International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET218617

Performance Analysis of Sense Embeddings in Multilingual WSD Framework

Mr. Prashant Y. Itankar* Dr. Nikhat Raza
Department of Computer Science and Engineering, MPU, Bhopal, Madhya Pradesh. India

ABSTRACT

Article Info
Volume 8, Issue 6
Page Number : 140-148

Publication Issue :
November-December-2021

Article History
Accepted : 01 Dec 2021
Published: 12 Dec 2021

Execution of Word Sense Disambiguation (WSD) is one of the difficult undertakings in the space of Natural language processing (NLP). Age of sense clarified corpus for multilingual WSD is far off for most languages regardless of whether assets are accessible. In this paper we propose a solo technique utilizing word and sense embeddings for working on the presentation of WSD frameworks utilizing untagged corpora and make two bags to be specific context bag and wiki sense bag to create the faculties with most noteworthy closeness. Wiki sense bag gives outer information to the framework needed to help the disambiguation exactness. We investigate Word2Vec model to produce the sense back and notice huge execution acquire for our dataset.

Keywords :- Natural Language Processing; Word sense disambiguation (WSD)

I. INTRODUCTION

Expanding requests by the client to get to message information in different languages opens up the entryways of multilingual NLP and WSD has ended up being a critical stage in execution improvement of numerous NLP frameworks. The current exploration is more centered on monolingual WSD, precision of WSD frameworks is a long way from being palatable and multilingual WSD has not accomplished agreeable outcomes because of inadequate asset accessibility [7]. The accessibility of multilingual word references has improved sense disambiguation utilizing multilingual substance which portrays the requirement for multilingual WSD [23]. It additionally opens up an alternate method of drawing closer multilingual WSD by utilizing BabelNet, a wide ontological design investigating semantic information.

Depending just on multilingual information based framework might hamper the development of WSD frameworks and however multilingual word references give wide inclusion investigating the interconnected metaphysics structure, different issues actually still need to be worked out, for example, formal people, places or things are not piece of the word reference and relationship between's most successive words and uncommon logical words need word reference inclusion. Outside information as far as crude text is required which is given utilizing word and sense inserting [22]. Our examination utilizes word and sense embeddings to make a semantic word cloud by planning a wiki sack notwithstanding the sense pack. Wiki pack is planned utilizing Wikipedia as it is the biggest reference book which covers the greater part of the data set fundamental for disambiguation. It gives world information to the WSD framework notwithstanding the corpus utilized for the review, by bringing the connected words conclusion in the inserting space and bodes well portrayal additionally boosting the disambiguation precision.

II. LITERATURE REVIEW

Word2Vec model [1-10] gives an effective tool to assessing vector model utilizing the corpus. Scientists utilized this model to configuration word embeddings via preparing corpus as it brings semantically comparable words conclusion in the vector space and acceptable outcomes are accomplished. A sense sack was made [2] utilizing word reference assets like synset individuals, model sentences, hypernymy and hyponymy subsets. A review was introduced on WSD [11] featuring the inspiration for tackling the vagueness of words and giving portrayal of the assignment. The idea of WSD in multilingual setting [12-13] presents by utilizing huge exhaustive ontological organization Babelnet. Creators take the interpretations of the equivocal word in different dialects as word uncertain in one language may not be questionable in different languages. Accuracy accomplished was 54.3% when tried on SemEval 2010 dataset. Examines communicating implications as far as summarizes in and novel methodology was proposed to remove rewords by turning through eight languages while separating faculties in the info language.

It is observed that very little work is accounted for on WSD in a multilingual setting supposedly and it should be investigated utilizing different best-in-class WSD strategies. Outer assets presently being accessible is utilized to bring multilingualism into WSD framework and investigate different languages which help in the disambiguation cycle.

