DOI: 10.21917/ijsc.2023.0403

BERT AND INDOWORDNET COLLABORATIVE EMBEDDING FOR ENHANCED MARATHI WORD SENSE DISAMBIGUATION

Sandip S. Patil, R.P. Bhavsar and B.V. Pawar
School of Computer Sciences, K.B.C. North Maharashtra University, India

Abstract

Ambiguity in word meanings is a long-standing challenge in processing natural language. Word sense disambiguation (WSD) deals with this challenge. Prior neural language models make use of recurrent neural network and architecture with long short-term memory. These models process the words in sequence, are slower and not truly bi-directional, so they are not able to capture and represent the contextual meanings of the words, hence they are not competent in contextual semantic representation for WSD. Recent, Bi-Directional Encoder Representation from Transformers (BERT) is long short-term memory-based transformer model that is deeply bi-directional. It uses attention mechanisms, which process and use the relevance of the entire context at a time in both directions, so it is well suited to leverage the meanings in distributed representation for WSD. We have used BERT for obtaining contextual word embedding of context and sense gloss of Marathi language ambiguous word. For this purpose, we have used 282 moderately ambiguous Marathi words catering to 1004 senses distributed over 5282 Marathi sentences harvested by linguists from online Marathi websites. We have calculated semantic similarity between the pair of context and gloss embedding using Minkowski distance family and cosine similarity measures and assigned plausible sense to the given Marathi ambiguous word. Our empirical evaluation shows that the cosine similarity measure outperforms and yields an average disambiguation accuracy of 75.26% for the given Marathi sentence.

Keywords:

BERT, Distributional Semantics, Neural Language Modeling, Transfer Learning, Word Sense Disambiguation

1. INTRODUCTION

Ambiguity in word meanings is a long-standing challenge in the processing of natural languages. This problem is addressed by word sense disambiguation (WSD) techniques, which resolve the ambiguities in the word meanings and assign the most appropriate sense to the ambiguous word by looking at the context [1]. In the literature, various shades of WSD approaches for various languages are elaborated [2], [3]. Knowledge-based WSD has smooth portability but is knowledge lean. Supervised WSD approaches perform better in terms of accuracy but need computational resources, which consequently limit scalability and portability. The gap between prior knowledge and the availability of the labeled data in supervised approaches is bridged by unsupervised WSD, which generates distributional semantics by exploiting contextual features. Real-valued representations of distributed word semantics and contextual relations are called ‘contextual word embedding’. Polysemous words have different representations in different contexts.

The Transformer encoder is used to generate the contextual word embedding. Word embedding has enormous applications in various NLP tasks and, likewise, it has great potential for unsupervised WSD, especially for digitally resource-scarce languages like Marathi. Marathi is the official language of Maharashtra and Goa states in India and is ranked 10th in the world’s most spoken languages. A more advanced version of the transformer encoder model, i.e., BERT, is a stack of bi-directional encoders [4]. We have used BERT with a multi-head attention mechanism for obtaining contextual word embedding from the AI4Bharat IndicBERT Marathi Model and CFIL IIT Bombay’s IndoWordNet synset [5]. These word embedding are used for Marathi WSD as a downstream application. Detailed discussions on BERT and WSD using similarity measures with prior art, BERT philosophy, and empirical results are discussed in the following sections.

2. PRIOR ART

Distributional semantic representations help in generating the word embedding(s), which give real-valued context dependent vector representation for a word in different contexts. Recently of-the-shelf contextual word embedding(s) for different natural languages have been developed and made available for research and commercial purposes. These have proven to be a notable contribution for various downstream NLP tasks like MT, WSD, QA, IR etc. Previously, off-the-shelf static word embedding techniques like Word2Vec, FastText and GloVe were used for generating word embedding(s), but the advent of transfer learning resulted in contextualized (dynamic) word embedding techniques like ELMo and BERT. Contextual representation of a word is the cognitive view of a distributional semantics [6].

In the light of static word embedding, Raganato et al. [7] proposed bi-LSTM based sequence learning for English WSD, in which the words are labeled with WordNet synset or manually (if WordNet equivalent is not available), this approach reported the accuracy of 73.4% for English WSD. Heo et al. [8] have used Word2Vec technique to obtain static embedding for English. They constructed sense tagged corpus and evaluated the corpus on k-NN-based WSD for English, which reported the accuracy of 63%. Uslu et al. [9] have proposed FastText word embedding technique for German WSD. The FastText-based German WSD yielded the accuracy of 81%. Kageback and Salomonsson [10], Vial et al. [11] and Luo et al. [12] and Luo et al. [13] have used bi-LSTM-based GloVe embedding technique for English WSD, which reported the accuracy of 73.4%, 67.1%, 73.2% and 70.6%, respectively. Since these WSD approach uses GloVe, it cannot disambiguate the unknown or out-of-vocabulary ambiguous words.

As static word embedding ignores variability of word meaning in different contexts and obtains same embedding for the same word pair in different contexts [14], so it is not able to capture the polysemy and thus limits the performance for WSD. In dynamic contextualized word embedding techniques like ELMo [15] and BERT [4] both aims to capture the context dependent meaning of the word and addresses the issue of polysemy for WSD. Kumar et al. [16] have used ELMo model for English WSD, they have represented the vocabulary words as well as out-of-vocabulary words in the context of ambiguous words and WordNet gloss, so this ELMo based WSD provides the ability to disambiguate seen as well as unseen words and has reported the accuracy of 31.2% even for rare senses.

