ORIGINAL ARTICLE

SAMRIDDHI Volume 14, Issue 4, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

ULMFiT Embedding(s) for Context and Extended Gloss Intersection for Marathi Word Sense Disambiguation

Sandip S. Patil*, R. P. Bhavsar, B. V. Pawar

School of Computer Sciences, K.B.C. North Maharashtra University Jalgaon, M.S. India.

Abstract

Ambiguities in the word meanings makes all the natural language processing (NLP) tasks very difficult, word sense disambiguation (WSD) is used to resolve these ambiguities. Now a day's NLP-based human assistive systems are in demand, in which machines are expected to resolve word sense ambiguities. Today, due to the availability of machine readable dictionaries knowledge-based WSD approaches have become popular; it explores semantic relations between the contextual features and possible glosses of the given ambiguous word. Inductive transfer learning-based language models have great potential to represent the different semantic features of the word, which can be used in various NLP tasks. Universal language model fine-tuning for text classification (ULMFiT) is a popular transfer learning model used to embed various semantic features in digitally resource scare and morphologically rich language like marathi. In this reported work, the ambiguous words from the Marathi input sentence is extracted and have obtained its possible synset and glosses from IndoWordNet, these glosses are then extended using hypernym and hyponym relations. We have obtained the word embedding of marathi context and extended glosses using ULMFiT model. For the test run, we have crafted the test-bed of 6000 marathi sentences of 280 moderately ambiguous words harvested from marathi websites, which caters for 1200 senses. The winner sense is declared based on the maximum intersection score between the pair of context and gloss embedding. We have obtained the average accuracy up to 57.10% for our dataset.

Keywords: Marathi Word Sense Disambiguation, Lexical Relations, Neural Langauge Modeling, ULMFiT Model, Word Embedding.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i04.20

Introduction

Ambiguities in the word meanings is a classical problem in natural language processing (NLP), it occurs due to the presence of ambiguous words in the sentence and it makes all the NLP tasks very difficult. Humans use their world knowledge, cultural background, and previous experience to resolve such ambiguities. Due to advent of NLP-based human assistive systems and there growing demand, the machines are expected to tackle this issue, but due lack of human-like intelligence it is not able to handle this issue, hence it causes a big challenge as well as research opportunities in NLP. Word sense disambiguation (WSD) task deals with this challenge. WSD has many applications in various NLP applications like sentiment analysis, machine translation, information retrieval, question answering, text analysis, text entailment, semantic role labeling etc. rely on WSD.[1] Knowledge-based WSD approaches are simple and have greater coverage and scalability, which are knowledge lean. Due to the availability of machine readable dictionaries these WSD approaches became popular. Knowledge-based overlap WSD uses intersection between the sentential features of context and gloss. The sense which having maximum semantic related-based intersection is selected as the contextually appropriate sense. Marathi is a digitally resource scarce and morphologically rich Indian language, due to its resource scarce nature, to model it semantically using the existing RNN and LSTM-based language models, it imposes constraints as these models needs huge amount of data, so RNN and LSTM-based models overfit to NLP dataset and faces catastrophic forgetting. ULMFiT is the robust transfer learning model, which works on any number of documents of any size(s) and

Corresponding Author: Sandip S. Patil, School of Computer Sciences, K.B.C. North Maharashtra University Jalgaon, M.S. India, e-mail: sspatiljalgaon@gmail.com

How to cite this article: Patil, S. S., Bhavsar, R. P., Pawar, B. V. (2022). ULMFiT Embedding(s) for Context and Extended Gloss Intersection for Marathi Word Sense Disambiguation. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology,* 14(4), 126-132.

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

Conflict of interest: None

does not requires additional domain knowledge (labels), so it addresses these issues faced in RNN and LSTM.[2] In this reported work, we have studied ULMFiT model and used it to represent the various word features of Marathi deeply. The technique presented is an application of transfer learning in conjunction with IndoWordNet that defines word synsets along with extended gloss. In the proposed approach, various lexical semantics and the contextual intersections among the pair of context and extended gloss embedding(s) are measured for the task of marathi WSD. The following sections describes prior art followed by the detailed discussion on ULMFiT model, the brief description of marathi context and extended gloss embedding, semantic relatedness-based intersection and then used it for the task of marathi WSD.

Prior Art

Depending on the granularity of corpora, WSD approaches are classified into two categories knowledge-based and machine learning-based.[3,4] Most of the well-known WSD approached are studied in Patil et al. 2020.[4]