III. PROPOSED METHODOLOGY

The review is centered around successfully settling WSD and to work on the precision of the framework. The WSD framework takes input as regular language and our review utilizes English language as information. The information sentences go through the pre-handling stage where the info sentence is separated into words and grammatical forms labels are joined to each word. The info is then taken care of to the WSD framework and utilizes multilingual word reference to perform disambiguation. The methodology used to perform sense disambiguation in the word reference utilizes information based methodology Adapted Lesk calculation.

Given a text w1 w2...wn of n words, our decision of multilingual word reference permits us to look at the implications (i.e., sparkles) of words that are associated with the words to be disambiguated through the different connections characterized in BabelNet. This gives a more extravagant wellspring of data and, likewise further develops disambiguation exactness. The gleams of the different synsets that the word has a place with, just as the shines of those synsets that are identified with them through the relations are used to perform sense disambiguation. We disambiguate each in turn considering the likeness between the shine related to each feeling of the objective word wi in the cosmology and the unique circumstance. The significance whose gleam has the most elevated similitude is chosen. The setting could be addressed by a subset of encompassing words or the entire text where the word happens.

The sense stock utilized for our review is BabelNet word reference, an extremely huge multilingual semantic organization constructed depending on both WordNet and Wikipedia. BabelNet is chosen as shines are more extravagant and contain text from Wikipedia. It is multilingual, and can be applied to a few dialects, and it likewise contains data about named substances, along these lines a calculation utilizing BabelNet could be conceivably used to disambiguate substances.

A. Algorithm

```
Define Input: a word sequence μ = (W1, W2…, Wn)
ambiguous word t € μ with k senses ws1, ws2…. wsk
taken from BabelNet
Features (S, G, H, HP, HO, HG, HOG) from the
BabelNet BNT
Output: A distribution of scores for the senses of
target word t
S 🡐 Synonym_BNT(t)
Ctx 🡐 μ - t
FB_CTX = {y| y € Features(cs_i)}
G= {V, E}
For each FB_CTX:
δ score [] = {score (FB_CTX)}
For each S:
Synset_score [] = {score (S)}
Δ Score = Σ( δ score [], Synset_score [])
Global_Score 🡐min (Δ Score)
return Global_Score
```

Algorithm begins by representing a sentence in terms of sequence of words **μ** and the different senses of the ambiguous word **t** are collected in *S* represented as synonym set from the BabelNet. Context words are collected in *Ctx* which includes all the words other than the target word *t*. Context words are the clue words which helps in sense disambiguation. A feature bag is generated which represents the ontological features such as Hypernym, Holonym and Hyponym of context words from the BabelNet. Ontological features of target word are not considered as synset definition of target word is sufficient for disambiguation and features of target word generates too many scores which complicates the task. Feature bag of context words generates a score for each feature of the sense using the formula shown below.

$$score_j = \sum p \in paths\,(sj)\, \frac{1}{e^{length(p)-1}}$$

The score represents the distance of the target word from the clue word in the ontology represented as a graph G (V, E) where V represents the vertices and E represents edges of the graph. The generated score is added to the score of the synset of the target word. This generates a Δ Score and minimum distance is calculated to represent the correct sense which is the global score returned as the answer.

B. Working

Consider the example *Alex deposited money in the bank.* The feature bag is generated for the context words *money* and *deposited* using BabelNet for ontological features. The synset id for the target word is also generated. The scores for the context words are represented below for the different features.

deposited
bn:00083093v deposit - hypernym -bn:00088815v 00088815
bn:00083093v deposit - hyponym - bn:00092571v 00092571
money
bn:00055644n money - hypernym - bn:00054159n 00054159
bn:00055644n money - subclass_of - bn:00050571n 00050571
bn:00055644n money - subclass_of - bn:05130811n 05130811
money - part_holonym - bn:00055652n 00055652
bn:00055644n money - hyponym - bn:00011380n 00011380
bn:00055644n money - hyponym - bn:00016448n 00016448

bn:00055644n money - hyponym - bn:00046466n
00046466
bank
bn:00008364n 00008364 bn:00008371n 00008371

For simplicity of explanation, we consider the word *deposit* and its score representation. The score bn: 00083093v represents the synset identification number of the sense *deposit* meaning "putting into bank account" along with its parts of speech which is 'verb'. The *bn* stands for Babel score. The score hypernym bn: 00088815v represents the hypernym of sense *deposit* which means "transfer possession of something concrete or abstract to somebody". The score δ generated for the context words and synset_score generated for the target word are represented in an array shown below.

