (intersection) between the glosses (definitions/meanings) of the target and nearby words, or by finding out the exact occurrence of the nearby word's sense in the hierarchy (hypernyms) of the target word's senses. The above two parameters are modified by computing intersection using not only the glosses but also by including the related words. Also the intersection is computed for the entire hierarchy of the target and nearby words. It also includes a third parameter 'distance' which measures the distance between target and nearby words. The proposed approach incorporates more parameters for calculating similarity, which has not been attempted by any of the previous approaches. It scores the senses based on the overall impact of three parameters i.e. intersection, hierarchy and distance and then chooses the sense with the highest score. The correct sense of Telugu polysemous word would be identified with this approach*

***Keywords--** Knowledge based Approach for Word SenseDisambiguation of Telugu Language*

I. INTRODUCTION

In the last two decades, the NLP community has witnessed an increasing interest in ma- chine learning based approaches for automated classification of word senses. This is evident from the number of supervised WSD approaches that have spawned. Today, the supervised approaches for WSD possibly are the largest number of algorithms, used for disambiguation. Supervised WSD uses machine learning techniques on a sense-annotated data set tclassify the senses of the words. There are a number of classifiers also called word experts that assign or classify an appropriate sense to an instance of a single word. The training set for these algorithms consist of a set of examples, where the target word is manually tagged with sense from a reference dictionary. The supervised algorithms thus perform target-word WSD. Each algorithm uses certain

[1]*RA1611003011234, Department of Computer science and Engineering, SRM Institute of Science and Technology Chennai, India, ranjith.3579@gmail.com*
[2] *RA1611003011281, Department of Computer science and Engineering, SRM Institute of Science and Technology Chennai, India, k.inithreddy8@gmail.com* ,
[3] *Assistant Professor, Department of Computer science and Engineering, SRM Institute of Science and Technology Chennai, India, Senthilm6@srmist.edu.in*

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

features associated with a sense for training. This very fact forms the common thread of functionality of supervised algorithms. In this section we will discuss the notable supervised algorithms for sense disambiguation in the literature.

II. LITERATURE SURVEY

Natural Language Processing, Sentiment Analysis and Clinical Analytics,February 2019,DOI: 10.1016/B978-0-12-819043-2.00003-4 By Adil Rajput

Recent advances in Big Data has prompted health care practitioners to utilize the data available on social media to discern sentiment and emotions expression. Health Informatics and Clinical Analytics depend heavily on information gathered from diverse sources. Traditionally, a healthcare practitioner will ask a patient to fill out a questionnaire that will form the basis of diagnosing the medical condition. However, medical practitioners have access to many sources of data including the patients writings on various media. Natural Language Processing (NLP) allows researchers to gather such data and analyze it to glean the underlying meaning of such writings. The field of sentiment analysis (applied to many other domains) depend heavily on techniques utilized by NLP. This work will look into various prevalent theories underlying the NLP field and how they can be leveraged to gather users sentiments on social media. Such sentiments can be culled over a period of time thus minimizing the errors introduced by data input and other stressors. Furthermore, we look at some applications of sentiment analysis and application of NLP to mental health. The reader will also learn about the NLTK toolkit that implements various NLP theories and how they can make the data scavenging process a lot easier.

A novel approach to word sense disambiguation inBengali language using supervised methodologyNov 2019,Alok Ranjan Pal,Diganta Saha,Niladri SekharDash,Antara Pal

An attempt is made in this paper to report how a supervised methodology has been adopted for the task of Word Sense Disambiguation (WSD) in Bengali with necessary modifications. At the initial stage, four commonly used supervised methods, Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naïve Bayes (NB), are dev

Incorporating HowNet-Based Semantic RelatednessInto Chinese Word Sense Disambiguation, January 2020DOI: 10.1007/978-3-030-38189-9_38In book:Chinese Lexical Semantics, pp.359-370

This paper presents a semi-supervised learning method that incorporates sense knowledge into a Chinese word sense disambiguation (WSD) model. This research also effectively exploits HowNet-based semantic relatedness in order to leverage system performance. The proposed method includes Sense Colony task for improving context expansion and semantic relatedness calculating for sense feature representation. To incorporate sense knowledge into WSD, this paper employs the Semantic relatedness in a semi-supervised label propagation classifier. This research demonstrates state-of-the-art results on word sense disambiguation tasks.