During our study, it is observed that knowledge-based WSD has been investigated over the period of three decades. A knowledge-based WSD approach uses distributional similarity, thesaurus-based similarity and overlapped approach.[1,3,4] Distributional similarity uses distributional context, while thesaurus-based similarity uses relations in lexical semantics. Overlapped-based WSD approach uses the maximum overlap between the context and gloss of an ambiguous word, accordingly it selects the winner senses.[1] M. Lesk [5] proposed 'Lesk algorithm' for gloss overlapped-based WSD for English. Lesk algorithm looks the overlap between the words in the dictionary definitions with the text surroundings. The only resource required by the algorithm is a set of dictionary entries, one for each possible word sense, and knowledge about the immediate context where the sense disambiguation is performed. Lesk is a surface level work, it treats only the overlap among the context and sense bag, where sense glosses are fairly short and are not able to provide the sufficient distinctions between relatedness, it is susceptive to the exact gloss of the synset and hence presence or absence of certain word empirically change the result. Banerjee and Pedersen [2003][6] extended Lesk approach; they have used the relationship in the machine readable dictionary WordNet and extended features of the gloss overlapped for english WSD, they measured the relatedness of two word concepts from the english WordNet and proved that synset relations greatly improve the performance of lesk for english WSD by 19%. They have used overlap-based detection and scoring mechanism, it may generates tie among the score of multiple scenes, in this case it reports all tie score sense as a winner sense and it also assumes that target word must be at the center always, both are not the case of empirical evidence. Patil et al. [2021],[1] used path-based semantic and information content similarity measures for English WSD, Marathi sense repositories like WordNet and IndoWodNet are not suitable for structural similarity measures. In the literature, it is observed that most of the efforts have been done for the disambiguation of international languages, the efforts in this regards for Indian language like marathi are in prior stage.

The neural-based language models are used to represent the syntax and semantic information in the word sequence of natural language; they explore plenty of features from the granularity of corpora, which are discussed in the next section.

Representation of Word Sequence using Neural Language Models

The representation of syntax and semantic information in the word sequence using language models has attracted inductive transfer learning in various NLP tasks. Inductive transfer learning language models uses various neural architecture, which are pre-trained continuous language models.[7] These models generate intermediate valued representations called 'word embedding'. word embedding is the real-valued representation of the word, which allows the words to have similar representation for similar meanings and different representation for different meanings with respect to the context.[8] Word2Vec[9], GloVe[10] and FastText[11] provides a single and independent contextual representation for the same word in different contexts and it becomes static in nature. In the representation of polysemy, it is necessary to accommodate the complete sentential context, so to handle the issues of missing context information, Peters et al. (2018)[12] proposed bi-LSTM deep contextualized word representation called as ELMo[12-14] leads more steps of computations and applicable to only the fixed dimension encoding and embedding. LSTM and its parent RNN requires millions of documents for pre-training,[10,15] but NLP systems are limited in dataset, so it faces the problem of catastrophic forgetting, so limits the applicability in NLP also existing transfer learning methods for NLP required task-specific modifications and in NLP, as we know the existing domain and the domain of interest are different to solve this challenge. Ruder and Howard (2018)[2] improved transfer learning and developed ULMFiT, robust generative pre-training learning model, it does not requires additional in domain documents and labels also works on varying documents number and size, hence it addresses the issues of overfit and catastrophic forgetting in RNN and LSTM-based models.[2] In this study, ULMFiT is used to encode and represent the hypernym and hyponym-based extended gloss and the context for the task of marathi WSD.

Motivation

Marathi has more word sense ambiguity, which makes all the marathi NLP tasks very difficult.[4,16] Now a days, due to the availability of sense repositories, overlapped based approaches becomes popular, but existing overlapped-bases WSD approaches[5] and[6] limited in the clue and scores for

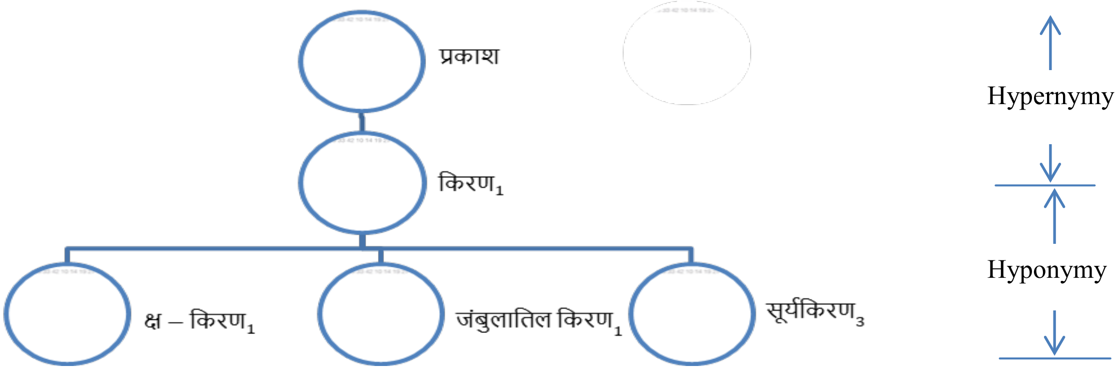

Figure 1: Hypernym and Hyponymy Relation for the sense किरण$_1$(Kirana).