δ score =
[88815, 92571, 54159, 50571, 5130811, 55652, 11380, 16448, 46466, 3510935, 16660687, 18175630, 4915814, 14777497, 4273111, 14055744, 3406922, 4982974, 725725, 15672138, 7749766, 2340094, 2656554, 3115187, 12856654, 17863888, 5154, 12054, 36839, 71099, 74975, 77506, 3340, 16448, 34202, 46675, 70995, 77506, 336097, 1969490, 2124867, 2476555, 3255581, 3662212, 3758092, 3779823, 2283385]

Synset_score=
[8364, 8371, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

The array of Synset_score shows that only two senses of the target word are generated namely 'finance' sense and 'river' sense. The summation of the above two arrays result into Δ Score and the minimum score is selected to represent the correct sense of the target word shown below.

Δ Score =
[177082448, 177082147, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

The global score is calculated by picking the minimum of the two scores as lesser the distance between two words, more the words are closer to each other in ontology. The global score represents the index position of the Δ Score which is minimum which is the correct sense of the target word 'bank' meaning "A building in which the business of banking transacted".

Global score=
[1.0, 2.718281828459045, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, .0,0.0,.0,0.0,0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

C. Results

The goal of our experiments is to initially establish a competitive baseline using unsupervised learning algorithms and the best combination of features that yield the highest accuracy to boost the disambiguation system. WSD framework evaluation is performed on SENSEVAL-2 corpus for English language for English lexical sample task.

Experiments were conducted with Monolingual input to the WSD system with English language in concern. System was tested for 25 polysemous ambiguous nouns and results were evaluated based on the global score. Owing to multiple senses, system generated various scores from which we selected the minimum score as the winner sense. Tested for 2000 instances, for polysemy nouns, average accuracy observed was 40%. As BabelNet is a huge ontological network, testing the system for huge instances results in generative score and less accuracy. For many instances, our system generated similar score failing to pick up the winner sense which also disturbs the accuracy of WSD system. First sense is picked up as the winner sense in case of multiple similar scores. It is also observed that proper nouns from our instances were not part of the dictionary definitions which failed to generate proper scores. Also, dictionary definition being short lacks strong clues which fail the disambiguation accuracy. Words not part of BabelNet failed to generate the features and resulted in inaccurate results and in some cases the context words were not strong enough for disambiguation which drifted apart from the target word resulting in lower accuracy.

It is seen that synset individuals alone are not adequate for distinguishing the right sense. This is on the grounds that some of synsets have an extremely short definitions and synset individuals alongside the shine further develops results as gleam individuals are more straightforward in characterizing the sense. The other explanation is to cut down the effect of subject float which might have happened on account of polysemous synset individuals. Additionally, it is likewise seen that utilizing hypernymy, hyponym and holonymy alongside their synset shine gives better execution improvement in deciding the most incessant feeling of a word. The best outcomes were accomplished when every one of the elements were joined together.

TABLE I PERFORMANCE ANALYSIS USING BABELNET FEATURES

Features	Global score	Accuracy
S	0.0869	34
S+G	0.1923	27
S+G+H	0.1666	33
S+G+H+HP	0.0588	38
S+G+H+HP+HG	0.3333	42
S+G+H+HP+HG+ HO	0.0526	47
S+G+H+HP+HG+ HO+HOG	0.5238	52

Features of Babelnet senses are extracted from the synset(S), gloss of synset member (G), hypernymy (H), hyponymy (HP), synset gloss of Hypernymy-hyponymy relation (HG), holonymy (HO) and gloss of holonymy (HOG). We tested these features on 2000 instances and results are represented by taking the maximum of the global scores received represented in Table 5.4. It is observed from the table above that combining all the features of Babelnet senses together gives us an improved accuracy of 50%.

