Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 10664 - 10674

Research Article

Implementation of Neighboring Word Feature Technique using CNN for Word Sense Disambiguation for Marathi Language

Swati G.Kale[a], Dr.Ujwalla Gawande[b], Varsha Nagpurkar[c]
[a,b] Yeshwantrao chavan college of engineering/Department of information technology, Nagpur, India
[c] St.Francis Institute of Technology/Department of computer Engineering Mumbai,India
Email: [a]swati79kale@gmail.com, [b]ujwallgawande@yahoo.co.in

Abstract

Basically, WSD is a linguistic method for automatically defining the true sense of a word in the particular sense of context. WSD analysis is worldwide challenge in AI-computational field and NLP. The various operations of WSD system include Information retrieval system (IRS), Lexicography, Speech & Text Processing, Machine Translation. In every language of this world a single word may have different meaning in various context of sentence. For the human being it is very easy and simple to determining the exact meaning after reading and analyzing the complete sentence. Our research work includes, identifying the appropriate sense of lexeme for already available context where it has been used and finding exact relationship with the lexicons which can be done using NLP techniques This paper attempt to develop and implement a unique technique for WSD of different words in Marathi language, which is one of the most popular regional Indian languages. In the experiment, we have developed and implement our novel knowledge based techniques with wordnet of Marathi language. Our proposed approaches uses knowledge based word sense ambiguity algorithm for WSD in Marathi language. The proposed unique approach is capable of identify and solve the ambiguity and provides sense of ambiguity is very less amount of time. Accuracy achieved for proposed approach is 87.22%.

Keywords: *Word Sense disambiguation, Natural Language processing, neural network, wordnet, Marathi words*

Introduction

Recently NLP attracted lots of researchers for offering inter-displinary fields of research that includes language based study and processing of that linguistically with computer system that generates more new teeth known as computational linguistics [6]. The main motive of NLP is to deal with the advance developments of practical life applications and usage of that software applications using artificial intelligence system. Basic functionality of these types of systems is Morphological analysis, phonological analysis, syntactic analysis, semantic analysis, pragmatic analysis[7]. In these types of application system ambiguity plays a very classic and important role .As we all know that in this world there are majority of languages available and been used by local region people. In every language there are lots of words which contain different act of sense in different context or lexemes. Same word may have different meanings and sense in two different context[8]. These types of words are known as ambiguous word. A single word may have more than one meaning in the sentence where it has been used. Let us take an example of word "Pratima" in context it's like "Pratima is playing with her friends" and "Maa durga Pratima is very giant and beautiful" here the word "pratima" in first sentence is a name of a girl who is playing with her friends where as in second sentence same word "pratima" have meaning "statue" of lord maa durga. The actual meaning of Pratima is "Statue", but is represents different meanings in the context where it has been used[9].

A single word may have various different meaning with the sentence it has been used. Identifying the appropriate meaning of these types of ambiguous words with the particular context is known as WSD. WSD can be defined as ability of a system to determine the exact sense of the ambiguous phrase used in a particular sentence context..WSD depends on knowledge data stored in database.WDS system has following functionalities: Initially the system will take input as collection of sentences (lexemes) or words. Then different NLP methodologies will be applied that will make use of any one of the source of the data resources to determine the exact and best – suitable sense of the phrase where the word has been used or written. Word that has different meaning with variability in the context where they have been used & identifying what meaning of the word is to be intended in that context. This is one of the major problem that is faced by every NLP system framework which is also called as a "Lexical – Semantic Ambiguity"[10]. This Lexical Semantic Analysis(LSA)is a mostly used methodologies in NLP which can be used and depicted by various types of jobs execution, various sources of knowledge databases, AI applications, computational systems, language-based systems, assumptions.