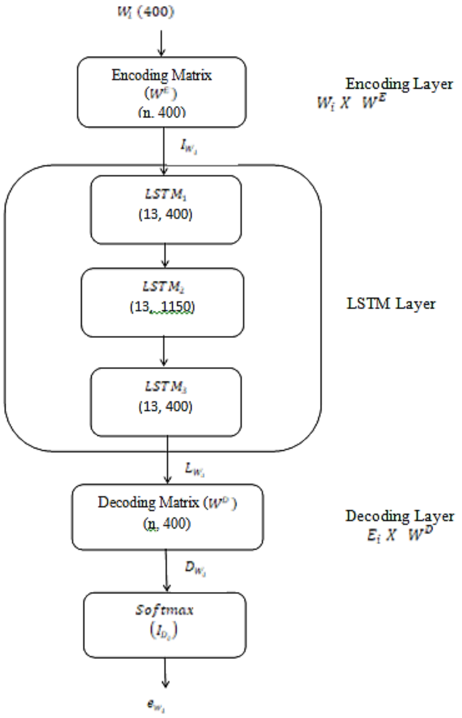

Figure 2: Transfer Learning in ULMFiT

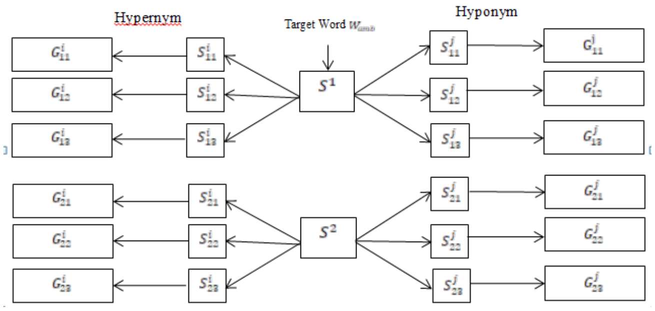

Figure 3: Extension of Gloss using Hypernyms and Hyponyms Relations.

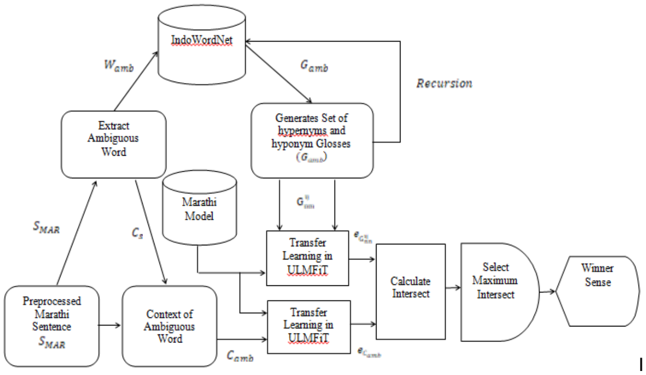
IndoWordNet
W_{amb}
G_{amb}
Recursion
Extract Ambiguous Word
Generates Set of hypernyms and hyponym Glosses (G_{amb})
Marathi Model
S_{MAR}
C_s
G^n_{amb}
Transfer Learning in ULMFiT
Transfer Learning in ULMFiT
Calculate Intersect
Select Maximum Intersect
Winner Sense
Preprocessed Marathi Sentence S_{MAR}
Context of Ambiguous Word
C_{amb}
e_{amb}

Figure 4: Architecture of Extended Gloss Intersection Measures for Marathi WSD.

the short context and gloss, consequently the possibility of generating the same score for more glosses is increased and it limits the overall performance of the disambiguation task. A possible solution to this problem is to incorporate the lexical relations like hypernym and hyponym and extend the gloss and then represent the gloss and context using neural-based ULMFiT contextual embedding, so that it will improve the coverage and generalization ability of the context and gloss intersection and get the variety of the clues and matching scores for the possible sense of the given ambiguous words.

Extraction of Gloss and Extended Gloss from IndoWodnet

The proposed WSD system accepts the marathi sentence, preprocess it and extract the most ambiguous word from the given input sentence. For this ambiguous word, it then explore the IndoWordnet and extract it's all possible synsets and recursively for each synset it extracts hyponymy relations like hypernym and hyponym synsets and their glosses, these hypernym and hyponym glosses for the particular synset is called as 'extended gloss'. In marathi sentence l ;çdk'k, vfXup e[; L=kr vkg, ;keG l ;kp fdj.k 'kjhjkyk dkjM djrkr, T;keG 'kjhjkr 'khryrk vkf.k fiÙk nj gkrs (Sūryaprakāśa, agnicē mukhya strōta āhē, yāmulē sūryācē kirana śarīrālā kōradē karatāta, jyāmulē śarīrāta śītalatā āni pitta dūra hōtē/sunlight is the main source of fire, so the sun's rays dry the body, which removes coldness and bile from the body), for the ambiguous word fdj.k (Kirana), IndoWodnet has two senses, çdk'kfdj.k (Prakāśakirana/ray of light) and ipxxrhy ikp çe[k u|kidh ,d (Pancagangētīla pāca pramukha

Table 1: Accuracy, Precision, Recall and F_1 Score of Intersection-based WSD framework