D. Wiki sense bag creation using word embeddings

The observations depict that the dictionary alone is not a sufficient resource to perform disambiguation and there is a need to provide world knowledge to the system as dictionary definitions are limited and failed to perform disambiguation. The context and target words if drifted apart complicate the disambiguation process. The words related in context needs to be close to disambiguate a sense. The similar words can be kept together using word and sense embeddings which also represents world knowledge to the system. Clue words from the context help to disambiguate a target word; it is this dependency that the system fails to understand. This dependency needs to be identified and the system should be trained for understanding the relationship between the context word and the ambiguous words. This is handled using word embeddings which are immensely popular in NLP.

Word embeddings are based on distributional hypothesis which works under the assumption that similar words occur in similar contexts. Word representations mean converting plain text into numeric form as machine learning and deep learning architectures are not capable of processing text in raw form. It represents embedding continuous vector space with lesser dimensions and word embedding will be trained using *word2Vec* tool as described by Mikolov. The training proceeds by presenting different context-target words pair from the corpus thus preparing an ensemble model for all the ambiguous words in the vocabulary.

Sense specific word embeddings are generated which represents word bag and disambiguation is performed using cosine similarity calculation between the training and testing vectors. The similarity measure is calculated by considering the cosine similarity between the word representation of context vector and sense bag representation.

$$\text{Cos}|(\text{vec}(w)\ \text{vec}(SB)) = \frac{\sum_{i=1}^{n} w * SB}{\sqrt{\sum_{i=1}^{n} w^2} \ * \ \sqrt{\sum_{i=1}^{n} SB^2}}$$

Where vec(w) is the word embedding for word *w*, SB represents the sense bag and vec(SB) is the sense embedding representing the combined score of ontology bag and the wiki sense bag.

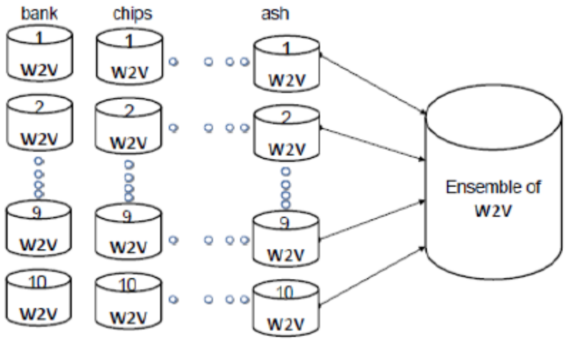

Figure 1. Ensemble method for word2vec

Word specific sense embedding generated satisfactory results and improvising the system by providing encyclopedic knowledge will not only boost the disambiguation accuracy but will also make the system more intelligent. Encyclopedic knowledge is provided to the system using Wikipedia by creating a wiki sense bag of encyclopedic data. Wiki sense bag is created which generates sense embeddings that is a vector representation of Wikipedia of ambiguous words. The aim is to build an intelligent system which will help to disambiguate ambiguous words and word and sense embeddings are the best way to provide world knowledge to the machine as machine lacks the world knowledge which humans possess.

Experiments were performed on SENSEVAL-2 dataset whose aim is to encourage more languages for WSD task. The evaluation was performed on the gensim word2Vec model as the aim was to provide world knowledge to the system for improvement in accuracy Results were evaluated using the F1 score and received promising results for gensim and wiki sense bag evaluation.

TABLE II PERFORMANCE COMPARISON OF SENSEVAL -2 DATASET

System	F1 Score
MFS baseline	59.88
Chaplot[48]	60.5
UMFS[65]	52.34
Our approach	60.2

The graph showing the performance comparison of the state-of-the-art methods with our approach is presented below.