In figure 1, the resources used for WSD is depicted

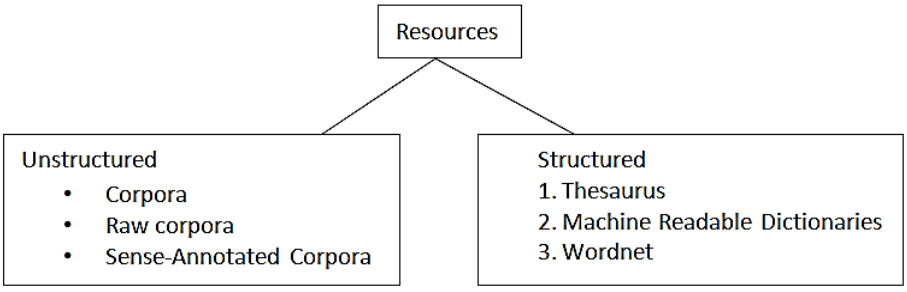


Figure 1: Resources Used for WSD

As we all know that analyzing the correct meaning or sense of word in a particular context will be very easy for human being as they can use their own common sense, but for machine it is very difficult as machines don't have capacity to think. Instead machine work on only logic and methods. It does not have its own calculating and thinking capability. Till now various procedures that can be used to handle this WSD issue are supervised, knowledge based and unsupervised[11]. These methodologies have been discussed by various researchers in various research papers.

1. **Supervised Approach:** In this method, WSD is done with the use of already existing data set that has been generated in previous exceptional steps. These learning data sets consist of sense related to that word or phrase.
2. **Knowledge–Based Approach:** This approach is completely depend on various external sources of data (information) such as online Wikipedia, semantic database, dictionaries, some machine readable - documents, etc. to generate or develop sense definitions of the phrase in the context where it has been used.
3. **Unsupervised Approach**: In this approach word sensing is done in two steps. In the first step clusters will be developed for the sentences using some typical clustering algorithim, and then these sentencedclusters are grouped with some relevant sense with the help of some language developer or expert.

In Asian world languages like Hindi, Bengali, Tamil, Punjabi, Marathi, Malayalam etc., the various linguistics sources are not sufficient. The databases such as online dictionaries, thesaurus, online information are not sufficient data bases for sense referencing, as a result already existing approach are not able to generate a good result[12]. In this research paper, a Neighboring Word Features which is based on CNN system is proposed for WSD in Marathi language.In our experiment, the test data has been generated from Marathi text corpus, developed in project of the government of India. Our experimental result produces 87.22% accuracy in sensing the word disambiguity.

Revıew of Lıtreture

In this section we have presented various reviews of literature of already exiting approaches presented by various researchers. [1]Explores the lesk algorithm for resolving WSD issues in phrase context. This has been the initial first machine understandable language form dictionary based technique. This technique has been completely based on the cross - overlapping of the word's meaning in dictionary definition for particular phrase in a context where it has been used or written. Basically this lesk algorithim works on selecting a very short phrase (word) from the context which consists of ambiguous phrase. After that the definition of the word from the dictionary (GLOSS) for every senses of that ambiguous phrase has been compared with the definition of another phrase in that specific context. After that ambiguous phrase is designated with that specific sense, whose gloss value is having the peak frequency value with the gloss value of another word of the phrase? In [2] author describes the heuristic approach for WSD where the heuristic tress has been developed for the collection of ambiguous word (phrase). The heuristic properties have been evaluated to determine the sense of the ambiguous word. Basically three various heuristic values have been used for estimating WSD system properties such as (1) Mostly used sense, (2) One specific sense for collection of word and (3) One sense for per phrase. The first heuristic approach works by identifying each and every similar sense

that a phrase may have & it is generally correct sense that occurs very often than the others. Second method stores the meaning of all words and its occurrence for a prescribed context. Third approach selects sense for collection of different phrases in a particular context and presumes the phrase which is nearest to that word sense which gives the strong and static signal to the sense of an ambiguous phrase. In [3] represent a hybrid approach for WSD which is based on structural and semantic interconnection methodology. This technique use more than one knowledge sources like word-net, Wikipedia, corpus data or any small corpora database. This method allows the phrase to capture necessary data which is available in encoded form in all types of word net as well as draw syntactic generations which form basic tagged corpus. This technique has got much sustainability in very less execution time for resolving WSD by using general purpose methods and classifiers [4]. In [5]machine learning based methodology (i.e. Supervised has been discussed by the author. It suggest that extracting the exact "Sense Definitions" or "Different Usage patterns" from corpus database gives more accurate result of Word sense disambiguance. Although in majority of supervised techniques which gives excellent result are not a general purpose WSD based system but are a particularly word specific classifier. In[5] author discussed Unsupervised and semi-supervised machine learning algorithm that has capability just like a supervised algorithim. The actual fact is that these two type of algorithm are able to work with much less or non-tagged data(Information) which make them eligible as a popular algorithm for regional languages like Marathi, Hindi, Bengali. This algorithm has only one disadvantage that it is difficult make this model as a general purpose wide coverage model.

