Amharic Sentence-Level Word Sense Disambiguation Using Transfer Learning

Neima Mossa[1(✉)] and Million Meshesha[2]

[1] Faculty of Computing, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia
neimamussa32@gmail.com
[2] School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Abstract. Word sense disambiguation (WSD) plays an important role, in increasing the performance of NLP applications such as information extraction, information retrieval, and machine translation. The manual disambiguation process by humans is tedious, prone to errors, and expensive. Recent research in Amharic WSD used mostly handcrafted rules. Such works do not help to learn different representations of the target word from data automatically. Moreover, such a manual disambiguation approach looks at a limited length of surrounding words from the sentence. The main drawback of previous works is that the sense of the word will not be detected from the synset list unless the word is explicitly mentioned. Our study explores and designs the Amharic WSD model by employing transformer-based contextual embeddings, namely AmRoBERTa. As there is no standard sense-tagged Amharic text dataset for the Amharic WSD task, we first compiled 800 ambiguous words. Furthermore, we collect more than 33k sentences that contain those ambiguous words. The 33k sentences are used to finetune our transformer based AmRoBERTa model**.** We conduct two types of annotation for our WSD experiments. First, using linguistic experts, we annotate 10k sentences for 7 types of word relations (synonymy, hyponymy, hypernymy, meronomy, holonomy, toponymy, and homonymy). For the WSD disambiguation experiment, we first choose 10 target words and annotate a total of 1000 sentences with their correct sense using the WebAnno annotation tool. For the classification task, the CNN, Bi-LSTM, and BERT-based classification models achieve an accuracy of 90%, 88%, and 93% respectively. For the WSD task, we have employed two experiments. When we use the masking technique of the pre-trained contextual embedding to find the correct sense, it attains 70% accuracy. However, when we use the FLAIR document embedding framework to embed the target sentences and glosses separately and compute the similarities, our model was able to achieve 71% accuracy to correctly disambiguate target words.

Keywords: Word sense disambiguation · Transfer learning · Neural network · Pre-trained language model · Natural language preprocessing · Morphological analyzer · Amharic WSD

B. H. Woldegiorgis et al. (Eds.): ICAST 2022, LNICST 455, pp. 227–238, 2023.
https://doi.org/10.1007/978-3-031-28725-1_14

1 Introduction

Natural language processing (NLP) is a field of artificial intelligence that assists computers in understanding, interpreting, and manipulating human language. Natural language is now being used to exchange information among humans and has now reached the extent of being an evolution criterion for technology (Reta 2015). To properly access and understand the information on the internet, there is a need for people all over the world to be able to use their language. This requires the existence of NLP applications such as machine translation, information retrieval, information extraction, and others. These downstream NLP applications rely on tools such as word sense disambiguation for their reasonable performance.

Most of the words in natural languages are polysemic, which means that they have several meanings (Hassen 2015). Amharic is one of the languages that have many words with multiple meanings. It is like other Semitic languages with a morphologically complex structure (Senay 2021). The ability to recognize the meaning of a word from its context and solve the ambiguity is one of the most difficult problems in natural language processing (Alian et al. 2016). Ambiguity is defined as a word, term, notation, sign, or symbol interpreted in more than one way (Mindaye et al. 2010). Word Sense Disambiguation is a hard and challenging task in NLP, intending to determine the exact sense of an ambiguous word in a particular context (Huang et al. 2019). When WSD is used in conjunction with other NLP approaches, it improves the efficiency of identifying accurate keywords for use as features in classification, searching, and many more NLP application (Senay 2021).

Knowledge-based, corpus-based, and hybrid machine learning methods are the main categories of approaches for WSD tasks (Pal and Saha 2015). Knowledge-based WSD approaches are based on different knowledge sources such as machine-readable dictionaries (WordNet), thesauri, etc. LESK, semantic similarity, selection preference, and heuristic are the main algorithms for knowledge-based approaches. There are two sets of data for training and testing in supervised approaches. This approach to WSD systems employs machine learning techniques based on manually created sense-annotated data. The training set, which consists of examples related to the target word, could be used to learn a classifier. The supervised approach includes techniques such as Naïve Bays, decision lists, and K-nearest neighbor algorithms. Unsupervised WSD methods do not rely on external knowledge sources, machine-readable dictionaries, or sense-annotated data sets, rather, they use the information found in un-annotated corpora to differentiate the word meaning.

