IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DEEP LEARNING: UNRAVELING WORD MEANINGS WITH LANGUAGE MODELS

[1]Dr. Amit P. Patil, [2]Ms.Chhaya S. Patil, [3]Mr.Vishal A. Pawar

[1]Assistant Professor, [2]Assistant Professor, [3]Assistant Professor
[1,2,3]Department of Computer Science,
[1,2,3]RCPET's Institute of Management Research and Development, Shirpur, India

Abstract:
In Natural Language Processing (NLP), the term "word sense disambiguation" (WSD) pertains to the process of determining the specific meaning or sense of a word based on the context in which it is employed. The term polysemy refers to words that take on a different meaning depending on their context in a sentence. Tie, Bank, Interest, and Book are some examples of polysemy words. In Natural Language Processing, the task of WSD remains an open problem. WSD is easy for humans but challenging for automatic systems.

Several WSD systems have been proposed, and it enables discrete word features to be extracted. These methods use a classifier that is trained using surrounding words and collocations in order to identify the words. By incorporating continuous words of surrounding words, this classifier can be improved. Recent improvements haven't been noticed by any of them. This is due to the fact that all systems use word representations that are independent of the context in which they are used.

Recent research has demonstrated that contextualized word embeddings enhance several NLP tasks. BERT (Bidirectional Encoder Representations from Transformers) contains pre-trained contextualized word representations. BERT identifies the word that is most likely to be in a word that has been hidden in a sentence.

In this paper we are giving introduction to transformers in NLP and the BERT (Bidirectional Encoder Representations from Transformers) model. We also explained here the limitations of transformer and work done using BERT model for WSD.

IndexTerms **- BERT, NLP, Deep Learning, Language Models**

I. INTRODUCTION

Word Sense Disambiguation is employed when a word with more than one meaning appears in a text. Determining the appropriate context for a word in a sentence or other piece of writing is a task carried out by a computer system. Ambiguity is common in natural language. Lexical ambiguity is dealt with by word sense disambiguation (WSD), this is referred to as polysemy in the sentences. Examples of polysemy terms are Bank, Book, Interest, Mouse, etc. Humans can quickly determine a word's appropriate meaning given the context, but WSD is a difficult work for automatic machines, and it is still unsolved in NLP. To appropriately determine the senses, systems must first apply various algorithms. Take these sentences as an example. Bass is one of my favorite foods, and she plays bass frequently. A musical instrument and a food-related item, or anything that may be eaten, are the two meanings of bass, respectively.

Three procedures are used in the fundamental WSD task. Because there are several words in a context, we must first rank their senses. Ranking is therefore required to understand the significance of each word in the context. The second step is to choose the window to be considered, which means the length of the sentence is considered in the same way that Bag of Words. Finally, there is the use of a knowledge base such as WordNet. WordNet is the most used sense inventory in English. WordNet is a lexical database created specifically for NLP. For the WSD task, a pre-trained contextualized word representation, such as the BERT model, is used. Consider the sentence "Artificial intelligence should always [MASK] humans," which contains a word that has been masked. According to BERT, the most likely word in the masked position is "help." This demonstrates that BERT has a profound comprehension of a wide range of sentences or contexts, and it is clear that this understanding will be very beneficial to the WSD system. The system takes a sentence and an ambiguous word as input, and it outputs the word's target sense.

Word Sense Disambiguation is an ongoing research topic in natural language processing, with various approaches being used. Determining how much information to employ for the most accurate disambiguation is a difficult part of the WSD endeavor. Another difficulty is determining whether the word in a particular context should be disambiguated for a more general sense or for a finer sense. Furthermore, it can be challenging to pinpoint a word's senses as well as the level of information that each

meaning represents in terms of usage. Word Sense Disambiguation (WSD) is an open problem in computational linguistics concerned with determining which sense of a word is used in a sentence. The solution to this problem has ramifications for other computer-related writing, such as discourse, search engine relevance, anaphora resolution, coherence, and inference. The human brain is very good at understanding word meanings. It reflects this neurological fact that natural language is constructed in a way that demands so much of it [1]. To put it another way, the success of human language reflects the inherent potential offered by the neural networks of the brain, as well as a contribution to its shaping.