Table 1: Comparative analysis

Sr.No	Technique Name	Advantage	Disadvantage
1.	Lesk Algorithim-Dictionary based technique	1.Capable of selecting very short phrase	1.Not suitable for finding long ambiguous word o phrase. 2.This method cannot be used in developing general purpose model
2.	Heuristic Approach:	1. It gives accurate result than previous approach. 2.It works on stepwise manner using three different approach. 3. Easy to understand.	1. It has needs more execution time to sense the correct meaning of the ambiguous word. 2. As uses three procedures it need more time to implement. 3. It cannot be used for developing wide coverage classifier model.
3.	Hybrid: It uses Structural Semantic Interconnections	1. Uses Combination of more than one knowledge sources hence gives more precise result. 2. This technique is most suitable for developing general purpose classifier model.	1. It requires more memory for process execution. 2. As it uses more than one knowledge data source it need more processing time.
4.	Supervised Machine Learning Algorithim	1. It gives the accurate result by determining the exact sense of word or phrase. 2. It is an mostly used algorithim.	1. It need more “usage pattern” to identify the exact sense hence need more storage memory. 2.It requires more time to give result.
5.	Unsupervised or Semi-Supervised Machine Learning Algorithim	1. It has capability of working like a supervised ML Algorithim hence it has same result as a supervised ML algorithim. 2. It is mostly used to develop wide coverage classifier model.	1. It requires more storage memory. 2. It requires more execution time.

Iıı. Database Used Wıth The Proposed Work

Marathi Wordnet: Marathi wordnet is basically an online available semantic words dictionary, which is used for getting or determining semantics information of typical Marathi words (Dash2011). This wordnet gives all various information related to Marathi words and the relationship status that exist between these words. This Marathi wordnet has been created and implemented with the use of some specific language development tools available in IIT, Bombay (Indian Institute of Technology, Bombay). Here, in this Marathi wordnet a user may find any Marathi word and its original meaning. Also it will provide each and every grammatical category of that word like noun, adjective, verb or adverb which has been searched by user. One should note that a single word can be appeared in more than one grammatical category and every grammatical category may have different or more than one sense of that word. Hence, this Marathi wordnet also provides data for this type of words and its related

grammatical category. It also provides related sense of every category for the particular word which has been searched by an user in Marathi WorldNet With collective information of each and every grammatical category and their different sense of Marathi words, it also consists of a set of various information for the various Marathi word in the wordnet as given below:

1. Hierarchical semantics representation (Ontology)
2. Various examples where a word can be used
3. Similar words with their meanings (Synonyms)
4. Relationship of Semantic and lexical analysis
5. Part of Speech
6. Meanings of every word
7. Data-Id
8. Word Count
9. POS Tagger

Till date Marathi wordnet contains 36,890 words that cover every lexical category such as verb, noun, adjectives, adverb, etc.

Preprocesseıng Modules Used

The proposed system is having various modules which are having their independent working algorithms. All these modules are combined together in one main system to implement our proposed work.

Stage 1] - An algorithm has been developed for suffix removal, tokenization and stop word

Stage 2] - An paring algorithms has been developed for reading data suffix file, index file.

Stage 3]- An ambiguity detection algorithm.

Proposed Work

Design and development of Neighboring Word Features based Convolution Neural Network (NWF-CNN) for WSD for Marathi language:

In this section, complete proposed approach is presented, that combines all the above modules to obtain the desired output. The proposed approach basically works on neural network known as NWF - CNN. The approach has following exceptional steps as given below and also depicted in the form of block diagram as figure 2:

1. Feature extraction
2. Feature selection
3. Define hidden layer
4. Define Input layer
5. Define output layer
6. Apply filters
7. Average/max pooling
8. Standard. Deviation
9. Back Propagation

