AMERICAN JOURNAL OF

EDUCATION AND TECHNOLOGY (AJET)

ISSN: 2832-9481 (ONLINE)

VOLUME 1 ISSUE 2 (2022)

Indexed in

PUBLISHED BY: E-PALLI, DELAWARE, USA

American Journal of Education and Technologies (AJET)

Volume 1 Issue 2, Year 2022
ISSN: 2832-9481 (Online)
DOI: https://doi.org/10.54536/ajet.v1i2.531
https://journals.e-palli.com/home/index.php/ajet

Sentence Level Amharic Word Sense Disambiguation

Dereje Senay Merawi[1*], Tesfa Tegegne Yalewu[2], Yitbarek Worku Tamir[1]

Article Information

Received: August 31, 2022

Accepted: September 17, 2022

Published: September 20, 2022

Keywords

Deep Learning, Natural Language Preprocessing, WordNet, Word Sense Disambiguation

ABSTRACT

Lexical ambiguity, phonological ambiguity, structural ambiguity, referential ambiguity, semantic ambiguity, and orthographic ambiguity were all types of Amharic ambiguity. The other ambiguities were out of this research because the study focuses on lexical-semantic, orthographic, and semantic ambiguities. Until now, some experts have been researching the Amharic word sense disambiguation system. Recent research, on the other hand, did not take into account antonym, troponymy, holonomy, and homonym relationships in the WordNet; this problem was overcome by this study. Using a Deep Learning method, we are developing an Amharic word sense disambiguation model. We use a design science research strategy to close the gap, starting with problem identification and concluding with final communication. 159 ambiguous words, 1214 synsets, and 2164 sentence datasets were used to create three distinct Deep Learning algorithms in three separate experiments. Using the given dataset, the overall performance of the model is measured using performance metrics in precision, F-measure, and confusion matrix. In this study, LSTM, CNN, and Bi-LSTM obtained 94 percent, 95 percent, and 96 percent accuracy respectively in the third experiment, based on performance measurement.

INTRODUCTION

Natural Language Preprocessing is an essential part of computer science, in which computational linguistics and machine learning are broadly used. The Machine learns the syntax and semantics of human language; processes and gives the output to the user (Heo *et al.*, 2020). (Pesaranghader *et al.*, 2018) WSD is a crucial problem in Natural Language Processing. However, the task is useful for applications such as machine translation and question answering. The problem defines as, a reasonable sense of a word that can have multiple meanings or seems to be ambiguous in a given context (Heo *et al.*, 2020).

Amharic is a Semitic language that is part of the Afro-Asiatic family and is spoken by the majority of Ethiopians. The Ethiopian Orthodox Tewahedo Church uses an alphabet called (fidäl) that was inherited from the Geez language. Its words are a linguistically ordered mixture of phonemes and their orthographic representations. It is like other Semitic languages and has a morphologically complex structure. Even though six types of ambiguity occurred in Amharic languages, this research focused on the lexical, orthographic, and semantic ambiguity of Amharic ambiguous words due to problems solved by the word sense disambiguation model. Native speakers of the language can disambiguate Amharic words or phrases easily. But the same task becomes highly complex and knotty when done by machines (Abid *et al.*, 2018). It is useful for many applications such as event extraction, document indexing, information retrieval, theme extraction, machine translation, semantic annotation, cross-document referencing, and semantic web search. This study dataset considers the following type of ambiguities, the word **ሃረግ**(phrase) is ambiguous and has three synsets, three glosses, and eight relations.

ሀረግ1(shrubs): **ከአጠገቡ በሚገኙ ዛፎችና ቋሚ ነገሮች ላይ እየተጠመጠመ እና እየተሳበ የሚያድግ እንደ ሃረግ ሬሳና፣ ጥሮ አይነት እጽዋት።** (It grows on twigs and shrubs and grows on trees and shrubs)

ሀረግ2(design): **ለጌጥ ሲባል በጥርብ እንጨት፣ በመጽሃፍትና ሐውልቶች ላይ እንደ ሃረግ እየተለጣጠፈ የሚሰራ ንድፍ።** (Design for decorative wood carvings, books and statues)