As BERT uses bi-LSTM transformer architecture, it is purely bi-directional, so achieved better performance in different NLP tasks including WSD [17]-[21]. First time, Du et al. (2019) [17] fine-tuned BERT embedding for WSD, they have trained multiplayer perceptron (MLP) classifier on BERT embedding(s) for English WSD, MLP based WSD approach disambiguates only the linearly separable embedding(s) of the given polysemy word, so not suitable for fine grained sense disambiguation and reported the accuracy of 76.3%. Hadiwinoto et al. (2019), [19] have incorporated nearest neighbor sentences' matching on pre-rained BERT embedding for English and Chinese WSD, apart from the target sentence, neighbor matching–based WSD also needs embedding of left and right neighboring sentences, consequently slower the performance of WSD, reported the accuracy of 80% and 89.5% for English and Chinese WSD respectively. Vial et al. [20] have explored various WordNet relations on contextualized BERT word vectors for the task of English WSD, this approach uses softmax neural classifier, so needs large sense annotated corpora and reported the accuracy of 90.6%. Huang et al. [18] have applied the neural-based binary classification task on the pair of context and gloss BERT embedding for English WSD, it does not allow flexible membership to the probable senses also it does not explores more than four sense for the given target word and reported the accuracy of 77%. Stoeckel et al. [21] have applied the SenseFitting classification task on the BERT embedding of gloss and relation over the context of the target for German WSD, they have fixed the context window size, so limits the applicability of distributional semantics for WSD and reported the accuracy of 56.28%. Bevilacqua and Navigli [22] constructed graph and adjacency matrix from the WordNet relations and employed of-the-shelf BERT embedding[s], trained on English embedding[s] for cross lingual WSD and evaluated the system on French, German, Italian and Spanish WSD, the cross lingual WSD system is over–relies on corpus supervision, reported the accuracy of 80% for seen and un-unseen synsets.

For digitally resource scare Indian languages like Hindi and Marathi, Bhingardive et al. [23], have used E-M likelihood on the Word2Vec static word embedding technique for cross lingual Hindi-Marathi WSD on Health domain, these word embedding technique not only fails to capture the distributional semantics but also unable to handle the polysemy and out-of-vocabulary words, consequently results poor performance for all PoS WSD and reported the accuracy of 60.94% and 61.30% for Hindi and Marathi WSD respectively.

All the attempts for neural WSD in the literature proved that, incorporating gloss knowledge into supervised WSD approach is helpful, but still not achieved much improvement, because it may not be able to fully explore the context and gloss representations. In this paper, we focus on how to better leverage BERT-based gloss, context representation and similarity measures in an un-supervised neural-based Marathi WSD system.

3. MOTIVATION

Marathi is a morphologically rich, highly inflectional, but digitally resource-scarce Indian language. As compared to other Indian languages, it has more word-sense ambiguity [24]. WSD is a classical step, needed in all the Marathi NLP tasks. In the literature, it has been observed that most of the authors disambiguated words in English, French, Chinese, etc., foreign languages. Efforts in this regard for Marathi are at a preliminary stage. Contextualized word embedding like BERT provides a condensed set of features for downstream NLP applications, so contextual word embedding will be an important resource for the Marathi WSD task. In this study, we have explored various similarity measures on pre-trained BERT context and gloss embedding for downstream Marathi nouns, verbs, adverbs, and adjectives with WSD.

4. TRANSFORMER ENCODER

Transfer learning is a kind of unsupervised learning that makes use of prior knowledge and solves problems that require similar expertise for other problems/domains [25]. A transformer is a basic architecture for transfer learning, it consists of two processing units, viz., encoder and decoder.

Encoder maps the input representation (for WSD, sequence of input symbols) into intermediate representation (contextual embedding), while decoder generates output representation (sequence of target symbols) from intermediate representation [26].

4.1 SELF-ATTENTION MECHANISM IN TRANSFER LEARNING

As discussed in the prior art, static word embedding(s) techniques are not able to express the variability of contextual meaning as they ignore relevance of context words. Self-attention mechanism captures the relevance of each word with remaining words, so it can express the variability of contextual meaning better and generates more relevant contextual representation for WSD [27], [28]. Self-attention mechanism calculates the attention score for every pair of words and updates the meaning score of the word each by combining global score of the entire context.

Self-attention score calculation between the contextual words is demonstrated in Fig.1. Let C_s=[W_1,W_2, W_{amb}, W_n] be the sequence of n number of words and C_{amb} be the sentential context vectors of the word W_{amb}, where $W_{amb}{\in}R^{512}$. The aim of self-attention is to use the global context information and encode all the entities of C_{amb}. Randomly initialize three learnable query, key and value weight matrices W^Q, W^K, W^V, where W^Q, W^K, $W^v{\in}R^{512\times64}$ are used to project all the vectors in C_{amb} into get Query [Q_n], Key [K_n] and Value [V_n] by matrix multiplication of $W_1{\times}W^Q{=}Q_1$, $W_1{\times}W^K{=}K_1$, $W_1{\times}W^V{=}V_1$, $W_2{\times}W^Q{=}Q_2$, $W_2{\times}W^K{=}K_2$, $W_2{\times}W^V{=}V_2$, $W_n{\times}W^Q{=}Q_n$, $W_n{\times}W^K{=}K_n$, $W_n{\times}W^V{=}V_n$ and so on.