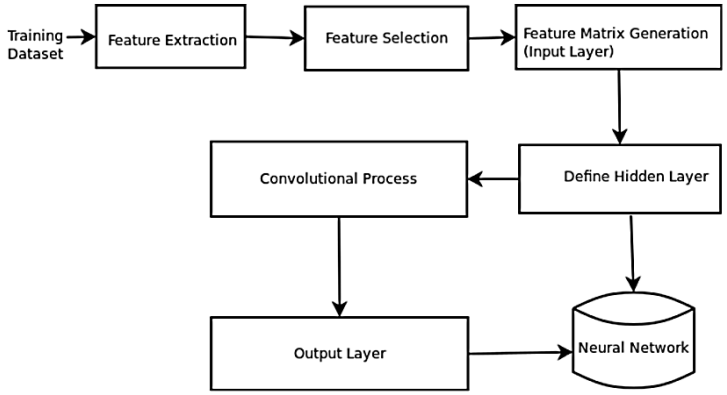


Figure 2: Processing block of Neighboring Word Features based Convolution Neural Network (NWF-CNN).

Working of each block is given below in detail:

Feature Extraction:

It is a basic building block of any text (data) processing system. Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier. Feature extraction process will automatically enhance the accuracy of the system. This feature extraction block will develop new advanced feature with the help of various combinations and transformations functions into the original set of feature of an input data (text). Selecting the document information will reflect the data on content word and calculation of weight is known as text feature selection. Here a basic set of feature of the text is selected by applying different effective techniques for reducing the actual dimensional set of featured space, hence it enhances training and inference speed of the system. The text feature extraction includes various sub processing block like text fusion, text filtering process, text mapping and clustering techniques. Typical techniques of feature extraction needs general handcraft text features. Designing of hand-craft techniques is a very lengthy process. The following features are extracted for every word in a sentence.

- Word
- Root
- Suffix
- Replacer
- POS
- Ontology

Feature Selection:

This processing block plays a very important role in text mining, like pattern classification, machine learning and data analysis. A good feature selection technique will comparatively reduced the price of feature measurement and enhance the classifier efficiency and accuracy of system. Hence the feature selection will provide a good pre-processing tool for resolving the text categorization.

For example:

Input given to feature Extraction Block="***श्यामघरीजातहोता***"

The following feature are selected for inputted sentence:

- FS ={Fwf, Fwp,Fwa,Fwn,Fwl}
- Fwf:-Feature of first word
- Fwp:-Feature of previous word
- Fwa:-Feature of Ambiguous word
- Fwn:-Feature of next word
- Fwl:-Feature of last word

Feature Matrix Generation:

- Feature matrix generation plays a very important role in text processing of proposed technique. Here as shown in figure 3, Layerd processing of the text is done.the first text matrix in monocolor form, represent a complete text that is given as input to the block. As an example, the sample input text is ""***श्यामघरीजातहोता***"".

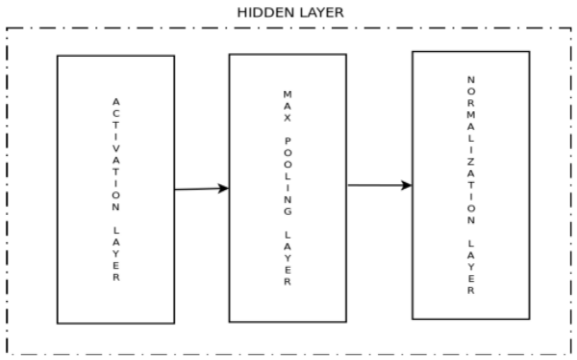


Fig 3: Layered processing of text

After doing the maxpooling and normalization functions the output generated is given in following figure 4.

For text processing in input layer, following calculative method are done:

1. Activation function:

$ai=\sum Fwfi*f1j + \sum Fwpi*F2j$

2. filter Matrix of hidden layer

For N=n gram filter Matrix

N=2(2*9)

N=3(3*9)

Layer is calculated as

R=En-N+1

3. Left position= $\frac{1-|Awp-w2p|}{\text{length of sentence}}$
4. Right position= $\frac{1-|Awp-w2n|}{\text{length of sentence}}$
5. Recall
6. Precision
7. F1 score (F-score or F-measure)

VI. Experimental Results and Conclusion

In this research, an approach is proposed for WSD by using wordnet for Marathi. Here small Marathi corpus is considered which has 4000 ambiguous sentences. WSD has been evaluated using various parameters such as recall, precision, F1-Score, and execution time and hence system accuracy has been determined.