The field of natural language processing (NLP) and machine learning has presented significant challenges within computer science and the associated information technology. Numerous methodologies have been explored to address these challenges, including dictionary-based approaches that utilize lexical resources, supervised machine learning techniques that employ classifiers trained on manually sense-annotated examples, and unsupervised approaches that infer word senses by grouping word occurrences. Among these methodologies, supervised learning has emerged as one of the most successful algorithms thus far.

II. THE TRANSFORMER MODEL IN NLP

Transformers are one invention that has propelled natural language processing to new heights in the last three years. Transformers are semi-supervised machine learning models that are mostly employed with text input and have largely supplanted RNN (recurrent neural networks) in NLP tasks.

The Transformer is a unique architecture in NLP that tries to handle long-range relationships while resolving tasks that involve sequences to sequences. In this context, the term "transduction" pertains to the conversion of input sequences into output sequences. The Transformer architecture addresses the issue of handling dependencies between input and output by utilizing attention and recurrence mechanisms in a comprehensive manner.

The Transformer architecture comprises two key components: an encoder and a decoder. The encoder is responsible for processing input sequences, while the decoder is employed during training to process target output sequences and predict subsequent items in the sequence. The embeddings, the positional encoding block, and the multi-head attention blocks are the components of a transformer that are particularly crucial. DistilBERT, T5, BERT, GPT-2, and other newer creations are some of the most popular transformer models.

Transformer's Limitations

- The attention mechanism in NLP-based systems has a limitation in that it can only process text strings of a specific length. Therefore, before inputting the text into the system, it needs to be divided into segments or chunks of a predetermined size. This ensures that the attention mechanism can effectively process and analyze the input text.
- Contextual fragmentation occurs as a consequence of text chunking, leading to the loss of significant contextual information. When a sentence is divided in the middle or without considering sentence boundaries and other semantic divisions, a substantial amount of context is compromised.

III. RELATED WORK

While the challenge of Word Sense Disambiguation (WSD) remains to be addressed, significant advancements have been made in this field over the years. One notable progress is the utilization of unsupervised or minimally supervised models alongside highly supervised models, which represents a crucial development in computational techniques compared to earlier decades. This transition has been greatly influenced by the emergence of machine learning and the advancement of AI-based algorithms. The following section highlights several notable approaches that researchers have undertaken in recent years.

The authors Faralli et al., (2012) of this model leverage domain glossaries produced through repeated bootstrapping rather than an annotated corpus to present an unsupervised strategy for Domain WSD with a minor restriction on the uniqueness of the chosen relations. It offers answers for two different levels of sense granularity: one that is exceedingly fine-grained and one that is relatively coarse-grained. They employ domain boosted PPR algorithms and Personalized PageRank (PPR) to choose the most appropriate gloss. Both algorithms gave good results as 69% and 80% F1-scores [2].

A notable contribution in the field of Word Sense Disambiguation (WSD) was made by Basil et al. (2014), who introduced an unsupervised WSD algorithm. This algorithm aimed to enhance the existing knowledge-based WSD algorithms by improving the gloss-context overlapping technique. The authors employed annotated data known as Babel-Net, which combines WordNet and Wikipedia. By generating semantic vectors based on the Expanded Gloss in the Distributional Semantic Space and calculating their cosine similarity, this technique focused on identifying the highest similarity rather than relying solely on simple overlapping. The algorithm achieved a 70% F1-score measure [3].