Further the dot products between Query [Q_n] and transpose of Key [K_n^T] are used to generates the attention scalars S_{nn} between the two context word vectors in C_{amb}, attention scalars S_n has the vales in different ranges, so we needs to normalized it in the range of 0 to 1 by softmax function as given in Eq.(1), which generates attention weights w_{nn} between the two word vectors in C_{amb}.

$$w_{nn} = \text{Softmax}\left[S_{nn}\right] = \frac{Q_n K_n^T}{\sqrt{d_k}} \tag{1}$$

where d_k is the dimension of vector K. The weighted summation of attention weights w_{nn} and values V_1, V_2, V_3,…, V_n are then used to generate the contextualized word embedding e_n for the word vectors in C_{amb} as shown in Eq.(2):

$$e_n = \sum_{i=1}^{n} w_{nn} V_n \tag{2}$$

Where n is the number of input context vectors

The Fig.2 shows the working of self- attention in transfer learning, here W_1 is the vector of *word*$_1$, W_2 is the vector of *word*$_2$.

As shown in Fig.1, S_{11} is the relevance of word W_1 to describe itself, S_{22} is relevance of word W_2 to describe itself, S_{12} is relevance of word W_1 to describe word W_2 and S_{21} is relevance of word W_2 to describe word W_1 and so on, where $S_{11}=Q_1 \cdot K_1^T$, $S_{12}=Q_1 \cdot K_2^T$, $S_{13}=Q_1 \cdot K_3^T \ldots S_{nn}=Q_n \cdot K_n^T$ and so on …

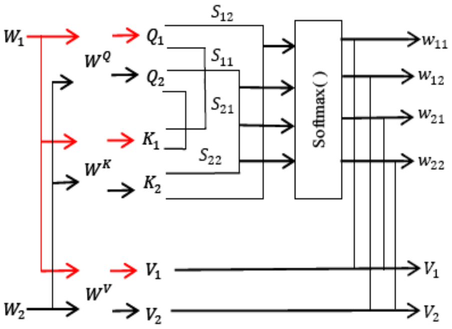

Fig.1. Working of Self-Attention in Transfer Learning

So, $S=(S_1,S_2,S_3,\ldots,S_n)$ are the relevancies of all the word with each other in the context, where s_1 is the relevance of all word to describe the word W_1.

$$S_1=(S_{11},S_{12},S_{13},\ldots,S_{1n})$$
$$S_2=(S_{21},S_{22},S_{23},\ldots,S_{2n})$$
$$S_3=(S_{31},S_{32},S_{33},\ldots,S_{nn})$$

As S_n vales are in different ranges, so we need to normalize it. Softmax(S_n) function is used to normalize the range of 0 to 1.

$$w_{11}=\text{Softmax}[S_{11}],\ldots,w_{1n}=\text{Softmax}[S_{1n}]$$
$$w_{21}=\text{Softmax}[S_{21}],\ldots,w_{2n}=\text{Softmax}[S_{2n}]$$
$$w_{31}=\text{Softmax}[S_{31}],\ldots,w_{3n}=\text{Softmax}[S_{3n}]$$
$$w_{n1}=\text{Softmax}[S_{n1}],\ldots,w_{nn}=\text{Softmax}[S_{nn}]$$

So, $w=(w_1,w_2,w_3,\ldots,w_n)$ are the weights of all the relevance with each other in the context, where w_1 is the weight used to calculate the contextual embedding for the word W_1, w_1 is the weight used to calculate the contextual embedding for the word W_2 and so on.

$$w_1=(w_{11},w_{12},w_{13},\ldots,w_{1n}), \text{ where } \sum_{i=1}^{n} w_{1i} = 1$$

$$w_2=(w_{21},w_{22},w_{23},\ldots,w_{2n}) \text{ where } \sum_{i=1}^{n} w_{2i} = 1$$

$$w_3=(w_{31},w_{32},w_{33},\ldots,w_{nn}) \text{ where } \sum_{i=1}^{n} w_{ni} = 1$$

So, the contextualized embedding(s) for the given input word vectors W_1, W_2, W_3, W_4,…W_n will be calculated using the weighted sum of all V_n.

$$e_{W1} = w_{11}V_1 + w_{12}V_2 + w_{13}V_3 + w_{14}V_4 + \cdots + w_{1n}V_n$$
$$e_{W2} = w_{21}V_1 + w_{22}V_2 + w_{23}V_3 + w_{24}V_4 + \cdots + w_{2n}V_n$$
$$e_{W3} = w_{31}V_1 + w_{32}V_2 + w_{33}V_3 + w_{34}V_4 + \cdots + w_{3n}V_n$$
$$e_{W4} = w_{41}V_1 + w_{42}V_2 + w_{43}V_3 + w_{44}V_4 + \cdots + w_{4n}V_n$$
$$e_{Wn} = w_{n1}V_1 + w_{n2}V_2 + w_{n3}V_3 + w_{n4}V_4 + \cdots + w_{nn}V_n$$

where,

e_{W1} has the influence from [W_1, W_2,….W_n], defined by w_{1n}

e_{W2} has the influence from [W_1, W_2,….W_n], defined by w_{2n}

e_{W3} has the influence from [W_1, W_2,….W_n], defined by w_{3n}

e_{Wn} has the influence from [W_1, W_2,….W_n], defined by w_{nn}

So, e_{W1}, e_{W2}, e_{W3},…, e_{Wn} are the contextualized representation of words W_1, W_2,….W_n respectively, which no longer points in the same direction that they were initially pointing to. They probably point in different directions influenced by their neighbors, it means whatever score you have, whatever your surroundings are, the neighbors will have strong influence on the contextual representation of each word.