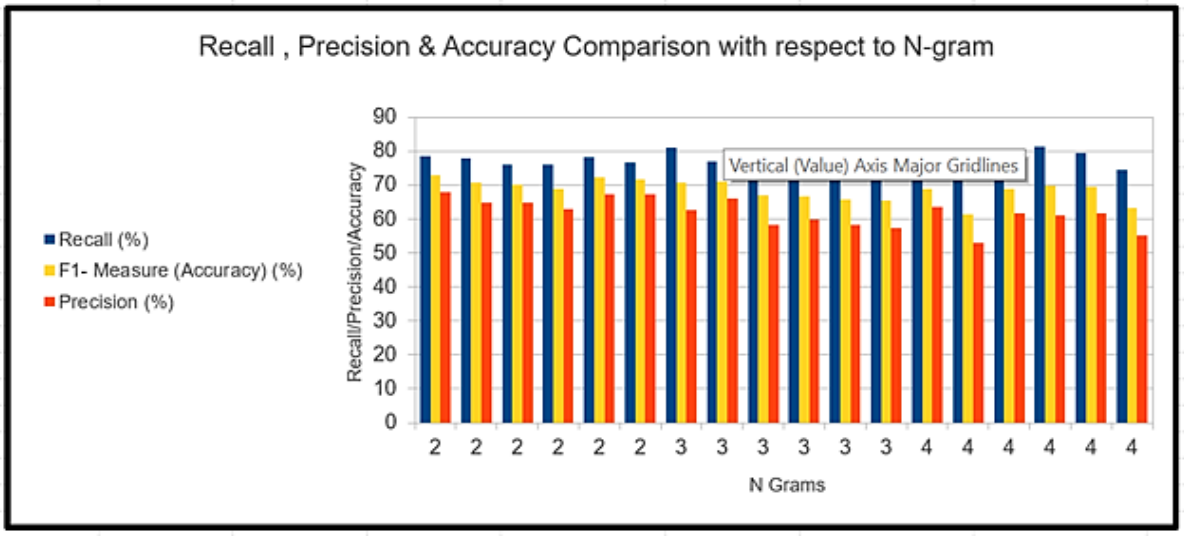
Selected Features

Word	Left Position	Right Position	Ambiguity	Noun	Adjective	Verb	Adverb	Pronoun	Helping Verb P...
First Word	0.5	0	1	1	0	0	0	0	0
Previous Word	0.75	0	1	1	0	0	0	0	0
Ambiguous W...	1	1	1	1	0	1	0	0	1
Next Word	0	0.75	0	0	0	1	0	0	1
Last Word	0	0.75	0	0	0	1	0	0	1

Figure 4 : Feature generated matrix

Sr.	N-Gram	Filters #	Testset Prediction Time (secs)	Average Prediction Time/Sentence (secs)	Recall (%)	Precision (%)	F1- Measure (Accuracy) (%)
1	2	4	17.604	0.083828571	78.57	67.77	72.77
2	2	8	17.686	0.084219048	77.84	64.93	70.8
3	2	12	18.259	0.086947619	76.11	64.93	70.08
4	2	16	18.74	0.089238095	76	63.03	68.91
5	2	20	18.408	0.087657143	78.02	67.3	72.26
6	2	24	18.856	0.089790476	76.76	67.3	71.72
7	3	4	6.827	0.032509524	80.98	62.56	70.59
8	3	8	2.609	0.01242381	76.8	65.88	70.92
9	3	12	2.759	0.013138095	78.85	58.29	67.03
10	3	16	2.814	0.0134	75.06	59.72	66.49
11	3	20	2.304	0.010971429	75.46	58.29	65.78
12	3	24	2.495	0.011880952	76.1	57.35	65.41
13	4	4	17.291	0.082338095	75.28	63.51	68.89
14	4	8	16.959	0.080757143	72.73	53.08	61.37
15	4	12	17.212	0.081961905	77.84	61.61	68.78
16	4	16	17.118	0.081514286	81.13	61.14	69.73
17	4	20	17.273	0.082252381	79.27	61.61	69.33
18	4	24	17.035	0.081119048	74.36	54.98	63.22

Fig 5:Accuracy with respect to n-gram model.