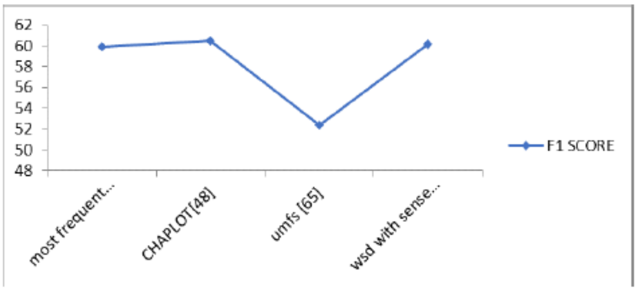

Figure 2. Performance comparison using SENSEVAL-2

It is observed that SENSEVAL-2 dataset is more generalized and suited for application-based natural language processing and not specific to WSD which fails to generate satisfactory performance. The data set should be more refined and suited to WSD work.

E. Neutral language code generation

To resolve the efforts needed for multilingual MT, our research aims to generate a neutral language which is independent of language pair and hence applicable for multilingual MT as it provides an open platform for handling MT in various languages. The source language is processed and converted into neutral language code for translation represented in figure 6.2 and the target language is generated from the neutral language code. Neutral language helps for accurate translation as it contains all the information needed for translation about a particular word like its correct sense identification, parts of speech, number, gender, case, tenses etc. encoded. This information is vital to the translation process as it captures the knowledge about the natural language efficiently and accurately. The neutral language code appears at the top of the MT pyramid and assumes that it is possible to convert source texts into representations common to more than one language and from such representation's texts are generated into other languages. The translation is thus in two stages: from the source language to the neutral language and from the neutral language to the target language. It requires an analyzer for each source language and a generator for each target language.

Neutral language generation for a language is not an easy task as it requires handling the ambiguities associated with the source language. Lexical semantic ambiguity is efficiently handled using WSD which play an effective role in the translation task. The ambiguity resolution leads to building a framework for multilingual translation where information can be directly translated from any source language to any target language using neutral language code. Words after disambiguation are converted into unique representation making it ready for translation and this unique representation termed as neutral language code is formed using a binary combination of 30-bit unique code where each bit represent significant information about the disambiguated polysemy noun.

TABLE III PERFORMANCE COMPARISON OF SENSEVAL -2 DATASET

Noun	Code
BAT चमगादड़$_{HIN}$ চামচিকে$_{BG}$ वटवाघूळ$_{MAR}$	**0001000110101100101011110011XX** 0001 – UNIQUE IDENTIFICATION 000- PARTS OF SPEECH 1101- TYPE OF NOUN 011- NUMBER 001- GENDER 010- PERSON 11110011- TENSE INFORMATION IF VERB XX – RESERVED BITS

Table III represents 30-bit code for polysemy noun '*bat*'. First four bits represents the sense identification of word '*Bat*' and the next three bits represents its parts of speech, the next four bits represent its type whether it is a common noun or proper noun. Noun represents a number, gender and person which are represented in consequent bits. The number represents singular or plural number like *bat* or *bats*. Gender is masculine, feminine or neuter gender and a person is first, second or third person. There are certain words which have more than one part of speech. The given word being a verb, its information is stored along with its cases and tense information in the next seven bits. Last two bits are reserved bits for performing any kind of modification to the code as the framework is open for future modification to the neutral language code.

IV. CONCLUSION

The ontological structure of the dictionary using the knowledge-based approach also boosts the disambiguation accuracy and observed an overall accuracy of 40%. Limitations of the dictionary-based approach and lack of world knowledge leads to the creation of word and sense embeddings which are useful in NLP for disambiguation task as it gives promising results. The results discussed showed that sense embeddings after sense bag creation helped to improve the disambiguation accuracy and came close to baseline accuracy. Neutral language code is unique in the sense that it covers all the information other than sense identification and parts of speech. Additional information is encoded into the code like various factors of a noun which is crucial for translation. Innovative neutral language is independent of any language irrespective of multiple senses belonging to target language reducing the complexity. The design of innovative neutral language compels to bring various natural languages under one platform and design an open framework for multilingual MT. The open framework will benefit the researchers working on a new language as WSD is effectively resolved and improves the effectiveness of translation from any language to any other language using neutral language. Unique code assigned to each word in the context after ambiguity resolution is universal for English as well as the target language aimed for translation and kept in the knowledge base. The enriched knowledge base will boost the further processing of natural languages and this unique language code shall be used in the future to strengthen the knowledge base further and used for MT. The advantage is the uniqueness of the code which will help to improve various applications of NLP in the future.