5. BERT FOR MARATHI CONTEXTUAL WORD EMBEDDING

The Transformer encoder is faster than the conventional RNN and LSTM-based neural architectures because it uses bi-LSTM neural architectures and represents into intermediate representation. If we stack these transformer encoders, it will enhance the intermediate representations and can handle the long word sequence as it explores the entire text, which is helpful to improve the contextualized embedding. It enhances the intermediate representations and can handle the long word sequence as it explores the entire text, which is helpful to improve the contextualized embedding. Therefore, it will become suitable to leverage context and gloss distributed representation, which is required for downstream NLP tasks like WSD [26], [29]. This stacked transformer encoder architecture is called BERT. BERT is being claimed to be a breakthrough for NLP and it is being used in various NLP tasks like Neural Machine Translations, Question Answering, Sentiment Analysis, Text Summarization and many more [4].

In this work, we have used BERT for Marathi WSD. To generate the pre-trained word embedding for Marathi, the BERT requires the vocabulary, for which we have pre-trained the BERT encoder on the AI4Bharat IndicBERT Marathi model [30].

5.1 WORKING PHILOSOPHY OF MARATHI WSD FRAMEWORK

The Working philosophy of proposed WSD model is shown in Fig.2 and pseudo-code in sub-section 5.2. It identifies the most ambiguous word W_{amb} from the given input sentence i.e. $C_s=[W_1,W_2,W_3,\ldots W_{\text{amb}}\ldots W_n]$ and C_{amb} be the sentential context for W_{amb}, then we explores IndoWordNet and extract possible synset m glosses for W_{amb}, which are denoted by $G_{amb} = \{G_1,G_2,G_3\ldots..G_m\}$. We have calculated the word vectors for each of the G_i [$i = 1\ldots.m$]. We have calculated word embedding[s] for

C_{amb} and each gloss vector G_i from BERT encoder, $e_{Camb} = \{e_{W1}, e_{W2}, e_{Wn}\}$ and $e_{Gi}=\{e_{Gi1}, e_{Gi2},\dots e_{Gin})$, where $e_{Wn})$ and e_{Gin} are real numbers (ranging from 0 to 1) respectively. We have calculated the various distance measures i.e. Manhattan distance, Euclidean distance, Chebychev distance from Minkowski family and Cosine similarity measure between the word embedding[s] pairs of C_{amb} and G_{amb}, which are $\{[e_{Camb}, e_{G1}], [e_{Camb}, e_{G2}], [e_{Camb}, e_{G3}]\dots [e_{Camb}, e_{Gn}]\}$. Then we have conducted the polling between all the similarity measures and the gloss, whose pair got the maximum similarity poll, is declared as a plausible sense for the ambiguous word given in Marathi sentence.

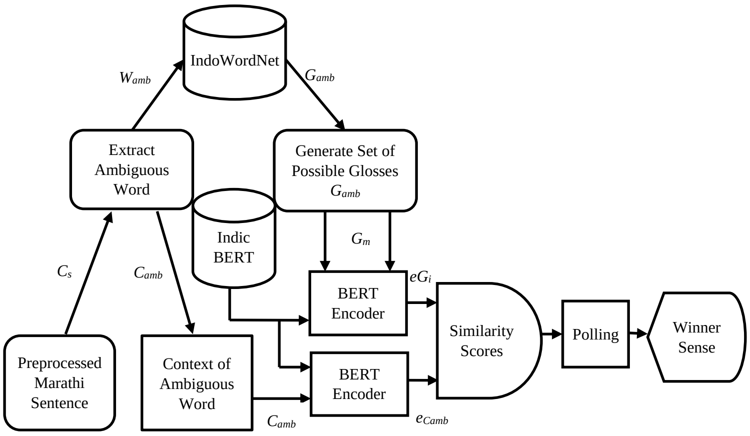

Fig.2. BERT-based Marathi WSDPseudo-code for Proposed Methodology

Requirements: Marathi_Model, Marathi_Vocabulary, IndoWordNet:IWN

Input: Marathi Sentence S_{MAR}

```
1: Start
2: BERT_Model←Pre-BERT_T(Marathi_Mod)
3: $M_{Voc}$←Preproce(Marathi_Vocabulary)
4: Word←GenFreq($M_{Voc}$)
5: Sent ←Preprocess($S_{MAR}$)
6: for each word[i] in Sent:
      MFS[i]=GetSynFromIWN(word[$i$], IWN)
7: AW=Max(MFS[$i$]) where $i$ maximum
8: Gloss = GetSynFromIWN($W_{amb}$, IWN)
9: $V_i$ ← Encoding[Sent, Gloss]
10: $C_i$ ← BERT_Model[Sent]
11: $G_i$ ← BERT_Model[Gloss]
12: $S_i = V_i V_i^T$
13: $w_i$ = Softmax($S_i$)
14: for each $i$ in $V_i$:     //Context Embedding
       Calculate $e_{Camb} = w_i V_i$
15: for each $i$ in $G_i$:    //Gloss Embedding
       Calculate $e_{Gi} = w_i G_i$
16: for each $e_{Gi}$:
       Score← 0
       $Score_{New}$←(1-$d_{norm}$($e_{Camb}$,$e_{Gi}$)) //for Dist.
       $Score_{New}$← $\arg\max_w \text{Cos\_}Sim\left(e_{C_{amb}}, e_{G_i}\right)$
          $= \frac{e_{C_{amb}} \cdot e_{G_i}}{\left|e_{C_{amb}}\right|\left|e_{G_i}\right|}$ //for Cosine sim.
       If $Score_{New}$>$Score$
       Then $Score$←$Score_{New}$
       Return ($Score$[$i$], $Gloss$[$e_{Gi}$])
17: End
```