Recently, contextual embedding methods like BERT, ELMO, and GPT-2/3 learn sequence-level semantics by considering the sequence of all the words in the input sentence (Chawla et al. 2019). These methods are characterized by their high performance, and the ability to extract a lot of information from raw text. These recent language models, especially the BERT model is trained to predict the masked word(s) of the input sentence (El-razzaz et al. 2021). To weigh, the relationship between each word in the input sentence and the other words in the same sentence, BERT learns self-attention by giving a vector for each word. The vector represents the relationship of one word with other words in the input sentences and is used to generate word embedding. In this work,

we have employed AmRoBERTa, a RoBERTa model trained for Amharic (Yimam et al. 2021).

2 Related Works

The research by Kassie (2009) tried to demonstrate WSD for Amharic language using semantic vector analysis. A total of 865 words were selected from the Ethiopian Amharic language legal statute documents. Instead of using sense-tagged words, the researcher evaluates WSD using pseudo-code words (artificial words). The developed algorithm outperformed the one used by Lucene, according to their comparison of the two. The achieved result is an average precision and recall of 58% and 82%, respectively. The author recommended developing resources such as Corpora, Thesaurus, and WordNet, that could be useful to advance the research in information retrieval, and word sense disambiguation.

Mekonnen (2010) conducted the Amharic WSD study using a corpus-based, supervised machine-learning approach. The author used the Naïve Bayes algorithm for Amharic WSD to classify a word to its correct sense using Weka 3.62 package in both the training and testing phases. A total of 1045 English sense examples for the five ambiguous words were gathered from the British National Corpus (BNC). The dictionary is used to translate the sense illustrations back into Amharic. For each sense of the ambiguous word, a total of 100 sentences were collected where the accuracy achieved ranged from 70% to 83.5% for all classifiers.

Assemu (2011) tried to develop corpus-based Amharic WSD through the use of unsupervised machine learning. A total of 1045 English sense examples for the five ambiguous words were gathered from the British National Corpus (BNC). Using the Amharic-English dictionary, the sense examples were converted to Amharic and prepared for experimentation. The result showed that the accuracy of unsupervised Amharic WSD is state-of-the-art result than the supervised machine learning approach, with an accuracy of 83.2% and 70.1%, respectively. For better Amharic WSD, the researcher recommended using linguistic tools like the Thesaurus, Lexicon from WordNet, machine-readable dictionaries, and machine translation tools.

Wassie (2014) utilized a semi-supervised learning strategy, and present a WSD prototype model for Amharic words. Unsupervised machine learning approach for clustering based on instance similarity and supervised machine learning approach after unlabeled data are applied. To cover all the senses of each target word available, annotated corpora are highly insufficient. The development of the Adaboost Bagging and ADtree algorithms perform at 84.90%, 81.25%, and 88.45%, respectively. The author concludes that Semi-supervised learning using bootstrapping algorithm performs better.

The research by Hassen (2015) developed an Amharic WSD knowledge-based approach based on WordNet to extract knowledge from word definitions and relationships between words and senses. They manually created the Amharic WordNet for this study and chose 2000 words, including ambiguous words. They carried out two tests to compare Amharic WordNet's impact with and without a morphological analyzer, and the results showed an accuracy of 57.5% and 80%, respectively. A two-word window on either side of the ambiguous word is sufficient for Amharic WSD, according to their

research into the optimal window size. In this experiment, they have concluded that Amharic WordNet with a morphological analyzer can have better accuracy than without a morphological analyzer. They recommended automatic the development of Amharic WordNet and to apply a hybrid approach.

Tesema, Tesfaye and Kibebew (2016) applied supervised machine learning techniques to a corpus of Afaan Oromo language to automatically gather disambiguation information. This method is known as a corpus-based approach to disambiguation. To determine the prior probability and likelihood ratio of the sense in the provided context, they have utilized the Naïve Bayes approach. A total of 1240 Afaan Oromo sense examples were gathered for the chosen five ambiguous words, and the sense examples were manually tagged with their appropriate senses. The author used a corpus of Afaan Oromo sentences based on the five selected ambiguous words to acquire disambiguation information automatically. The system attains an accuracy of 79%, and it was discovered that the Afaan Oromo WSD can handle four words on either side of an ambiguous target word.