The Lesk's algorithm used by overlap based approach can be stated as if W is a word creating disambiguation, C be the set of words in the context collection in the surrounding, S be the senses for W, B be the bag of words derived from glosses, synonyms, hyponyms, glosses of hyponyms, example sentences, hypernyms, glosses of hypernyms, meronyms, example sentence of meronyms, example sentence of hypernyms, glosses of meronyms

then use the interaction similarity rule to measure the overlap and output the sense which is the most probable having the maximum overlap

A Robust Learning Approach for Text Classification , January 2007 By Viet Ha-Thuc,Padmini Srinivasan Previous learning approaches often assume that every part of a positive training document of a class is relevant to that class. However, in practice, it is often the case that only one or a few parts in the training document are really relevant to the class. To overcome this limitation, we propose another learning approach based on relevance-based topic model, an extension of well-known Latent Dirichlet Allocation. In this approach, the real relevant parts in each document are automatically determined byits statistical correlation to the rest of the positive training set. And only these parts contribute to the final results. Therefore, the approach is robust to "impurities" in the training sets. In addition, the approach exploits the "bag-of-words" assumption to rearrange words in an appropriate order that could reduce the computational complexity of learning algorithm. 1 Introduction Machine learning techniques are popular in text classification. However, most of previous approaches, including both supervised ([9][7]) and semi-supervised learning ([11]), assume that every part of a positive training document of a class is relevant to that class. This assumption is, nonetheless, not true in many cases. For instance, when one builds a training set for topic "machine learning", a positive example could be a paper about speech recognition that uses some machine learning technique and another positive document could be an overview article about artificial intelligence that contains the term "machine learning". Therefore, in such cases only one or a few parts in the document are really about machine learning.

III. WSD USING ROGET'S THESAURUS CATEGORIES

Roget's thesaurus is an early Nineteenth century thesaurus which provides classifica- tion or categories which are approximations of conceptual classes. This algorithm by Yarowsky (1992) uses precisely this ability of Roget's thesaurus to discriminate between the senses using statistical models. The algorithms observes following:

IV. BILINGUAL WSD

The limited performance of monolingual approaches to deliver high accuracies for all- words WSD at low costs created interest in bilingual approaches which aim at reducing the annotation effort. Here again, the approaches can be classified into two categories, viz., (i) approaches using parallel corpora and (ii) approaches not using parallel corpora.

The approaches which use parallel corpora rely on the paradigm of Disambigua- tion by Translation, described in the works of Gale et al. (1992), Dagan and Itai (1994), Resnik and Yarowsky (1999), Ide et al. (2001), Diab and Resnik (2002), Ng et al. (2003), Tufis et al. (2004), Apidianaki (2008). Such algorithms rely on the frequently made observation that a word in a given source language tends to have different translations in a target language depending on its sense. Given a sentence-and-word-aligned parallel corpus, these different translations in the target language can serve as automatically acquired sense labels for the source word.

In this work, we are more interested in the second kind of approaches which do not use parallel corpora but rely purely on the in-domain corpora from two (or more) languages. For example, Li and Li (2004) proposed a bilingual bootstrapping approach for the more specific task of Word Translation Disambiguation (WTD) as opposed to the more general task of WSD. This approach does not need parallel corpora (just like our approach) and relies only on in-domain corpora from two languages. However, their work was evaluated only on a handful of target words (9 nouns) for WTD as opposed to our work which focuses on the broader task of all-words WSD. Supervised algorithms train a model based on the annotated corpus provided to it. This corpus needs to be manually annotated, and the size of the corpus needs to be large enough in order to train a generalized model.Semi-supervised, also known as minimally supervised algorithms make some assump- tions about the language and discourse in order to minimize these restrictions. The common thread of operation of these algorithms are these assumptions and the seeds used by them for disambiguation purposes.This section presents two such approaches, based on two different ways to look at the problem, namely Bootstrapping and Monosemous Relatives.

V. LINS APPROACH

Lin (1998) clusters two words if they share some syntactic relationship. More the relation, more close the words are situated in the cluster. Given context words w1, w2,,wn and a target word w, the similarity between w and wi isdetermined by the information content of their syntactic features.