ሀረግ3(phrase): **ሁለትና ከሁለት በላይ ቃላቶቸን የያዘ ማሰሪያ አንቀጽ የሌለዉ የአረፍተ ነገር ከፋይ።** (A sentence without a binding clause containing two or more words). Its relation used as a dataset for this research was

Hypernym: **እጽዋት በዛፎች ላይ የሚጠመጠሙ ሃረጎችንም ይይዛሉ።** (Plants also contain phrases that are wrapped around trees)

Hyponym: **ሃረግ በዛፎች ላይ የሚጠመጠም የእጽዋት አይነት ማለትም ነው።** (shurbs is a type of plant that grows on trees)

From the above example hypernym, **እጽዋት**(plant) hyponym **ሃረግ**(shrubs)the family of shrubs plants. It indicates how many Amharic ambiguous words cause communication barriers because of various meanings based on different contexts. To infer senses, information-based systems use language tools such as dictionaries, thesauri, and knowledge graphs. On the other hand, supervised methods use an annotated training dataset to train a computer to predict a meaning given the target word and its context. To improve WSD efficiency, semi-supervised approaches to Word Sense Disambiguation combine manually generated training sets with a large corpus of unlabeled data.

To disambiguate a given sentence either with WordNet or without WordNet by contextual embedding. The mechanisms of WordNet development in manual, semi-automatic, or automatic approaches. Besides, this study was followed by a manual WordNet development

[1] Information Technology, Faculty of Technology, Debre Tabor University, Ethiopia
[2] Computer Science, Institute of Technology, Bahir Dar University, Ethiopia
[*] Corresponding author's e-mail: radereje6@gmail.com

strategy of hand-crafted lexical databases. In addition to knowledge-based word sense disambiguation models, Deep Learning approaches are also effective for disambiguating ambiguous words in different languages. LSTM is the one Deep Learning approach that is a gated type form of the recurrent neural network that is used in sequence modeling to efficiently capture long-term dependencies (Yang & Mitchell, 2019). The underlying methodology allows the model to duplicate its state between time steps while avoiding non-linearity constraints. Unlike the classic RNN model, which uses logistic functions to compute the gradients, the LSTM model uses multiplicative gates to better compute the gradients. (Zobaed *et al.*, 2021) Bi-directional LSTM (Bi-LSTM) is a form of LSTM in which the state at each time step combines the states of two LSTMs enunciated as forward and backward LSTM.

The following points are the contribution of this research.

A. We included antonym, troponym, holonym, and homonym relations or senses of ambiguous words not covered by the other researchers.

B. In this research, we prepared 2164 Amharic sentences and Amharic WordNet which has 1214 glosses in 159 ambiguous words.

C. We developed word2vec models that handle the vocabulary and context by CBOW within a size of 29.8 MB and 310,138 sentences within 478,797 vocabulary sizes as embedding layers.

LITERATURE REVIEW

(Wassie *et al.*, 2014) employed Amharic WSD by using a semi-supervised approach. They applied both unsupervised and supervised algorithms for clustering based on instances similarity and classification after the unlabeled data respectively. However, annotated data are costly and even limit the weight related to unsupervised.

The researchers (Assabie, 2014; Dureti, 2017) attempted to use the WSD system by using WordNet. The ambiguous word and its senses are located in a database and used as a knowledge base to disambiguate a given word. The WordNet hierarchy contains a maximum of three senses for a word and it is implemented on the Lesk algorithm. The system expected to disambiguate all open class words in a given sentence such as nouns, adjectives, adverbs, and verbs. Nevertheless, it disambiguates only one frequently occurring target word in the WordNet for an input sentence. In addition to that, the Lesk algorithm includes multiple words having multiple senses that were considered at once a problem during the execution of the system.

(Mieraf, 2019) attempted the Amharic hierarchical word sense disambiguation system by using WordNet at the sentence level in all classes of a word. To disambiguate Amharic sentences used context-to-gloss overlap and augmented semantic space approaches. The most popular algorithm extended by Lesk (Banerjee & Pedersen, 2002) was used for word sense disambiguation in Amharic language. One of the methods applied in this research was augmented semantic space is highly dependent on the number of words that exist in the WordNet since 17 words are only included in the WordNet. It also depends on several context words in the sentences which also exist in the WordNet.