V. REFERENCES

[1]. Bhingardive, S., Singh, D., Murthy, R., Redkar H., Bhattacharya, P., “Unsupervised Most frequent sense detection using word embeddings”, Proceedings of the 2015 Conference of the North American Chapter of the Association of Computational Linguistics:Human Language technologies, Denver, Colorado., 2015.

[2]. Bhingardive S., Shaikh S., Bhattacharyya P.; "Neighbours Help: Bilingual Unsupervised WSD Using Context." ACL, 2013.

[3]. Mikolov T., Kai C., Greg C., Jeffery, D.;“Efficient Estimation of Word representations in vector space”, In Proceedings of workshop at ICLR. 2013.

[4]. Fu R. Guo J., Qin B., Che W., Wang H., Liu T.; "Learning semantic hierarchies: A continuous vector space approach." IEEE Transactions on Audio, Speech, and Language Processing 23.3, 2015 pp. 461-471.

[5]. Schutze H.; “Word space”, Advances in neural information processing systems, 1993, pp. 895-902.

[6]. Chen X., Liu Z., Sun M.; “ A unified model for word sense representation and disambiguation”, Proceedings of the 2014 conference on empirical methods in natural language processing(EMNLP) Qatar, 2014, pp. 1025-1035.

[7]. Iacobacci I., Pilehvar M., Navigli R.; “ Embeddings for word sense disambiguation: An Evaluation study”, Proceeedings of the 54th Annual meeting of the Association for Computational Linguistics, Germany, 2016, pp. 897-907.

[8]. Trask A., Michalak P., Liu J.; “Sense2vec- A fast and accurate method for Word sense disambiguation in neural word embeddings”, ICLR, 2016, pp. 1-9.

[9]. Taghipour K., Ng H.; “Semi-supervised word sense disambiguation using word embeddings in general and sepcific domains”, Human Language Technologies: The 2015 Annual conference of the North American Chapter of the ACL, Colorado, 2015, pp. 314-323.

[10]. Sugawara H., Takamura H., Sasano R., Okumura M.; “Context representation with word embeddings for WSD”, Conference of the Pacific Association for computational lingusitics PACLING 2015, pp 108-119.

[11]. Navigli R.; “Word sense disambiguation : A survey”, ACM computing surveys, Vol 34, No. 2, Article 10, 2009.

[12]. Navigli R., Ponzotto P.; “Multilingual WSD with just a few lines of code: the Babelnet API”, Proceedings of the 50th Annual meeting of the Association for Computational Linguistics, Korea, 2012, pp. 67-72.

[13]. Aziz W., Specia L.; “Multilingual WSD-like constraints for Paraphrase extraction”, Proceedings of the 17th Conference on computational Natural language learning, Bulgaria, August 2013, pp 202-211.

[14]. Montoyo A., Romero R., Vazquez S., Calle C., Soler S.; “The role of WSD for Multilingual natural language applications”, International conference on Text, Speech and Dialogue, Springer, 2002, pp. 41-48.

Cite this article as :

Mr. Prashant Y. Itankar, Dr. Nikhat Raza, "Performance Analysis of Sense Embeddings in Multilingual WSD Framework", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 6, pp. 140-148, November-December 2021. Available at doi : https://doi.org/10.32628/IJSRSET218617
Journal URL : https://ijsrset.com/IJSRSET218617