5.2 WSD USING MINKOWSKI DISTANCES FAMILY

Minkowski distance-based methods calculates the distance-based similarity between context and each gloss pairs embedding[s], as shown in Eq.(3), proper sense of the ambiguous word is declared on the basis of maximum similarity score for the context and either of the gloss.

$$Sim_{Score}(e_{Camb},e_{Gi})=1-d_{norm}[e_{Camb},e_{Gi}] \tag{3}$$

where, d_{norm} is the minimum distance among all the pairs of e_{Camb} and e_{Gi} is calculated in Eq.(4).

$$d_{norm}\left(e_{C_{amb}}, e_{G_i}\right) = \frac{Minkowski_{dist}\left[e_{C_{amb}}, e_{G_i}\right]}{Max\left[Minkowski_{dist}\left[e_{C_{amb}}, e_{G_i}\right]\right]} \tag{4}$$

where, $Minkowski_{dist}\left[e_{C_{amb}}, e_{G_i}\right]$, is the distance between given pair of context and gloss embedding, while $Max\left[Minkowski_{dist}\left[e_{C_{amb}}, e_{G_i}\right]\right]$ is the maximum distance among all the pairs of embedding, which is given by Eq.(5):

$$Minko\_D\_dist\left[e_{C_{amb}}, e_{G_i}\right] = \left(\sum_{i=1}^{n}\left|e_{C_{amb}} - e_{G_i}\right|^{\frac{1}{p}}\right)^{p} \tag{5}$$

where, n is the dimension and of the contextual embedding and gloss embedding and p is the order norm with common values of 'p' are:

p = 1 for Manhattan distance

p = 2 for Euclidean distance

p >2 for Chebychev distance

MFS_{amb} is the proper sense for the distance based similarity is given in Eq.(6):

$$MFS_{amb} = \arg\max_c Similarity_{Score}\left(e_{C_{amb}}, e_{G_i}\right) \tag{6}$$

5.3 WSD USING COSINE SIMILARITY

Cosine-based similarity calculates the similarity between context and each gloss pairs embedding[s] as shown in Eq.(7) and Eq.(8), proper sense of the ambiguous word is declared on the basis of maximum similarity score for the context and either of the gloss.

$$MFS_{amb} = \arg\max_c \text{Cos\_}Sim\left(e_{C_{amb}}, e_{G_i}\right) \tag{7}$$

$$\arg\max_c \text{Cos\_}Sim\left(e_{C_{amb}}, e_{G_i}\right) = \frac{e_{C_{amb}} \cdot e_{G_i}}{\left|e_{C_{amb}}\right|\left|e_{G_i}\right|} \tag{8}$$

where, $\left|e_{C_{amb}}\right|$ is the Euclidean-norm of context embedding which is defined in Eq.(9):

$$\left|e_{C_{amb}}\right| = \sqrt{e_{W_1}^2 + e_{W_3}^2 + \ldots + e_{W_n}^2} \tag{9}$$

$\left|e_{G_i}\right|$ is the Euclidean-norm of gloss embedding which is defined in Eq.(10):

$$\left|e_{G_i}\right| = \sqrt{e_{G_{i1}}^2 + e_{G_{i2}}^2 + \ldots + e_{G_{in}}^2} \tag{10}$$

6. EXPERIMENTAL SETUP, TEST BED AND EVALUATION STRATEGY

For the experimentation purpose, we have used a test bed of randomly picked 5285 Marathi sentences from 282 Marathi moderately ambiguous words harvested from Marathi websites (Heritage, News, Sports, History) catering around 1004 senses. To extract the synsets and glosses for the ambiguous word, we have used CFIL IIT Bombay's IndoWordNet sense inventory, for preprocessing and encoding the sentence and its glosses, we have used iNLTK. We have Pre-trained the BERT encoder on AI4Bharat IndicBERT Marathi Model and used it to generate the contextual embedding(s) for sentential context and glosses. We have used Minkowski family distance measures and Cosine similarity to calculate the similarities between the ambiguous word's context and glosses embedding. To declare the winner sense, we used a polling technique, which considers the maximum number of least distances and maximum similarity values.

In evaluating the performances, we have generated and compared the responses from the WSD approach with the linguist-decided answers in the test-bed and recorded the measures including, TP: number of relevant senses that have been correctly disambiguated, FN: number of relevant senses missed during disambiguation, FP: The number of irrelevant senses wrongly selected as relevant and TN: number of irrelevant senses that were correctly eliminated. Based on the observations and counts of the above parameters, we have estimated the classical measures like exactness (Precision of disambiguation), sensitivity (True sense recognition rate/Precision of disambiguation), specificity (True negative rate/Recall of disambiguation), disambiguation accuracy (Ability of disambiguation) and f_1-score (Harmonic mean of precision and recall of disambiguation).

7. RESULTS AND DISCUSSION

As stated earlier, after doing the test-runs on the testbed, we have calculated and reported sample classical measures like disambiguation accuracy and F_1–Score.

7.1 DISAMBIGUATION ACCURACY (DISAMBIGUATION ABILITY)

It is the percentage of ambiguous words in experimentation that are correctly disambiguated. It shows the overall disambiguation rate and the ability of disambiguation.