The algorithm works by counting overlap between glosses which makes it dependent on the length of glosses, and the exact wording of the number of related synsets and glosses. The WordNet didn't consider antonyms, homonymy, troponym, and holonomy relation of the words having single sense up to three senses were the main limitation of this study.

METHODOLOGY

In this study, we would follow the design science research approach. Design science has six key steps. Those are problem identification, motivation, objectives of the solution, design and development, demonstration,

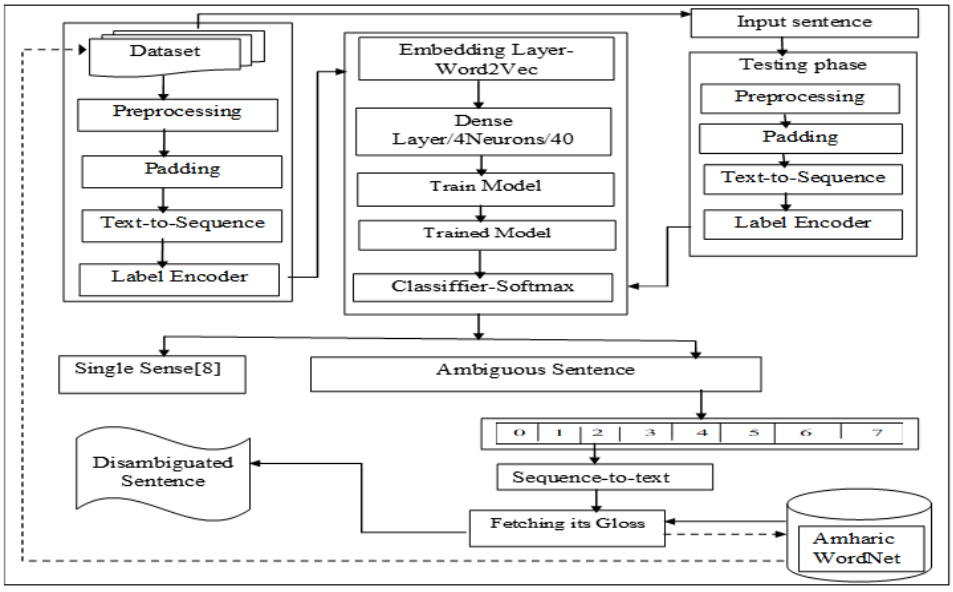

Figure 1: General Architecture of Amharic Word Sense disambiguation model

evaluation, and communication. This research would use these six design science steps to solve Amharic word sense disambiguation problem.

As presented in Figure 1 above describes the general architecture of the automatic Amharic word sense disambiguation model starts from collecting Amharic corpus from different sources. The model has three modules preprocessing module, model development modules, and fetching gloss (matching) with trained model weighting with WordNet vectors, if there is a matching vector then fetch its glosses and synsets from the WordNet. The research dataset started with finding ambiguous Amharic words from different sources: Amharic Sketch Engine, Amharic dictionary within its gloss or definition, Abissinica online dictionary, and then constructing sentences by using their definition for training and testing was the second task of the research. By using the ambiguous words within its definition and relation type prepared Amharic WordNet. And reviewed by the expert and take only the sentences selected by the reviewer as a dataset. The dataset used in this study was 159 ambiguous words and 1958 synsets with 1214 glosses to develop WordNet and 2164 sentences for training, testing, and validation of the Deep Learning model. The sentences used in testing were the parts of the dataset that are split from the total 2164 sentences dataset within a ratio of 0.8, 0.1, and 0.1 for training, testing, and validation purposes respectively.

1) Preprocessing: In this section, firstly the given corpus is tokenized, and raw text is converted into meaningful data using text preprocessing. Removing numbers and punctuation marks, even if different writing styles in Amharic script but follow uniform writing style (normalization), removing (suffix, prefix, infix) to control words that have the same stem or root but different structure.

2) Model Development: in this phase, we train LSTM, CNN, and Bi-LSTM Deep Learning language models and then develop the model and save it for offline use. Evaluating its classification accuracy would be the last task of this module by test dataset.