PoS	*Total Sent.*	*Total Synset*	*Accuracy in %*	*Precision in %*	*Recall in %*	*F_1 Score in %*
Noun	4560	875	57.3	63.0	62.4	62.7
Adj	1160	235	55.2	61.3	60.0	60.6
Adv	160	42	51.2	56.0	58.3	57.1
Verb	120	48	50.0	53.1	48.6	50.7
All	6000	1200	57.1	62.0	62.5	62.3

nadyāmpaikī ēka/ One major river in the Panchganga)For the synset çdk'kf dj.k (Prakāśakirana/ray of light) the hypernym is çdk'k (Prakāśa/light) i.e. f dj.k (Kirana), is a kind of çdk'k] mtM] vkyksd] nhfIr] rst] çHkk] vkHkk] |qrh (Prakāśa, ujēda, ālōka, dīpti, tēja, prabhā, ābhā, dyutī/light) and hyponyms are {k&f dj.k] tcqykrhr f dj.k] सू;f dj.k (Ksa-kirana, jambulātīta kirana, sūryakirana/x-ray) i.e. {k&f dj.k (Ksa-kirana/X-rays) is a kind of f dj.k (Kirana) and for the synset iapxaxrhy ikp çeq[k u|kaiSdh ,d (Pancagangētīla pāca pramukha nadyāmpaikī ēka/ One major river in the Panchganga) the hypernym is unh (Nadī/river) and hyponyms is tyjk'kh (Jalarāśī/water stream). Figure 1 shows the Hypernymy and Hyponymy for the Marathi word Kiran.

ULMFiT for Obtaining the Embedding(s) of Marathi Sentence

Inductive transfer learning models does not need to train from scratch, so they overcomes the issues of labeled data (18), in inductive transfer learning, the model is pre-trained on source task which is different than the target task.[19] ULMFiT is a inductive transfer learning-based language model.[2] In this work, to model the marathi langauge, ULMFiT is pre-trained on iNLTK marathi model.

Transfer learning in ULMFiT for Marathi Sentence

For modeling the marathi language, ULMFiT is pre-trained on iNLTKs' vocabulary of Marathi wikipedia articles. Figure 2 shows the detailed pre-training of ULMFiT model for marathi. To obtain the contextualized embedding of the given marathi sentence S_{MAR}, first it has to be tokenized and encoded with an integer and then prepares the vector of each token by one-hot vector method. Let $C_s = (W_1, \; W_2, \; W_3, \ldots\ldots W_{amb}, \ldots. W_n)$ be the sequence of n number of one hot vector, which get feed into the ULMFiT model. ULMFiT model will then prepare the sentential context C_{amb} of the 13 word vectors of the ambiguous word (6 from left and 6 from right) W_{amb}, where $C_{amb} = (W_1, \; W_2, \; W_3, \ldots\ldots W_{amb}, \ldots. W_{13})$. Pre-defined learnable encoding matrix W^E, where $W^E \in R^{n \; X \; 400}$, in which n is the number of unique vocabulary tokens in the iNLTK marathi model and embedding size equals 400, which is the 400d vector for each token. One hot encoding match each encoded token of the input sentence C_s gives a tensor C_{amb} , where $C_{amb} = (W_1, \; W_2, \; W_3, \ldots\ldots W_{amb}, \ldots. W_{13})$ of the 13 vectors, where each $W \in R^{1 \, X \, n}$, here 13 is the context window for the input sentence, if the context size of the input sentence is less than 13, then padding will get applied and if it is more than 13, then ULMFiT will considers 13 tokens only, in this way the model will understand each word of the input sentence. To obtain initial embedding(s) $I_E = (I_{W_1,} \; I_{W_2,} \; I_{W_3,} \ldots\ldots\ldots. I_{W_{amb}}, \ldots\ldots. I_{W_{13},})$ for each encoded token of the sentence S_{MAR}, Each one hot encoded tensor in $C_{amb} = (W_1, \; W_2, \; W_3, \ldots\ldots W_{amb}, \ldots. W_{13})$ is multiplied with the learnable encoding matrix W^E by $W_i \; X \;\; W^E = I_{W_i,}$, where i is the index of word W and $I_{W_i} \;\in\; R^{1 \, X \, 400}$, which generates 13 initial embedding(s), these initial embedding(s) I_E are then passed through the stack of three long short term memory (LSTM) architectures. $LSTM_1$, $LSTM_2$ and $LSTM_3$ which are cascaded to each other and are used to remember the sentential contextual information for long dependency and period (20), where $LSTM_1 \in R^{13 \times 1 \times 400}$, $LSTM_2 \in R^{13 \times 1 \times 1150}$ and $LSTM_3 \in R^{13 \times 1 \times 400}$, where 1150 are the number of hidden states in $LSTM_2$. The output sequence of the $LSTM_3$ is a encoded sequence called as L_{W_i} , where $L_{W_i} \in R^{1 \, X \, n}$ then L_{W_i} is multiplied with learnable pre-defined decoder matrix W^D by $L_{W_i} \; X \;\; W^D = D_{W_i}$, where $W^D \in R^{n \; X \; 400}$ and $D_{W_i} \in R^{1 \, X \, 400}$. The values in D_{W_i} are in different range, so $Softmax(D_{W_i})$ is used to normalized it into the range between the contextual probabilities [0 to 1] and this probability distribution is the contextual word embedding e_{w_i} for the given word W_i.

$$e_{w_i} = Softmax(D_{W_i})_{400} = \frac{e^{DW_i}}{\sum_{i=1}^{400} e^{DW_i}} \qquad 1$$

in this way every probable value in e_{w_i} indicates the probabilities of the next contextual token in the sentence based on the all previous contextual features, this value is called as contextual embedding(s) $e_{C_{amb}} = (e_{w_1}, \; e_{w_2}, \; e_{w_3} \ldots\ldots\ldots \; e_{w_{amb}}, \ldots.. e_{w_{13}})$ of each word sequence $C_{amb} = (W_1, \; W_2, \; W_3, \ldots\ldots W_{amb}, \ldots. W_{13})$ respectively by the ULMFiT model.