Siraj (2017) attempts to develop a system for word WSD that uses data from WordNet and tagged example sentences to determine the sense of ambiguous Amharic words. Information from WordNet was extracted using the LESK algorithm and Python programming. The WordNet is made up of 17 ambiguous words from various classes, along with developed synonyms and glossary definitions. Based solely on the Jaccard Coefficient and Cosine Similarity, Amharic WSD's accuracy performance reached 84.52% percent and 85.96%, respectively. The average accuracy of the Jaccard Coefficient with Lesk scores is 89.83% which is a better result, compared to cosine similarity with LESK (86.69%). The researcher suggests for future work to use the Adaptive LESK algorithm and improve the performance of the WSD system.

Mulugeta (2019) attempts to develop an Amharic WSD system that uses Amharic WordNet hierarchy as a knowledge base. They use context to gloss overlap augmented semantic space approach. Most previous research on Amharic WSD focused on verb class; yet, Mulugeta (2019) tried to solve all open classes (verb, noun, adverb, and adjective) by developing WordNet. The WordNet contains about 250 synsets and does not include all relationships for single-sense words in the WordNet. The main challenge in this study was the unavailability of lexicon resources (WordNet), and the stemmer algorithm used in the preprocessing does not cover all exceptions and has limitations in returning the root word. Experimental result shows that context-to-gloss followed by augmented semantic space has achieved the highest recall of 87% and 79% for three target words at word and sentence level respectively. And the highest average accuracy of 80% and 75% at word-level and sentence level are achieved by this approach. Their recommendation is to develop a better stemmer or morphological analyzer and fully constructed WordNet containing relationships for non-ambiguous words.

Tadesse (2021) proposed a machine learning based WSD model for the Wolaita language. A total of 2797 sense instances were gathered to complete the investigation. Language specialists assessed the acquired data before creating five datasets for five ambiguous words, including "Doona," "Ayfiya," "Aadhdha," "Naaga," and "Ogiya." They used quantitative and experimental research to discover the ideal machine combination algorithms for learning and methods for extracting features. AdaBoost classifier

utilizing BOW, TF-IDF, and Wor2Vec features as an extraction approach and the Support Vector Classifier, Bagging, Random Forest Classifier, and AdaBoost as classifier for the five datasets. In this study, precision and recall were used as the primary metrics for evaluation. Support Vector Classifier and Bagging classifiers with TF-IDF obtain an accuracy of 83.22% and 82.82%, respectively.

Recently, Senay (2021) has developed Amharic WSD by using a deep-learning approach. A total of 159 ambiguous words, 1214 synsets, and 2164 sentence datasets were used to create three distinct deep learning algorithms in three separate experiments. As a methodology, they used a design science research strategy. The author used different deep learning models for classification such as LSTM, CNN, and Bi-LSTM that are trained on the dataset using different hyperparameters. The results showed that LSTM, CNN, and Bi-LSTM obtained 94%, 95%, and 96% accuracy during the third experiment, respectively. But for disambiguation, they used handcrafted rules without applying any model. To increase the performance of the model, using lemmatization in the preprocessing, and using an attention mechanism are recommended.

Generally, Amharic word sense disambiguation was done by different researchers using different machine learning approaches. However, there is no easy and automatic Amharic word sense disambiguation, and there is no research that used the transfer learning algorithm for the disambiguation purpose. Generally, most of the literature tries to develop Amharic WSD but there is a gap in solving the problems of word sense. Most of them follow a manual approach for extracting word sense. Recent research used handcrafted rules or directly fetching the meaning of an ambiguous word from the synset list or in the WordNet but did not learn different representations from data automatically. The WSD developed by researchers requires manually labeled sense examples for every word sense. Previous researches also require defining features explicitly; but transfer learning algorithms aim to learn different representations from data automatically (Bouhriz et al. 2016); solve ambiguity problem based on sentence semantics. In this research, we attempt to employ transfer learning for Amharic WSD.

3 Amharic Language

Amharic is one of the northern Semitic languages in the part of the Afro-Asiatic families and it becomes a countless contribution in the area of literature in the 17th century up to the 19th century (Kebede at el. 1993). After Arabic, Amharic (አማርኛ) is the second most broadly spoken Semitic language (Gezmu et al., 2019). In addition, the language has a significant number of speakers in all regional states of the country (Salawu and Aseres 2015) and also in Canada, the USA, Eritrea, and Sweden (Mulugeta 2019).