3) Fetching Glosses: If the predicated result of the model is ambiguous then the sentences back changed to text and match with its WordNet. If no matching occurred the word was recorded in the WordNet without gloss and the sentences were added to a dataset within special labeling. These sentences wait for experts for labeling and the word gloss for future training.

DATA SOURCE

The following table I describes the number of ambiguous Amharic words, number of synsets, number of glosses, sentence datasets for the research.

Table 1: Describes the number of ambiguous Amharic words, number of synsets, number of glosses, sentence datasets for the research

Number of Ambiguous Words	Number of Synsets	Number of Gloss	Sentences in all Relations	Source of Data
159	1350	460	2164	Amharic dictionary, sketch engines, Abyssinica, and experts

RESULT

The training and validation loss and accuracy of the Bi-LSTM model are presented in Figure 2 and Figure 3 respectively. In the first experiment, there was overfitting, The overfitting was handled by model regularization and balancing the dropout rate since it determines which neuron is idle and active at a time. The table shows the overall performance of the Bi-LSTM model on the Amharic word sense disambiguation problem. This experimental result indicates the total performance of the model based on true positive and true negative using a random search hyperparameter selection algorithm.

Lastly, the performance of Bi-LSTM for Amharic word sense disambiguation problem confusion matrix Table is

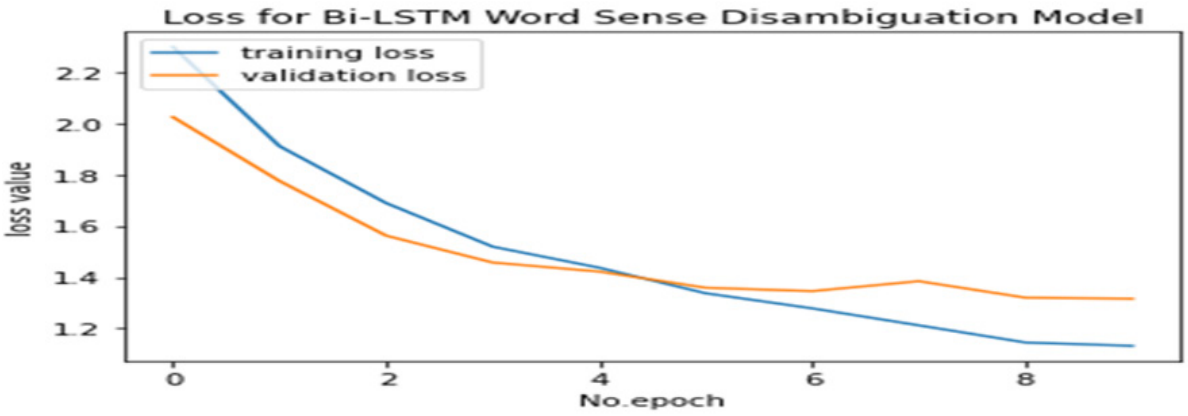

Figure 2: Training Loss and Validation Loss of Bi-LSTM model

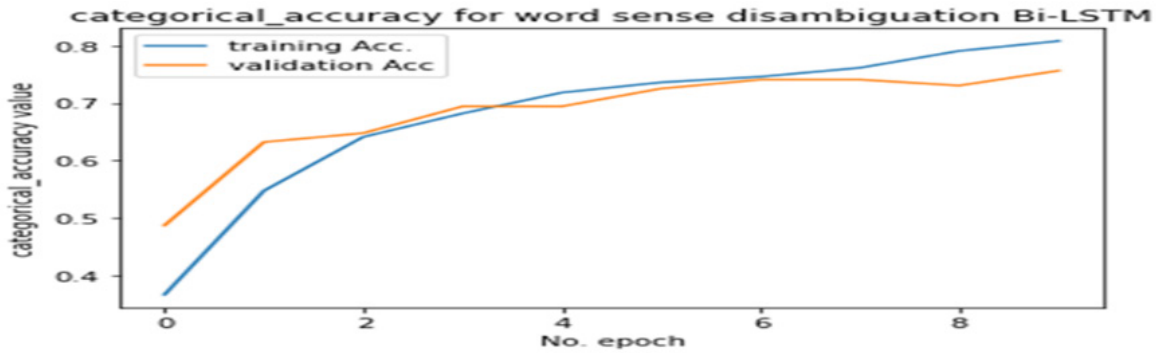

Figure 3: Training Data and Validation Data Accuracy of Bi-LSTM model

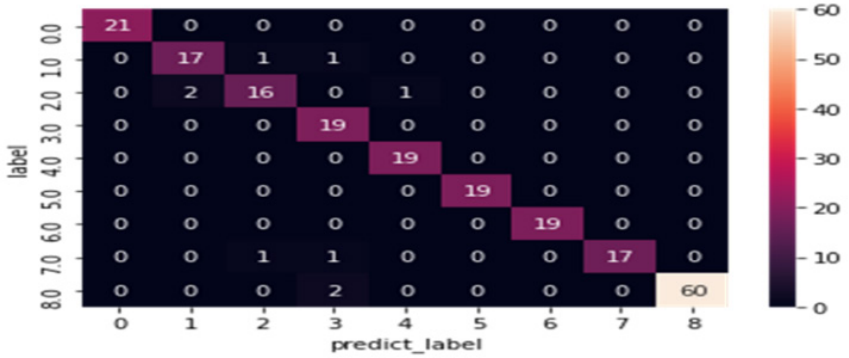

Figure 4: Confusion Matrix of Bi-LSTM model

shown in Figure 4.

To know the performance of the model, we made different error analysis mechanisms. Like mean absolute error (MAE), root mean square (MSE), and root mean square error (RMSE) from 216 test sentences we could calculate the error analysis of misclassified sentences the outperform Bi-LSTM model.

therefore,

$MSE=\sum^{n}_{(i=0)}(pi-ai)^2/9$.....(1)

$MSE=(21-21)^2+(17-19)^2+(16-19)^2+(19-19)^2+(19-19)^2+(19-19)^2+(19-19)^2+(19-17)^2+(62-60)^2$

$= (0+4+9+0+0+0+0+0+4+4)/9$

$= 21/9$

$= 2.33$

= 2.33 is our mean square error for 9 classes.

DISCUSSION