Yuan et al. (2016) proposed the utilization of two distinct neural models in their research. One model incorporated Long Short-Term Memory (LSTM) while the other employed a semi-supervised technique called Label Propagation (LP). The first algorithm, trained on unlabeled text, utilized similarity calculations to identify the context words within the corpus and predict their meanings. The second algorithm initially utilized labeled sentences and then expanded them by including unlabeled sentences. In a graphical approach, cosine similarity was employed to predict contexts using vertices representing all the sentences. In terms of the All-words F1 scores in Sem-Eval 2013, the LSTM technique achieved the highest scores, except for specific tasks. When combined with an LSTM language model, the LSTM LP classifier yielded the best results for nouns and adverbs [4].

Raganato et al. (2017) developed a WSD system that utilized a supervised approach, treating the task as a sequence learning problem, which differed from the conventional classification-based approaches focusing on individual words. The researchers employed three progressive algorithms: Bidirectional LSTM (BLSTM) architecture, an Attentive Bi-directional architecture incorporating an additional attentive layer, and a Sequence-to-Sequence (Seq2Seq) Model based on Recurrent Neural Networks, which employed separate encoder and decoder components. The results obtained from the BLSTM and Seq2Seq models were either superior to or on par with the performance of the best-performing supervised models [5].

IV. BERT WORD SENSE DISAMBIGUATION

The BERT framework, developed by Google AI, employs the techniques of pre-training and fine-tuning to create state-of-the-art language representation models for various tasks. These tasks include language inference, sentiment analysis, and question-answering algorithms. Using the BERT approach, this section of the research reviewed previous work in word sense disambiguation.
BERT utilizes a multi-layer bidirectional Transformer encoder as its architecture. The self-attention layer within BERT performs self-attention in both directions. Google has released BERT in two variants, providing users with different options to leverage the model's capabilities.

- BERT Base: 110M parameters total, layers of transformers=12.
- BERT Large: Total Parameters = 340M, Transformer layers-24.

In BERT, bidirectionality is achieved through pre-training on two tasks: Masked Language Model and Next Sentence Prediction. This approach allows the model to capture contextual dependencies in both forward and backward directions.

In a study conducted by Luyao Huang et al. (2019), a WSD system was developed utilizing three BERT-based models: GlossBERT (Token-CLS), GlossBERT (Sent-CLS), and GlossBERT (Sent-CLS-WS). These models were fine-tuned specifically for the WSD task by constructing gloss pairs as inputs. The training process involved utilizing the SemCor 3.0 English all words dataset, and the performance of the models was evaluated on various WSD datasets, including Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), and SemEval-2015 (SE15) English all words datasets. The experimental results demonstrated that the proposed models achieved state-of-the-art performance [6].

In the research conducted by Jiaju Du et al. (2019), a WSD system was introduced that incorporated various approaches for combining the BERT base model and the required classifier, referred to as BERT def. Additionally, a uniform classifier was trained using sense definitions, enabling the model to differentiate between unseen polysemes. The system was trained on the SemCor 3.0 dataset for English all words, with SE7 serving as the validation dataset. The performance evaluation was conducted on the SE2, SE3, SE13, and SE15 datasets.

Daniel Loureiro et.al performed extensive experiments on the BERT model and to train and test the BERT model they developed dataset CoarseWSD-20 noun ambiguity. They got more than 90% F1 score for disambiguating nouns. This work was done for the English language. They also concluded that feature extraction along with fine tuning the BERT model gave the best result of WSD [8].

Anu P C and Rameez Mohammed (2020) proposed a WSD system using BERT model for English language, this system perform disambiguation is two stages, first stage was to predict the word using BERT base model and in second stage used WordNet to find the senses of target word or predicted word. Masked Language Model of BERT was used for this system [9].

Vandenbussche, P., et.al (2021) treated WSD as a binary classification problem. The authors used the BERT base uncased transformer model for the English WSD. The target word was given in context (input sentence) and the sense of the word was separated by a special token ([SEP]). Originally, this configuration was used to predict whether sentences would follow one another in a text. Due to the transformer architecture learning power, they learned this new task by simply changing the fields' meanings and keeping the structure of the input data the same. They added a fully connected layer on top of the transformer model's layers with a classification function for predicting the target word in context matches the definition. In addition to being well suited to weak supervision, this approach also generalizes to word-sense pairs not previously encountered in training. Word-in-Context Target Sense Verification (WiC-TSV) dataset was used for training and testing purposes. This system gave the 75.1% F1 score [10].