7.1.1 PoS-wise Disambiguation Accuracy:

Part-of-speech-wise disambiguation accuracy by Minkowsky and cosine similarity for Marathi WSD is summarized in Table.1. From Table.1, it is observed that Minkowsky distance family and cosine similarity for Marathi WSD has obtained overall disambiguation accuracy of 72.02% across all PoS, among all, Cosine similarity yields highest disambiguation accuracy of 77.84% on nouns whereas, Chebyshev measures has lowest disambiguation accuracy of 52.59 % for verbs in testbed. Due to the depth of noun taxonomy in the present form of the IndoWordNet and 'blessings of dimensionality' the cosine similarity performs better for noun WSD, this is not the case with Minkowsky distance family as well as taxonomy of other PoS.

Table.1. PoS-wise Disambiguation Accuracy in %

PoS	No. of Sense	No of Test Sentences	Manhattan-based WSD	Euclidean-based WSD
Noun	740	4220	75.30	72.42
Adjective	240	956	71.40	69.16
Adverb	12	43	70.00	64.03
Verb	21	66	69.19	59.15
Overall	1004	5285	73.92	70.68
PoS	**Chebyshev-based WSD**		**Cosine-based WSD**	**Average Accuracy**
Noun	70.05		77.84	73.90
Adjective	65.33		72.16	69.51
Adverb	59.15		69.57	65.69
Verb	52.59		68.87	62.45
Overall	67.78		75.69	72.02

It is also observed that Cosine similarity has yielded highest WSD accuracy for all PoS categories except adverb and verb, but overall, Cosine based accuracy score dominates over the Minkowsky distance family also there is not any notable variation in the gain of disambiguation accuracy for all PoS categories.

7.1.2 Sense BW-wise Disambiguation Accuracy:

Sense bandwidth is the number of possible sense of the given ambiguous word, the sense band-width-wise disambiguation accuracy of all similarity is shown in Fig.4.

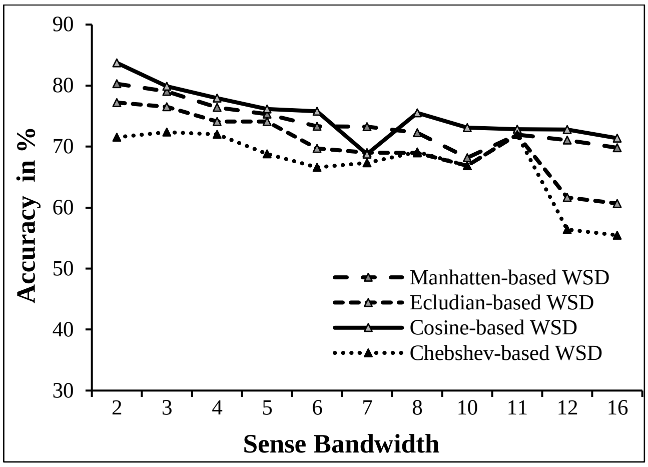

Fig.4. Sense BW-wise Disambiguation Accuracy

From Fig.4, it is observed that, the Minkowsky distance family and Cosine similarity in BERT-based WSD approach has obtained highest disambiguation accuracy of 83.72% by Cosine similarity on the ambiguous word having sense BW-2 whereas there is a lowest disambiguation accuracy of 55.47% by Chebyshev measures on the ambiguous word having sense BW-16. This is due to the blessings of dimensionality for cosine similarity performs better for WSD; this is not the case with Minkowsky distance family. Here the disambiguation accuracy is inversely proportional to the sense bandwidth. It is observed that higher the sense BW the lower the disambiguation accuracy. This phenomenon is not applicable for Cosine similarity from the gain of words of sense bandwidth from 07 to 08; here the disambiguation accuracy is getting higher gain.

From Fig.4, it is also observed that the disambiguation accuracy of the disambiguation process has the highest drop in gain of 21.60% in Chebyshev measures from sense BW-11 to BW-12. The reasons behind all these observations are described in detail in previous measure.

7.2 F_β–SCORE OF DISAMBIGUATION

It is the harmonic mean of precision (exactness) and recall (sensitivity) of sense disambiguation. Here we assume β=1.

7.2.1 PoS-wise F_1 Score of Disambiguation:

PoS-wise F_1–Score of disambiguation i.e. the harmonic mean of exactness and sensitivity of disambiguation by Minkowsky distance family and Cosine similarity for Marathi WSD is summarized in Table.2 and shown in Fig.5. Here F_1–Score represents the percentage of correct disambiguation of a particular PoS.

The disambiguation accuracy described in previous sub-section, is depends not only on the number of relevant sense correctly disambiguated but also the number of irrelevant sense correctly eliminated and in case of higher sense BW words has more number of irrelevant sense correctly eliminated leads higher accuracy of 52.16% of sense disambiguation and in present form of the IndoWordNet the depth of verb taxonomy is limited, so empirically this is not the case of rate of verb disambiguation and hence to decide the disambiguation ability the disambiguation accuracy is a bad metric.

In WSD the distribution of sense classes is unequal, the positive sense in which the system is interested are very rare as compare to negative, so here accuracy measure fails to decide the ability of WSD system, whereas, F_1-Score will takes the average mean of exactness and sensitivity i.e. it does not influenced by the more number of irrelevant sense correctly eliminated in increase sense BW and hence F_1-Score will be the correct measure. In present form of the IndoWordNet the depth of taxonomy has decreases like Noun, adjective, adverbs and then verbs, so their F_1-Score for noun to adjective to adverb and then verb, in this order the F_1-Score decrease, imperially it is true.