Graph 1:Recall ,Precision and accuracy

References

[1] M. Lesk,(1986) "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," Proceedings of SIGDOC

[2] Ekedahl, J., & Golub, K. (2004). "Word sense disambiguation using Wordnet and the Lesk algorithm. Projektarbeten "

[3] Moro, A., Raganato, A., &Navigli, R. (2014). Entity linking meets word sense disambiguation: a heuristicapproach. Transactions the Association for Computational Linguistics, 2, 231-244.

[4] Y. Heo, S. Kang and J. Seo, "Hybrid Sense Classification Method for Large-Scale Word Sense Disambiguation," in IEEE Access, vol. 8, pp. 27247-27256, 2020, doi: 10.1109/ACCESS.2020.2970436.

[5] Naseer, A., & Hussain, S. (2009). Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification. Center for Research in Urdu Language Processing, Lahore, Pakistan

[6] H. Seo, H. Chung, H. Rim,(2004) S. H. Myaeng and S. Kim, "Unsupervised word sense disambiguation using WordNet relatives," Computer Speech and Language, Vol. 18, No. 3, Pp. 253-273, 2004

[7] Reddy, S., Inumella, A., McCarthy, D., & Stevenson, M. (2010, July). IIITH: DoDomain-specificord sense disambiguation. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 387-391). Association for Computational Linguistics.

[8] Sharma, N., Kumar, S., & Niranjan, S. (2012). Using Machine Learning Algorithms for Word Sense Disambiguation: A Brief Survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume, 2.

[9] Trivedi, M., Sharma, S., &Deulkar, K. (2014). Approaches To Word Sense Disambiguation. International Journal of Engineering Research & Technology, 3(10), 645-647.

[10] P. Sachdeva, S. Verma and S. K. Singh,(2014) "An improved approach to word sense disambiguation," 2014 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Noida, 2014, pp. 000235-000240, doi: 10.1109/ISSPIT.2014.7300594.

[11] K. Samhith, S. A. Tilak and G. Panda,(2016) "Word sense disambiguation using WordNet Lexical Categories," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1664-1666, doi: 10.1109/SCOPES.2016.7955725.

[12] S. G. Kolte and S. G. Bhirud,(2008) "Word Sense Disambiguation Using WordNet Domains," 2008 First International Conference on Emerging Trends in Engineering and Technology, Nagpur, Maharashtra, 2008, pp. 1187-1191, doi: 10.1109/ICETET.2008.231.

[13] R. Liang, C. Luo, C. Zhang, T. Lei, H. Wang and M. Li,(2019) "Word Sense Disambiguation Based on Semantic Knowledge," 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT), Harbin, China, 2019, pp. 645-648, doi: 10.1109/ICEICT.2019.8846408.

[14] U. Farooq, T. P. Dhamala, A. Nongaillard, Y. Ouzrout and M. A. Qadir, (2015) "A word sense disambiguation method for feature level sentiment analysis," 9th International

Conference on Software, Knowledge, Information Management and Applications (SKIMA), Kathmandu, 2015, pp. 1-8, doi: 10.1109/SKIMA.2015.7399988.

[15] A. Guerrieri, F. Rahimian, S. Girdzijauskas and A. Montresor,(2016) "Tovel: Distributed Graph Clustering for Word Sense Disambiguation," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, pp. 623-630, doi: 10.1109/ICDMW.2016.0094.

[16] Samuel Sousa Evangelos Milios Faculty of Computer Science Dalhousie University Halifax, Canada eem@cs.dal.ca Institute of Science and Technology Federal University of São Paulo São José dos Campos, Brazil samuel.bruno@unifesp.br "Word sense disambiguation an evaluation study of semi-supervised approaches with word embeddings" 978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[17] Simone Conia Roberto Navigli Sapienza NLP Group Department of Computer Science Sapienza University of Rome "Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration"Proceedings of the 16th conference of the European chapter of the association for computational Linguatistic ,Pages 3269-3275 april 19-23 2021 Association for computational linguistic.