3.1 Amharic Writing System

The Amharic language has its own alphabet, known as ፊደል/fidäl, which was inherited from the Geez. ፊደል/Fidäl is a syllabary writing system in which the consonants and vowels coexist within each graphic symbol. Unlike most Semitic scripts such as Arabic and Hebrew, Amharic fidäl is written from left to right. The writing system consists of 231 core characters, 33 consonants, each of which has 7 orders depending on the vowel

with which it is combined, and some additional orders of ‘ፊደል’/ fidäl are called dikala hoheyat/ ዲቃላ ሆሄያት(Getaneh 2020).

To separate each word and sentence in a formal Amharic writing system, the main punctuation marks are discussed as follow. The Ethiopic comma (፡) to separate words, Ethiopic full stop (።) to end the sentence, Ethiopic semicolon (፤) to separate Amharic words or phrases with similar concepts, the Ethiopic double dash (፧) to separate Amharic sentences with a similar concept and Ethiopic question mark (?) to end the question are the main unique Ethiopic punctuation marks. Nowadays, the Ethiopian modern writing system uses a single space rather than an Ethiopic comma (፡) to separate words.

3.2 Ambiguity in Amharic Language

Different scholars define ambiguity in a different way. According to Mindaye et al. (2010), ambiguity is described as the attribute of being ambiguous, where a word, term, notation, sign, symbol, phrase, sentence, or any other form used for communication is deemed ambiguous if it can be understood in more than one manner. Amare (2001) also define ambiguity as the quality of any thought, idea, statement, or claim whose meaning, intention, or interpretation cannot be determined decisively by a set of rules or processes.

Based on the study of Amare (2001), there are six types of ambiguities in Amharic language, namely Lexical Ambiguity, phonological ambiguity, structural ambiguity, referential ambiguity, semantic ambiguity, and orthographic ambiguity. These ambiguities are summarized below.

Lexical Ambiguity: Lexical ambiguity occurs when a lexical unit falls into separate part-of-speech categories with different senses, or when a lexical unit has more than one sense, all of which fall into the same part-of-speech category (Abate and Menzel 2007).
Phonological Ambiguity: The placement of pause within the word may lead to phonological ambiguity. When speakers use pauses and without pauses during speaking leads to ambiguity (multiple meanings) of a word (Kassie 2009, Mekonnen 2010).
Semantic Ambiguity: It determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. Polysemy, idiomatic and metaphorical word relations in a sentence are causes of semantic Ambiguity (Siraj 2017, Hassen 2015).
Syntactic Ambiguity: Structural ambiguity can give more than one meaning by the order of the word and holds more than one possible position or arrangement in the grammatical structure of the sentence.
Orthographic Ambiguity: Geminate and non-geminate sounds are causes of orthographic Ambiguity. This type of ambiguity can be solved using the context meaning of the sentence (Kassie 2009, Assemu 2011).
Referential Ambiguity: This ambiguity arises when a pronoun stands for more than one possible antecedent. a pronoun is understood by default even if it is not written grammatically.

4 Methodology

Algorithm: For this research we compared three models CNN, BiLSTM, and BERT to classify weather the word is ambiguous or not. Our experimental result showed that

BERT has better result than CNN and BILSTM because BERT used self-attention-based transformer architecture, which, in combination with a masked language modeling target, allows to train the model to see all left and right contexts of a target word at the same time (Chawla et al. 2019). After identifying whether the word is ambiguous or not the next task is assigning the meaning of ambiguous word. So, to disambiguate the ambiguous word we apply the AmRoBERTa model with the flair document embedding technique. It is a recent transfer learning approach that gives better performance in the available datasets (Yimam et al. 2021).

Dataset Collection and Preparation. Since there are no labeled datasets available for Amharic word sense disambiguation, the main task for this thesis work is to prepare labeled datasets for WSD. We have collected 10k sentences and 800 ambiguous words from Amharic news, Amharic dictionary, Amharic Quran, Amharic Bible, Abissinica online dictionary and Amharic textbooks (from grade 7–12). A total of 33,297 sentences are used to finetune the AmRoBERTa model (transfer learning). The collected data passes through data preprocessing to prepare the data for experimentation. Data preprocessing is critical for improving the performance of the model. To make our data more suitable for the experiment, we use various data preprocessing techniques such as tokenization, stopword removal, special character removal, normalization, and morphological analysis.