ULMFiT for Marathi Context and Extended Glosses Embedding

As shown in Figure 3, the glosses of senses S^1 and S^2 will get extended using various *hypernymy and hyponym relations* and then pre-trained ULMFiT model is used to generate the embedding(s) of extended glosses and context of the ambiguous word.

As shown in figure 3, Let W_{amb} be the target ambiguous word extracted from the sentential sequence of Marathi context, $C_{amb} = (W_1,\ W_2,\ W_3, \ldots\ldots W_{amb}, \ldots\ldots W_{13})$ and IndoWordNet is used to explore and extract all the possible synsets of W_{amb}, which are $S^t_{ij} = (S^1_{11}, S^2_{22}, S^3_{33}, S^4_{44} \ldots\ldots\ldots, S^t_{ij})$ and glosses $G^t_{ij} = (G^1_{11}, G^2_{22}, G^3_{33}, G^4_{44} \ldots\ldots\ldots, G^t_{ij})$ where i and j are the hypernym and hyponym of the synset S^t_{ij}. Using lexical semantics, we have extracted hypernym(i) and hyponym(j) of synsets S^t_{ij}, where $S^i_{nm} = (S^i_{n1}, S^i_{n2}, S^i_{n3}, S^i_{n4} \ldots\ldots\ldots, S^i_{nm})$ is hypernym synsets, in which n is synset and m=0 to n (glosses), and $G^i_{mn} = (G^i_{n1}, G^i_{n2}, G^i_{n3}, G^i_{n4} \ldots\ldots\ldots, G^i_{nn})$ is the set of hypernym glosses and $S^j_{nm} = (S^j_{n1}, S^j_{n2}, S^j_{n3}, S^j_{n4} \ldots\ldots\ldots, S^j_{nn})$ is the hyponym synsets and $G^j_{nm} = (G^j_{n1}, G^j_{n2}, G^j_{n3}, G^j_{n4} \ldots\ldots\ldots, G^j_{nn})$ is the set of hyponym glosses.

In context and extended gloss contextual representation, the AWD-LSTM (weighted drop long term short term memory) reads the context sequence (C_{amb}), gloss sequences and generates the contextual embedding(s) $e_{C_{amb}}$, $e_{G^i_{nm}}$ and $e_{G^j_{nm}}$ of context and extended hypernyms gloss sequences G^i_{nm} and hyponyms gloss sequence G^j_{nm} respectively.

$$e_{C_{amb}} = (e_{w_1}, e_{w_2}, e_{w_3} \ldots\ldots\ldots e_{w_{amb}}, \ldots\ldots e_{w_{13}})$$

$$e_{G^i_{nm}} = (e_{G^i_{n1}}, e_{G^i_{n2}}, e_{G^i_{n3}} \ldots\ldots\ldots\ldots\ldots, \ldots\ldots e_{G^i_{n13}})$$

$$e_{G^j_{nm}} = (e_{G^j_{n1}}, e_{G^j_{n2}}, e_{G^j_{n3}} \ldots\ldots\ldots\ldots\ldots, \ldots\ldots e_{G^j_{n13}})$$, respectively.

Working Philosophy of Intersection-based Marathi WSD

Intersection between the embedding is the embedding that has contextual probabilities in common. The Intersection_ Score (O_t) calculates the probability score of each related sense (S^t) corresponds to the target word (C_{amb}). It calculates the intersection between the pair of context embedding $e_{C_{amb}}$ and extended gloss embedding $e_{G^{ij}_{nm}}$, which is the concatenation of hypernyms $e_{G^i_{nm}}$ and hyponyms $e_{G^j_{nm}}$ gloss representation of the particular sense in $S^t_{ij} = (S^1_{11}, S^2_{22}, S^3_{33}, S^4_{44} \ldots\ldots\ldots, S^t_{ij})$.

$O_t = f(e_{C_{amb}}, e_{G^{ij}_{nm}})$ is the intersection measure between the context and extended gloss

So, $$O_t = P(e_{C_{amb}} \cap e_{G^{ij}_{nm}}) \qquad 2$$

$$= P(e_{C_{amb}}) \times P(e_{G^{ij}_{nm}})$$

$$= e_{C_{amb}} \times e_{G^{ij}_{nm}}$$

In this way, the cumulative intersection score (O_t) will calculate the probability intersection of the pairs of the context and extended gloss embedding(s) like $P(e_{C_{amb}} \cap e_{G^{11}_{11}})$, $P(e_{C_{amb}} \cap e_{G^{22}_{22}})$, $P(e_{C_{amb}} \cap e_{G^{33}_{33}}) \ldots\ldots\ldots\ldots P(e_{C_{amb}} \cap e_{G^{nm}_{nm}})$, which is then used to assign the similarity score S_{MAX} to each sense in S^t, where $S_{MAX} = \max_{O_t}(O_1, O_2, O_3 \ldots\ldots\ldots\ldots O_n)$, which is the empirical sense identified by the Marathi WSD framework. Figure 4 shows the architecture of ULMFiT-based Marathi WSD framework.