Avi Chawla, et.al (2021) presented a comparative and analytical work on WSD using nine transformer model namely BERT, CTRL, OpenAI-GPT,Transformer-XL, OpenAI-GPT2, DistilBERT, ALBERT, ELECTRA and XLNet. K-NN classifier approach on CWEs (Contextualized Word Embeddings) for performing the WSD. These models were tested using 2 lexical sample datasets SensEval-2 and SensEval-3. The models were tested majorly on POS like Nouns, Verbs and Adjectives. They ended their investigation by claiming that, based only on the text encodings these models provide, the BERT, ALBERT and DistilBERT models demonstrate to be the most successful on the WSD challenge [11].

V. CONCLUSION

Natural language processing tasks can be handled by transformers, which are powerful deep learning models. Furthermore, we can fine tune pre-trained transformers to perform our own natural language processing tasks. Machine Learning for Natural Language Processing has undoubtedly made a breakthrough with BERT. As a result of its approachability and ability to allow fast fine-tuning, it is likely that in the future we will be able to use it in a wide range of practical applications. The utilization of advanced architectures like Transformers and BERT sets the stage for further advancements in the future. These breakthroughs are anticipated to incorporate even more advanced technologies, enabling significant progress in the field of NLP.

REFERENCES

[1] D. Yuan, J. Richardson, R. Doherty, C.Evans, and E. Altendorf, "Semisupervised word sense disambiguation with neural models," arXiv preprint arXiv:1603.07012, 2016.

[2] Faralli et al., 2012, July. "A new minimally-supervised framework for domain word sense disambiguation". In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1411-1422).

[3] Basile et al., 2014, August. "An enhanced lesk word sense disambiguation algorithm through a distributional semantic model". In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1591-1600)

[4] Yuan et al., 2016. "Semi-supervised word sense disambiguation with neural models." arXiv preprint arXiv:1603.07012

[5] Raganato et al., 2017, September. "Neural sequence learning models for word sense disambiguation". In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp.1156-1167).

[6] Luyao Huang, "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge", EMNLP-IJCNLP 2019,https://doi.org/10.48550/arXiv.1908.07245

[7] Jiaju Du, "Using BERT for Word Sense Disambiguation" ,arXiv:1909.08358v1,18 Sep.2019

[8] Daniel Loureiro, et. al, "Language Models and Word Sense Disambiguation: An Overview and Analysis", arXiv:2008.11608v1,26 Aug 2020

[9] Anu P C, Rameez Mohammed, "BERT Approach for Word Sense Disambiguation", International Journal of Advances in Engineering and Management (IJAEM) Volume 2, Issue 3, pp: 370-373

[10] Vandenbussche, P., Scerri, T., & Jr., R. (2021), "Word Sense Disambiguation with Transformer Models". Workshop on Semantic Deep Learning.

[11] Avi Chawla, et. al, "A Comparative Study of Transformers on Word Sense Disambiguation", arXiv:2111.15417v1,30 Nov 2021

[12] A. Vaswani, N. Shazeer, et. al, "Attention Is All You Need", (2017), 31st Conference on Neural Information Processing Systems.

[13] F. Chaubard, M. Fang, et. al, "Word Vectors I: Introduction, SVD and Word2Vec", (2019), CS224n: Natural Language Processing with Deep Learning lecture notes, Stanford University.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", (2018), arXiv.org.

[15] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", (2019), arXiv.org.

[16] C. Raffel, N. Shazeer, et. al, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", (2019), arXiv.org.

[17] A. Radford, J.Wu, et. al, "Language Models are Unsupervised Multitask Learners", (2019), OpenAI.