Dataset annotation: In our study, we selected annotators to keep the nature and behaviors of Amharic language texts and to acquire quality and reliable data. We annotate both relationship of the sentence and the sense of the word in the sentence. For the dataset annotation, we have done two different annotations. The first annotation is to know whether the data set contains all the selected relationships of a word or not. Therefore, we selected three Amharic language and linguistic experts to annotate the data. The experts annotated the relationship between the sentences.

The second annotation is for disambiguation or to know the sense of the word. For this task, we have also used the WebAnno annotation tool to annotate the ambiguous word in the sentence. We selected two annotators and one curator from Amharic language native speakers. The main advantage of the WebAnno annotation tool is getting the value for inter annotation agreement (such as Fleiss kappa, and Cohen's kappa) is easy. We used Cohen's kappa as a measure of inter-annotator agreement.

5 Result and Discussion

5.1 Experimental Result of CNN Model

We have trained the CNN model with 2 dense layers with sigmoid activation functions and binary_crossentropy loss functions We also used 0.00001 for the learning rate, 64 batch-size, and a dropout rate of 0.2, which are optimal for our experiment (Fig. 1).

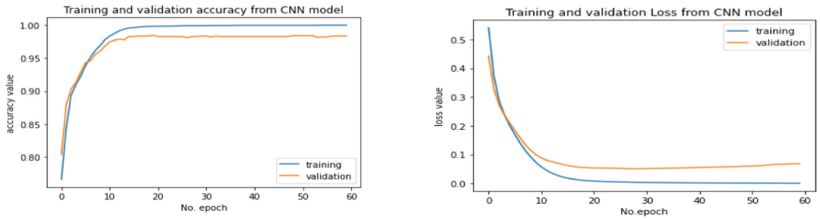

Fig. 1. Training and validation accuracy and loss graph for CNN model.

5.2 Experimental Result of BiLSTM Model

Experimental results of the Bi-LSTM model were analyzed and interpreted. We have trained the Bi-LSTM model with 2 dense layers with sigmoid activation functions and binary_crossentropy loss functions. We employed 64 neurons in the first dense, for a total of 128 neurons in both the forward and backward directions. We used, the maximum dropout rate of 0.2, the training epoch value of the model is 60, the learning rate that changes the weight of the training algorithm and we set the value of 0.00001. We set the batch-size to 64 (Fig. 2).

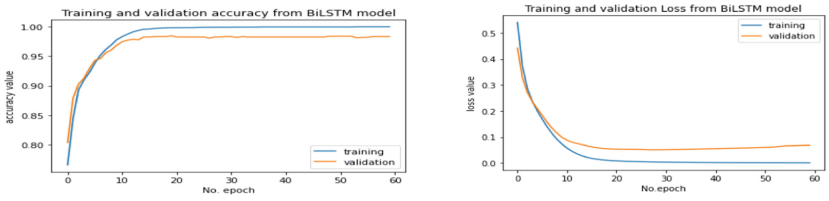

Fig. 2. Training and validation accuracy and loss graph from BiLSTM model

5.3 Experimental Result of BERT Model

We have used 60 epochs to train the model with a 0.00001 learning rate. To reduce overfitting, we set the dropout rate to 0.2. We have also used the Adam optimizer, RELU for the hidden layer, and Sigmond for the output layer is used as an activation function. To build the model we have used three dense layers, for the first dense we have used 64 neurons and a 0.2 dropout-rate. For the second dense layer we used 32 neurons. Lastly for the output layer we have used 2 neurons (Fig. 3).

For this research we select BERT for classification because BERT is better than both CNN and BiLSTM algorithms for semantic understanding.

5.4 Experimental Result of Disambiguation Model

We In our research, we have used the finetuned AmRoBERTa model with the FLAIR document embedding technique to disambiguate Amharic words in the given sentence.