During experiment 1, all three Deep Learning algorithms were tested within the specified constant hyperparameter setup using padding, embedding, split ratio, and vocabulary size. Specifically, from the variable hyperparameter optimizer = SGD, activation = tanh, batch size = 32, random state = 0, number of neurons = 100. In this experiment, we were using the activation function but this activation function had a problem of vanishing gradient around -1 and 1 (Nguyen *et al.*, 2021). The vanishing gradient problem restricts the model to reach easily into a global minimum. The activation function problem influences the outcomes of this experiment since SGD has a significant variation in model parameters. To achieve the same convergence as gradient descent, the learning rate must be gradually reduced. LSTM (88%) and Bi-LSTM (91%) in particular performed poorly in this experiment when it came to CNN (92%).

In the second experiment rather than the hyperparameter employed in experiment 1 its variable hyperparameters were random state=0, Optimizer = Adam, activation = ReLU, batch size=16, epochs =10, dropout=0.2, number of neurons =60. The activation function in this experiment had a problem of dead states around 0 and 1. Still, the batch size in this experiment used was largely related to the dataset. That is why the model performance has gone to a minimum related to other experiments. The dropout rate also factors into our experiment. Because we chose a 0.2 dropout rate in our experiment, this is inconvenient for our problem. The activation function was set to 0 for all negative values and dead state problems. This experiment achieved 82%, 88%, and 89% accuracy of LSTM, CNN, and Bi-LSTM respectively. The experimental result shows low performance in all algorithms due to the above reasons. During the last experiment, the constant hyperparameters were specified in both experiments 1 and 2. But the variables were random state=42, optimizer = Adam, activation = GELU, regularize =l2(0.001), batch size=8, epochs =10, dropout=0.39, number of neurons =40. Adam optimizers used in this experiment have a very good learning rate rather to the others. The activation function used in this experiment was GELU (gaussian rectified linear unit) has the potential to learn within mini-batch data by solving the dead state problem in ReLU. The random state is used to regularize and minimize the overfitting problems of the graph from 0 to increase 42. The batch size specified in this experiment was also comfortable for the models

because the dataset was small. To increase the model performance we used model regularize, small batch size within 0.39 dropout rate in our experiment.