Illustration of Idea

Consider the Marathi Sentence l ;çdk'k] vfXup e[; L=kr vkg] ;keGs l ;kp fdj.k 'kjhjkyk dkjM djrkr] T;keGs 'kjhjkr 'khryrk vkf.k fiÙk n j gkr (Sūryaprakāśa, agnicē mukhya strōta āhē, yāmulē sūryācē kirana śarīrālā kōradē karatāta, jyāmulē śarīrāta śītalatā āni pitta dūra hōtē/ Sunlight is the main source of fire, so the sun's rays dry out the body, which removes coldness and bile from the body)

In this sentence, the word fdj.k (Kirana/ray) is ambiguous and for this word, IndoWordnet has two senses which are;

$Synset_1$('fdj.k'): $Gloss_1$: ipxxrhy ikp çe[k u|kidh ,d (Pancagangētīla pāca pramukha nadyāmpaikī ēka) and $Synset_2$ ('çdk'kfdj.k): $Gloss_2$: l ;] pæ] fnok bR;knh rtLoh inkFkkikl u fu?kkyyh çdk'k'kykdk (Sūrya, candra, divā ityādī tējasvī padārthāpāsūna nighālēlī prakāśaśalākā)

For $Synset_1$, the IndoWordnet has 1 Hypernymy and 1 Hyponymy which are:

$Hypernym_1$: 'koVh leækyk tkÅu feG.kkjk ik.;kpk çokg (Śēvatī samudrālā jā'ūna milanārā

For $Synset_2$, the IndoWordnet has 2 Hypernymy and 6 Hyponymy which are:

$Hypernym_1$: çdk'k T;keGs fnl.k 'kD; gkr rs rÙo (Prakāśa jyāmulē disanē śakya hōtē tē tattva)

$Hypernym_2$: vkReckèk] ckèk] vkRekuHk rh] oLr] fo"k; bR;knhP;k Lo:ikph eukyk gk.kkjh tk.kho (Ātmabōdha, bōdha, ātmānubhūtī, vastu, visaya ityādīncyā svarūpācī manālā hōnārī jānīva)

$Hyponym_1$: {k&fdj.k ,[kk|k dBh.k oLr oj oxku ,yDV,upk ekjk d:u mRiUu gk.kkjk] deh rjx ykchpk] fo|rpcdh; fdj.k (Ksa-kirana ēkhādyā kathīna vastūvara vēgānē ēlēktŏnacā mārā karūna utpanna hōnārā, kamī taranga lāmbīcā, vidyutacumbakīya kirana)

$Hyponym_2$: tcykrhr fdj.k {kfdj.kki{kk tkLr rjxykchph fdj.k (Jambulātīta kirana ksa kiranāmpēksā jāsta tarangalāmbīcī kirana)

$Hyponym_3$: vojä fdj.k ,d çdkjp fdj.k (Avarakta kirana ēka prakāracē kirana)

$Hyponym_4$: स्;kp fdj.k] igkVph स्;fdj.k vx.kkr il: ykxyh (Sūryācē kirana, pahātēcī sūryakiranē anganāta pasarū lāgalī.)

$Hyponym_5$: pæfdj.k pækph fdj.k (Candrakirana candrācī kiranē)

$Hyponym_6$: xek fdj.k] dkgh inkFk bR;knhr u fu?k.kkjk ,d çdkjpk fdj.k (Gĕmā kirana, kāhī padārtha ityādīntūna nighanārā ēka prakāracā kirana)

Used ULMFiT model to generate the context $e_{C_{amb}}$ and extended gloss embedding $e_{G^{ij}_{nm}}$ which is the assembly of Hypernym $e_{G^i_{nm}}$ and Hyponym $e_{G^j_{nm}}$ of the particular sense.

$e_{C_{amb}}$ = ULMFiT_embedding (*Context*)

$e_{G^{11}_{nm}}$ =ULMFiT_embedding (*Extd_Gloss*$_1$)

$e_{G^{22}_{nm}}$ =ULMFiT_embedding (*Extd_Gloss*$_2$)

Measuring the intersections O_1 and O_1 between the pairs of $(e_{C_{amb}}, e_{G^{11}_{nm}})$ and $(e_{C_{amb}}, e_{G^{22}_{nm}})$, which is calculated as…

$O_1 = P(e_{C_{amb}} \cap e_{G^{11}_{nm}})$, where P is the probability of intersection.

$$= P\left(e_{C_{amb}}\right) \times \mathrm{P}\left(e_{\mathrm{G}_{\mathrm{nm}}^{11}}\right)$$

$$= e_{C_{amb}} \times e_{\mathrm{G}_{\mathrm{nm}}^{11}}$$

$$= 0.04005$$

$$O_2 = P\left(e_{C_{amb}} \cap e_{\mathrm{G}_{\mathrm{nm}}^{22}}\right)$$

$$= P\left(e_{C_{amb}}\right) \times \mathrm{P}\left(e_{\mathrm{G}_{\mathrm{nm}}^{22}}\right)$$

$$= e_{C_{amb}} \times e_{\mathrm{G}_{\mathrm{nm}}^{22}}$$

$$= 0.22414$$

Here $O_2 > O_1$, hence for the Marathi Sentence, ambiguous word fdj.k (Kirana) has the proper sense in the $Synset_2$ ('çdk'kfdj.k (Prakāśakirana).