The previous approach uses context vectors, which conflate senses of words, and thus, similarity of w with each wi can not be determined with that approach. Therefore, each word is represented in form of a vector. The information contentsare then found out using the syntactic features as mentionedpreviously

Word Clustering

Approaches Context vectors previously explained, are second-order representations of word senses, as in they represent the senses indirectly. The idea here is to cluster the senses based on word vectors, in order to draw out the semantic relationships between the words.

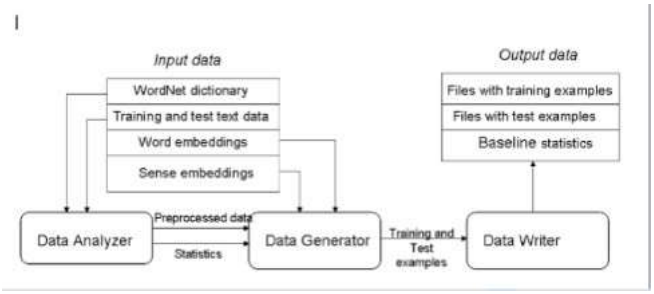
VI. PROPOSED SYSTEM

One of the important tasks of word sense disambiguation is to compute the similarity or relatedness of two words. Similarity of the target and nearby word is dependent on the following parameters in the proposed approach

Intersection between word families

Hierarchical Relationship (Hypernyms and Hyponyms) Distance

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020



VII. DATASETS

For this project, I use the publicly available Google Word Sense Disambiguation Corpora to train my sense-tagged word vectors as well as train and evaluate my LSTM models. The Google WSD Corpora, released on January 17, 2017, is one of the largest labeled WSD corpora, and consists of the popular SemCor and MASC datasets manually labeled with NOAD and WordNet senses.

Google commissioned the labeling of these datasets by having expert linguists label a small seed set used as a gold standard, and then having many other workers label the remainder of the datasets. In developing the corpus, Google prioritized having a high inter-rater reliability score to ensure high quality of tagged tokens. They achieved a Krippendorff's Alpha score of 0.869, implying the labeling are highly reproducible (usually a score above 0.67 is considered acceptable). However, as a result, despite having

1.1 million tokens, only 248k are polysemous tokens labeled with word senses, with many polysemous tokens instead being tagged with an ambiguous sense (since the reliability score for the tags on these tokens is lower than their standard).

For my experiments with LSTM models, I train the LSTM with at least the entire MASC dataset, with some experiments having the LSTM be trained with as much as ¾ of the SemCor dataset.

VIII. METHDOLOGY

Evaluation Metric

The testing set for evaluation my approaches consists of 30k tokens from the SemCor dataset. I use cosine similarity to evaluate the models. The fewer cosine similarity between the predicted sense vector and the actual sense vector, the more close these two vector at the vector space.

Experimental settings

The hyperparameter settings used during the experiments, presented in Table 1, were tuned on a separate validation set with data. The source code, implemented using Keras with TensorFlow (Abadi et al., 2015) backend, has been released as open source.

Embeddings

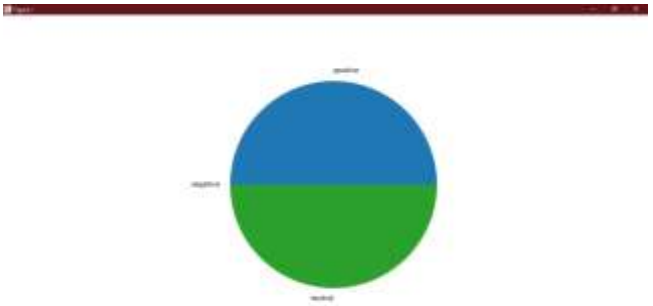
The embeddings are initialized using a set of freely available GloVe vectors trained on Wikipedia and Gigaword. Words not included in this set are initialized from N(0,0.1). To keep the input noise proportional to the embeddings it is scaled by σ_i which is the standard deviation in embeddingdimension I for all words in the embeddings matrix, is updated after each weight update.