Besides in this experiment, relatively all algorithms achieved a good average result for Amharic word sense disambiguation problems like LSTM (94%), CNN (95%), Bi-LSTM (96%). The model accessing WordNet performance was measured by the accuracy of 216 test dataset sentences; thus, 86 sentences were correct. The remaining 19 sentences were miss classified; due to the problem of the stemming algorithm. The stemmer over stemmed the given test dataset sentence. Therefore, its test accuracy was by retrieving its gloss from the WordNet if its prediction was less than eight is 94.5%. The result of the experiment shows that Bi-LSTM achieved higher results than the other two algorithms. The performance of our Bi-LSTM model is better than the previous models (16%); however, the models and datasets were different, and the comparison was done due to problem similarity.

CONCLUSION

Word Sense Disambiguation means selecting the right sense based on the surrounding context. Even if some researchers did Amharic word sense disambiguation system, they didn't include the holonymy, antonym, homonomy, and troponymy relation of the Amharic word in their study. Those relations were covered by this research. This research was done under a Deep Learning approach rather than the previous methods and consider the remaining four relations of the word listed above. The research collected 159 ambiguous words, 1214 glosses, and 2164 sentences from different sources. And then we are designing WordNet based on their relation. Using the given dataset, the overall performance of the model is measured using performance metrics in precision, F-measure, and confusion matrix. Based on performance metrics Bi- LSTM achieved the state-of-the-art result for disambiguating Amharic ambiguous words. Lack of resources and datasets were the main challenges of this study. We recommended the following points for future study:

1. Designing automatic Amharic word sense disambiguation model which includes the two meronomy relations the substance of and entailment.

2. Designing an Amharic word sense disambiguation model that considers structural ambiguity, phonological ambiguity, and referential ambiguity.

REFERENCES

Abid, M., Habib, A., Ashraf, J., & Shahid, A. (2018). Urdu word sense disambiguation using machine learning approach. *Cluster Computing, 21*.

Assabie, M. A. and Y. (2014). Development of Amharic Morphological Analyzer Using Memory-Based Learning. 9th International Conference on NLP, *PolTAL 2014, 32*, 1–13.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. *International Conference on Intelligent Text Processing and Computational Linguistics,* 136–145.

Dureti, S. B. (2017). *A Generic Approach towards all Words Amharic Word Sense Disambiguation*. Addis Ababa University.

Heo, Y., Kang, S., & Seo, J. (2020). Hybrid sense classification method for large-scale word sense disambiguation. *IEEE Access, 8,* 27247-27256.

Mieraf, M. (2019). Word Sense Disambiguation for Amharic Sentences using WordNet Hierarchy. Bahir Dar, Ethioipia: Unpublished Master thesis, Bahir Dar University.

Nguyen, A., Pham, K., Ngo, D., Ngo, T., & Pham, L. (2021, August). An analysis of state-of-the-art activation functions for supervised deep neural network. *In 2021 International Conference on System Science and Engineering (ICSSE*) (pp. 215-220). IEEE.

Pesaranghader, A., Pesaranghader, A., & Sokolova, M. (2018). *One Single Deep Bidirectional LSTM Network for Word Sense Disambiguation of Text Data* (pp. 96–107).

Wassie, G., Ramesh, B. P., Teferra, S., & Meshesha, M. (2014). A Word Sense Disambiguation Model for Amharic Words using Semi-Supervised Learning Paradigm. *Science, Technology and Arts Research Journal, 3*(3), 147-155.

Yang, B., & Mitchell, T. (2019). Leveraging knowledge bases in lstms for improving machine reading. *ArXiv Preprint ArXiv:1902.09091*.

Zobaed, S., Haque, M. E., Rabby, M. F., & Salehi, M. A. (2021). Senspick: sense picking for word sense disambiguation. *2021 IEEE 15th International Conference on Semantic Computing (ICSC),* 318–324.