Evaluation Strategy, Experimental Setup and Test Bed

For the experimentation, we have used a test-bed of randomly picked 6000 marathi sentences from 280 marathi moderately ambiguous words, harvested from marathi websites (heritage, news, sports, history) catering to around 1200 senses. To extract the synsets and their extended hypernym and hyponym glosses for the targeted ambiguous word, we have used IndoWordNet sense inventory (17), for pre-processing and encoding the input marathi sentence and its extended glosses, we have used iNLTK tools. We have pre-trained the ULMFiT model on iNLTK's marathi model and used it to generate the contextual embedding(s) for sentential context and extended glosses. We have calculated the probability of the intersection between the pairs of context and extended gloss and based on the maximum probability the winner sense for the given ambiguous word is declared. We have calculated the accuracy, precision, recall and F_1 score of proposed extended gloss intersection-based WSD on the given test-bed over different PoS categories in input sentence.

Results and Discussion

After doing the test run on 6000 open texts Marathi sentences, we have calculated accuracy, precision, recall and F_1 score matrices for nouns, adjectives, adverbs and verbs PoS categories as shown in Table 1. For our case, relevant senses are the sense identified by the marathi linguistics and retrieved sense are retrieved by the WSD framework, in other words the relevant senses are true positive and the retrieved sense are both true positive and false positive.

Precision represents the fraction of sense identified by the WSD framework are correct, it is the ratio of how many predicted senses are relevant with that of all. Recall represents the fraction of total proper sense correctly declared by the WSD framework; it is the ratio of how many relevant senses are correctly predicted. F_1 score is the weighted average of the ratio of how many predicted senses are relevant with that of how many relevant senses are correctly predicted and the accuracy represents the fraction of total relevant sense declared correctly relevant, non-relevant sense declared correctly non-relevant with that of all.

From Table 1, it is observed that, proposed WSD framework has obtained highest accuracy of 57.30% on nouns PoS category, whereas lowest accuracy of 50.00% on verb PoS category. In disambiguating the verb PoS, proposed WSD framework performs comparatively lower; it is because in present form of the IndoWordNet the depth of verb taxonomy is limited for hypernym and hyponym glosses, so contextual word embedding for the verbs fails in finding contextual words in extended gloss, so not able to capture the correct semantic information for the given target verb and leads to poor performance in all the measures. Due to unavailability of similar unsupervised computational approach for marathi open text WSD, we cannot compare the present work with other research works.

Phenomena Study and Error Analysis

For contextual representation, the proposed approach is using multivariate probabilistic distribution over the sequence of words, in which irrespective of context, the semantically similar words tend to have similar real-valued probabilistic representation. In traditional knowledge-based WSD the words were treated as atomic symbols, in which intersection were measured on the basis of lexical match, so there, we are not able to make the comparisons at semantic level, this is not the case in proposed WSD framework. The proposed WSD framework also attended all the sentential instances of a word in the context and extended gloss.

For example consider the sentence: vud fnolkP;k vH;klkus d,EiqVj dlk pkyokok gh dyk jktwyk voxr >kyh (Anēka divasāncyā abhyāsānē kŏmputara kasā cālavāvā hī kalā rājūlā avagata jhālī/ After several days of study, Raju learned how to operate a computer), here the word dyk (Kalā) is most ambiguous, in IndoWordnet this word has six senses which are:

Gloss 1: ,[kknh xks"V iqu%iqUgk djr jkfgY;kus rh dj.;klanHkkZr ;s.kkjh lgtrk (EKhādī gōsta punahpunhā karata rahālayānē tī karanyāsandarabhatā yēnārī sahajatā)

Gloss 2: lkSn;kZpk vuqHko ns.kkjh dyk(Saundaryācā anubhava dēnārī kalā)

Gloss 3: pæfcackpk lkGkok Hkkx (Candrabimbācā sōlāvā bhāga/ The sixteenth part of the crescent)

Gloss 4: Kku] vuqHko] f'k{k.k bR;knhP;k –"Vhdks.kkrwu tks oSf'k"Vî ok xq.kkP;k vk/kkjkoj ,[kknh O;ä h ,[kkns dk;Z ok inklkBh ;kX; ekuyh tkrs (Jñāna, anubhava, śiksana ityādīncyā drstīkōnātūna jē vaiśistya vā gunāncyā ādhārāvara ēkhādī vyaktī ēkhādē kārya vā padāsāthī yōgya mānalī jātē)

Gloss 5: dks.kr¢gh dykdkS'kY;kp] dykdqljhp] ;kaf=d oxSjs dke T;klkBh Kkuk0;frfjä dkS'kY; vkf.k ljkokph xjt vlrs (Kōnatēhī kalākauśalyācē, kalākusarīcē, yāntrika vagairē kāma jyāsāthī jñānāvyatirikta kauśalya āni sarāvācī garaja asatē)