Table.2. PoS-wise F_1–Score of Disambiguation in %

POS	Manhattan-based WSD	Euclidean-based WSD	Chebyshev-based WSD	Cosine-based WSD	Average F_1-Score
Noun	48.45	44.50	41.08	52.82	46.71
Adjective	45.06	42.32	35.87	46.18	42.36
Adverb	38.85	29.12	26.39	39.51	33.47
Verb	22.70	12.21	6.87	23.33	16.27
Overall	46.63	42.22	37.95	49.34	44.03

From Table.2, it is observed that, proposed WSD framework has obtained overall F_1-Score of disambiguation of 30.28%, highest F_1-Score of 31.02% on nouns, for adjectives it is 27.07% and for adverbs it is 21.11% whereas there is a lowest F_1-Score of 6.21% for the verb's disambiguation in test-bed.

From Table.2, it is also observed that, Minkowsky distance family and Cosine similarity in BERT-based WSD framework has obtained overall F_1-Score of disambiguation of 44.03%, highest F_1-Score of 52.82% by Cosine similarity on nouns whereas there is a lowest F_1-Score of 6.87% by Chebyshev measures for the verbs disambiguation, it is because of the nature of noun taxonomy in IndoWordNet.

7.2.2 Sense BW-wise F_1-Score of Disambiguation:

Sense BW-wise F_1-Score of disambiguation for Marathi WSD is shown in Fig.6. It shows, Cosine-based similarity yields highest F_1-score of 83.73% for the words have sense BW-2, whereas Chebyshev measures yields lowest F_1-score of 3.39% for the words have sense BW-2. Form the sense BW-wise exactness of sense disambiguation, for all similarity measures, we also observed that, higher the sense BW the lower the F_1-Score, it means in all similarity measures the F_1-Score of disambiguation is reciprocal to the sense band width of an ambiguous word and it is natural phenomena, which is common with human judgments, even the human is not able to discriminate the proper meanings of the words having once the number of sense increased.

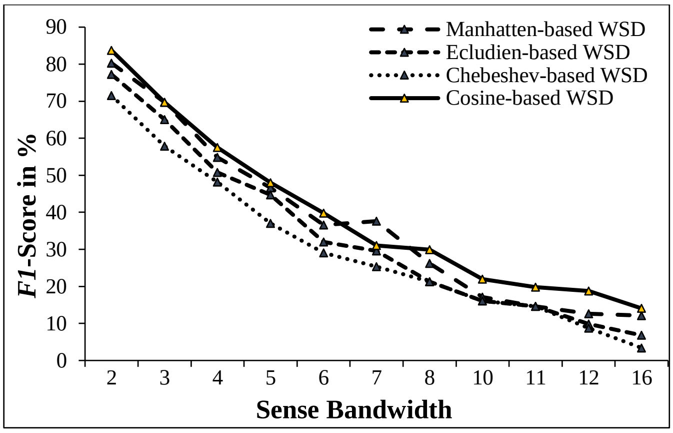

Fig.6. Sense BW-wise F_1 - Score of Disambiguation

8. PHENOMENA STUDY AND ERROR ANALYSIS

The disambiguation accuracy of proposed BERT-based approach is improved around by 8-12% for all PoS categories for Marathi ambiguous words of sense bandwidth up to 16 than the baseline accuracy of 60%, it is because, it is more suitable to represent the morphological richness of the Marathi language.

In computational semantics, the traditional representations treated semantics as an atomic symbol and limited in contextual size, which is sparse in nature, so it is not able to assemble the meaningful comparisons, whereas in the proposed BERT-based WSD approach, we are using distributional hypothesis for contextual representations, in which semantically similar words tend to have similar real-valued representation of magnitude, directions and vice versa, so it captures meaningful syntactic and semantic regularities in a very simple way.

To the best of our knowledge the accuracy obtained on Marathi noun WSD with this approach is better than the state-of-the-art. However, for verb WSD it fails because the present form of IndoWordNet does not have complete semantic information for verb topology, so the approach leads more number of irrelevant senses wrongly selected as a relevant and missed more number of relevant senses during disambiguation.

From the empirical results, it is also observed that, Cosine similarity measure yields higher disambiguation accuracy and F_1–Score for Marathi WSD over Minkowsky distance family measures; this is because of the 'blessings of dimensionality' to Cosine similarity and course of dimensionality to Minkowsky family measures.

The performance of Cosine-based WSD dominates that of distance-based WSD. It is because, as the dimensionality grows, it narrows the distribution of correlation between random embedding.

9. CONCLUSION AND FUTURE WORK

In this work, we have investigated the usefulness of attention-based transfer learning in BERT for contextualized distributed semantic representation and leveraged the meanings in glosses along with context for the task of Marathi WSD. Similarity scores between the context and each gloss embedding are used to declare the plausible sense of the ambiguous Marathi word. For the experimentation and evaluation of the proposed Marathi WSD, we have estimated the classical measures like disambiguation accuracy and F_1-Score for the given testbed. Results indicate that the proposed approach has demonstrated its usefulness for improving the Marathi WSD task effectively. The study also endorses the fact that Cosine similarity attains higher results than Minkowsky distance-based family measures due to the higher dimensionality of the representation. Improving the ontological structure in the IndoWordNet for adverb and verb categories will improve the Marathi WSD. This work can be reused for WSD tasks in contexts of variable size for other Indian languages covered in IndoWordNet.

ACKNOWLEDGEMENT

This research work is financially supported by the Rajiv Gandhi Science and Technology Commission (RGS&TC), Government of Maharashtra, through Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon by Ref. No. KBCNMU/ RGSTC/Sanction Order/84.

REFERENCES

[1] E. Agirre and P. Edmonds, "*Word Sense Disambiguation Algorithms and Applications*", Springer, 2007.

[2] S.S. Patil and B.V. Pawar, "Contrastive Study and Review of Word Sense Disambiguation Techniques", *International Journal on Emerging Technologies*, Vol. 11, No. 4, pp. 96-103, 2020.