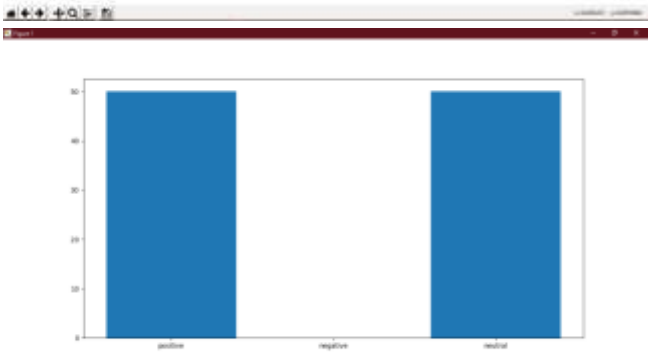
Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

IX. DATA PREPROCESSING

The only preprocessing of the data that is conducted is replacing numbers with a<number> tag. Words not present in the training set are considered unknown during test. Further, I limit the size of the context to max 140 centered around the target word to facilitate faster training.

X. RESULT





The minimum value loss tells about the rate of accuracy that has been concluded

XI. CONCLUSION

in this project we considers three parameters for calculating the similarity between target and nearby words. Similarity is calculated by computing intersection between word families along the entire hierarchy of the target and nearby word. Also the distance is combined with intersection and level to compute a score for all senses, corresponding to every target-nearby pair

REFERENCES

1. Manning, C. and Schutze, H. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.
2. Jurafsky, D. and Martin, J. Speech and Languag Processing. Prentice Hall, Upper Saddle River, NJ,2000.
3. BasakMutlum,WordSenseDisambiguationhttp://www.denizyuret.com/students/bmutlum/index. htm
4. Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, B. Slator, Providingmachine tractable dictionary tools, MachineTranslation 5, 99–154,1990.
5. Manish Sinha, Mahesh Kumar, Prabhakar Pande, Lakshmi Kashyap and Pushpak Bhattacharyya, , Hindi Word Sense Disambiguation, International Symposium on Machine Translation, Natural Language
6. Processing and Translation Support Systems, Delhi, India, November, 2004 http://www.cse.iitb.ac.in/~pb/papers/HindiWSD.pdf
7. M. Quillian, Semantic memory, in: M. Minsky (Ed.), Semantic Information Processing, the MIT Press, Cambridge, MA, pp. 227–270,1968.
8. J. Cowie, J. Guthrie, L. Guthrie, Lexical disambiguation using simulated annealing, in: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, pp.359–365,1992.
9. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French,David YarowskyACL1994
10. Navigli R 2009 Word sense disambiguation: a survey.ACMComput. Surv. 41(2): 1–69
11. Xiaojie W and Matsumoto Y 2003 Chinese word sense disambiguation by combining pseudo training data. In: Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, pp. 138–143
12. Sanderson M 1994 Word sense disambiguation and information retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and
13. Development in Information Retrieval, SIGIR'94, July03–06, Dublin, Ireland. New York: Springer, pp.142–151Eneko Agirre E and Edmonds P (Eds.) Word SenseDisambiguation: Algorithms and Applications
14. Lei Guo, Xiaodong Wang, Jun Fang, 'OntologyClarification by UsingSemantic Disambiguation', 978 - 1-4244-1651-6/08, IEEE 2008.
15. S. P. Ponzetto, R. Navigli, "Knowledge-rich WordSense Disambiguation rivaling supervised systems," Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1522-1531, 2010.
16. Roberto Navigli and Paola Velardi, "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 27, No. 7, pp. 1075-1086, July 2005.

17. Boshra F. ZoponAl_Bayaty and Dr.Shashank Joshi“Word Sense Disambiguation (WSD) and Information Retrieval (IR): Literature Review” ijarcsse,Volume 4, Issue 2,ISSN: 2277 128X, February. 2014.
18. Lucia Specia, Sujay Kumar Jauhar, RadaMihalcea, SemEval 2012 Task 1: English Lexical Simplification, in Proceedings of the SemEval-2012 Workshop on Semantic Evaluation Exercises, Montreal, Canada, June 2012.
19. Kulkarni, M.Comput. Eng. Dept., V.J.T.I., Mumbai, India,Sane, S. “An ontology clarification tool for word sense disambiguation”20113rd International Conference, IEEE - Electronics Computer Technology (ICECT),292 – 296, E-ISBN :978- 1-4244-8679-3, 8-10 April 2011.
20. RadaMihalcea, "Knowledge-Based Methods for WSD", e-ISBN 978-1- 4020-4809-2, Springer 2007.