Among all the senses gloss 1, 4 and 5 are very close to each other, so these are called as fine-grained senses, whereas at coarse grain level these sense are treated as one. At lexical level the intersection or similarity score will treat these glosses as a single gloss, the proposed WSD framework representing each word as a contextual probability, so it is treating all the six glosses distinctly at fine level. Hence for the given sentence, it is disambiguating the word dyk (Kala) as a gloss 5: dk.krgh dykdkS'kY;kp] dykdqljhp] ;kf=d oxSj dke T;klkBh Kkuk0;frfjä dkS'kY; vkf.k ljkokph xjt vlr (Kōnatēhī kalākauśalyācē, kalākusarīcē, yāntrika vagairē kāma jyāsāthī jñānāvyatirikta kauśalya āni sarāvācī garaja asatē). In this way, we have disambiguated the Marathi words at fine-grained level.

After calculating the intersections score between context and extended gloss embedding the highest score sense is assigned to the ambiguous word, if there is a tie in a highest score value the first sense occurring chronologically is assigned to the ambiguous word.

In the contextual representation of words, the uni-directional LSTM architecture of ULMFiT, it processes the word sequence in single direction with maximum 13 words context size, so it ignores compounded effect of words, which are yet to process, so limits the performance of WSD. Surely the bi-directional contextual representations will resolve these issues and will improve the speed, memorization and performance of WSD.

The performance of the proposed WSD approach is constrained by the coverage of IndoWordNet and the strength of the iNLTK's trained marathi model.

Conclusion and Future Work

In the reported work, we have investigated the effectiveness of ULMFiT-based transfer learning for contextualized distributed semantic representation and leveraged the benefits of meanings in extended gloss and its intersection with sentential context for the task of Marathi WSD. We have used multivariate probability distribution for quantifying the intersections. The proposed approach is empirically evaluated using traditional performance metrics like accuracy, precision, recall and F_1 score on the given test-bed. For detail investigation, we have measured the PoS category-wise performance. The results of our experimentation imply the feasibility of this approach for Marathi WSD task effectively. The results can be improved by strengthening the coverage of verb PoS category in IndoWordNet.

Acknowledgement

This research work is 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.

Conflict Of Interest

Authors and source of support does not have any conflict of interest

Reference

[1] Banerjee, S., & Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. Eighteenth International Joint Conference on Artificial Intelligence, (pp. 805-810). Mexico.

[2] Bhatt, B., & Bhattacharyya, P. (2011). IndoWordNet and its Linking with Ontology. ICON-2011: 9th International Conference on Natural Language Processing,.

[3] Bhingardive, S., Singh, D., V, R. M., & Bhattacharyya, P. (2015). Using Word Embeddings for Bilingual Unsupervised WSD. Proceedings of the 12th International Conference on Natural Language Processing (pp. 59–64). Trivandrum, India: NLP Association of India.

[4] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). EnrichingWord Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 135-146.

[5] Brownlee , J. (2017). What Are Word Embeddings for Text?, Machine Learning Mastery. Machine Leraning Mastrey .

[6] Faltl, S., Schimpke, M., & Hackober, C. (2019). Universal Language Model Fine-Tuning (ULMFiT) State-of-the-Art in Text Analysis.

[7] Hochreiter, S., & Schmidhuber, J. (1997, Nov). Long Short Term Memory. 9(8), pp. 1735–1780.

[8] Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification.

[9] Lesk, M. (1986). Autometic Sense Disambiguation Using Machine Readable Dictionories: How to Tell a Pine Cone from Ice Cream Cone. 5th Annual International Conference on System Documentation (pp. 24-26). Toronto: ACM.

[10] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2018). Learned in Translation: ContextualizedWord Vectors. arXiv:170800107v2.

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations ofWords and Phrases and their Compositionality. Neural and Information Processing System (NIPS), (pp. 1-9).

[12] Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic Regularities in Continuous SpaceWord Representations. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751). Atlanta, Georgia: ACL.

[13] Navigli , R. (2009, February). Word sense disambiguation: A survey. ACM Computing Surveys, 41(2), 69.

[14] Navigli, R. (2009). Word Sense Disambiguation: A Survey. ACM Computing Surveys, 41(2), 1-69.

[15] Pan, S. J., & Yang, Q. (2010, October). A Survey on Transfer Learning. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1345-1358.

[16] Patil, S. S., Bhavsar, R. P., & Pawar, B. V. (2020, June). Contrastive Study and Review of Word Sense Disambiguation Techniques. International Journal on Emerging Technologies, 96-103.

[17] Patil, S. S., Bhavsar, R. P., & Pawar, B. V. (2021). Path and Information-based Structured Word Sense Disambiguation. ICInPRo-2021. Bengluru: Springer CCIS.

[18] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for Word Representation. ACL, (pp. 532–1543). Doha, Qatar.

[19] Peters, M. E., Neumann, M., Iyyer, M., & Gardner, M. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter (pp. 2227–2237). New Orleans, Louisiana: ACL.

[20] Sherstinsky, A. (2021). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. Physica D: Nonlinear Phenomena, Special Issue on Machine Learning and Dynamical Systems, 1-43.

[21] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. CoRR, 1-9.