[3] S.S. Patil, R.P. Bhavsar and B.V. Pawar, "Path and Information Content based Structural Word Sense Disambiguation", *Proceedings Communications in Computer and Information Science*, Vol. 1483, pp. 341-352, 2022.

[4] J. Devlin and K. Toutanova K, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of International Conference on Machine Learning*, 2019.

[5] B. Bhatt and P. Bhattacharyya, "IndoWordNet and its Linking with Ontology", *Proceddings of International Conference on Natural Language Processing*, pp. 1-14, 2011.

[6] G. Boleda, "Distributional Semantics and Linguistic Theory", *Proceedings of International Conference on Machine Learning*, 2020.

[7] A. Raganato and R. Navigli, "Neural Sequence Learning Models for Word Sense Disambiguation", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 1167-1178, 2017.

[8] Y. Heo, S. Kang and J. Seo, "Hybrid Sense Slassification Method for Large-scale Word Sense Disambiguation", *IEEE Access*, Vol. 8, pp. 27247-27256, 2021.

[9] T. Uslu and W. Hemati, "Fastsense: An Efficient Word Sense Disambiguation Classifier", *Proceedings of 11th International Conference on Language Resources and Evaluation*, pp. 1042-1046, 2022.

[10] M. Kageback and H. Salomonsson, "Word Sense Disambiguation using a Bidirectional LSTM", *Proceedings of International Conference on Cognitive Aspects of the Lexicon*, pp. 51-56, 2016.

[11] L. Vial and D. Schwab, "Improving the Coverage and the Generalization Ability of Neural Word Sense Disambiguation through Hypernymy and Hyponymy Relationships", *Proceedings of International Conference on Machine Learning*, pp. 1-4, 2020.

[12] F. Luo and B. Chang, "Leveraging Gloss Knowledge in NeuralWord Sense Disambiguation by Hierarchical Co-Attention", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 1402-1411, 2019.

[13] F. Luo and Z. Sui, "Incorporating Glosses into NeuralWord Sense Disambiguation", *Proceedings of International Conference on Computational Linguistics*, pp. 2473-2282, 2018.

[14] V. Hofmann and H. Schutze, "Dynamic Contextualized Word Embeddings", *Proceedings of International Conference on NLP*, pp. 6970-6984, 2021.

[15] M.E. Peters and M. Gardner, "Deep Contextualized Word Representations", *Proceedings of Conference of the North American Chapter*, pp. 2227-2237, 2018.

[16] S. Kumar and P. Talukdar, "Zero-Shot Word Sense Disambiguation using Sense Definition Embeddings", *Proceedings of International Conference on Computational Linguistics*, pp. 5670-568, 2019.

[17] J. Du and M. Sun, “Using BERT for Word Sense Disambiguation”, *Proceedings of International Conference on Natural Language Processing*, pp. 1-7, 2019.

[18] L. Huang and X. Huang, “GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge”, *Proceedings of International Conference on Computational Linguistics*, pp. 3509-3519, 2018.

[19] C. Hadiwinoto, “Improved Word Sense Disambiguation using Pre-Trained Contextualized Word Representations”, *Proceedings of International Conference on Computational Linguistics*, pp. 5297-5306, 2019.

[20] L. Vial and D. Schwab, “Sense Vocabulary Compression through the Semantic Knowledge of WordNet for NeuralWord Sense Disambiguation”, *Proceedings of International Conference on Global Wordnet*, pp. 108-117, 2019.

[21] M. Stoeckel and A. Mehler, “SenseFitting: Sense Level Semantic Specialization of Word Embeddings for Word Sense Disambiguation”, *Proceedings of International Conference on Computational Linguistics*, pp. 365-366, 2019.

[22] M. Bevilacqua and R. Navigli, “Breaking Through the 80% Glass Ceiling: Raising the State of the Art inWord Sense Disambiguation by Incorporating Knowledge Graph Information”, *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pp. 2854-2864, 2020.

[23] S. Bhingardive and P. Bhattacharyya, “Using Word Embeddings for Bilingual Unsupervised WSD”, *Proceedings of International Conference on Natural Language Processing*, pp. 59-64, 2015.

[24] K. Saurav, D. Kanojia and P. Bhattacharyya, “A Passage to India: Pre-Trained Word Embeddings for Indian Languages”, *Proceedings of International Conference on Language Resources and Evaluation*, pp. 352-357, 2020.

[25] I. Guyon, G. Taylor and D. Silver, “*Unsupervised and Transfer Learning Challenges in Machine Learning*”, Microtome Publishing, 2013.

[26] A. Vaswani, L. Jones and A. Gomez, “Attention Is All You Need”, *Proceedings of International Conference on Neural Information Processing Systems*, pp. 6000-6010, 2017.

[27] S. Khan and F.S. Khan, “Transformers in Vision: A Survey”, *Proceedings of International Conference on Computational Linguistics*, pp. 1-28, 2021.

[28] S. Pan and Q. Yang, “A Survey on Transfer Learning”, *IEEE Transactions On Knowledge and Data Engineering*, Vol. 22, No. 11, pp. 1345-1360, 2020.

[29] I. Sutskever and V. Le, “Sequence to Sequence Learning with Neural Networks”, *Proceedings of International Conference on Neural Information Processing Systems*, pp. 1-9, 2014.

[30] V. Raghavan, M. Khapra and A. Kunchukuttan, “AI for Bharat”, Available at https://ai4bharat.org/indic-bert, Accessed on 2022.