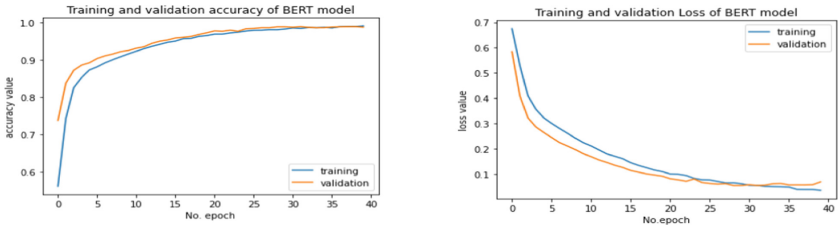

Fig. 3. Training and validation accuracy and loss graph from BERT model

AmRoBERTa fine-tuning: We fine-tuned the AmRoBERTa model using 33,297 sentences and 800 ambiguous words. When we train the model, we have used a maximum of eight contextual meanings for a single ambiguous word. Our experiment is conducted using an epoch of 200 and a batch_size of 64 using an NVIDIA GeForce RTX 1080/2080 Ti generations of GPU server, where each GPU has 12GB memory, with 32 CPU cores and 252 RAM to run our experiments. We have conducted our experiment with 100 and 150 epochs but the performance was not optimal. We set it to 200 epochs which is the optimal iteration for our data set. We have also experimented with batch-size of 32, 64, and 248. But we have selected batch-size 64 as the optimal batch size because when the batch-size is below 64 it takes more training time. When the batch-size is more than 64, there is faster training, but the performance is low.

AmRoBERTA with masking: AmRoBERTa model handles the context through masked language modeling by randomly masking the 15% of the sentence in each epoch of iteration. With a proper finetuning, our assumption is that, if we mask the ambiguous word, it should predict the correct word with the right sense. From the experiment, we take the following sentence predictions as an example.

Example: በአስተያየቱ ልክ ለለውጥ መትጋት ከምንም በላይ መሰረታዊ ነጥብ ነው ።From this sentence the ambiguous word ልክ (lik) is disambiguated as follow.

```
In [93]: sentt='በአስተያየቱ <mask> ለለውጥ መትጋት ከምንም በላይ መሰረታዊ ነጥብ ነው'
         predictions = fill_mask(sent)
         #print(prediction)
         for i in range(5):
             print(predictions[i]['token_str'])
```

```
ልክ
መጠን
ደረጃ
ትክክል
ሰርአት
```

Based on our experimental result, The sentence “ በአስተያየቱ ልክ ለለውጥ መትጋት ከምንም በላይመሰረታዊ ነጥብ ነው ።” the model masks the ambiguous word ልክ(lk) then the top 4 meaning of the masked word are predicted.

Word Sense Disambiguation with Flair embedding technique: For this experiment, we have used the finetuned pre-trained contextual model to disambiguate the correct sense of the ambiguous words. We have used the fine-tuned **AmRoBERTa** model with

the FLAIR document embedding technique. For the disambiguation task, we have followed a similar approach as Huang et al. (2019), where we have to prepare the target sentence and gloss sentence pairs. However, there is no WordNet for Amharic to employ for this task. Hence, we have selected 10 words that are previously annotated using the WebAnno annotation tool. These words are ዋና(Wana), መንገድ(Menged), ሳለ(Sale), አካል(Akal), ዋጋ(Waga), ገና(Gena), ቀና(Qena), ሀቅ(haq), ሃይል(Hayil), and ልክ(Lik). Then we constructed a gloss for 10 words, which contains the ambiguous word and possible senses with examples sentences. During disambiguation, we select a target sentence that contains ambiguous words where the sense is already annotated by the annotators. We use the FLAIR document embedding with the finetuned contextual pre-trained model to compute the similarity between the target sentence and the glosses. The sense which has a high similarity value with the target sentence would be the correct meaning of the ambiguous word. Based on the given sentence in the gloss, the model disambiguates the target word into its correct sense. Example 1 below shows a target sentence and glosses as examples.

Example 1: The sentence: "እያንዳንዱ ዋና ሃሳብ አንድ አንቀፅ ውስጥ ሰፋሯል።" is disambiguate as follow.

```
Target sentence: እያንዳንዱ ዋና ሐሳብ ራሱን በቻለ አንድ አንቀጽ ውስጥ ሰፍሯል ።

ጭብጥ: የአብሰባውን ዋና ሃሳብ አስረዳ ። ዋናውን ነገር ብቻ ንገሪኝ 0.5702
ዐይነተኛ: ተወካይ ሀኖም የተለየ ልብስ ለብሰ ዋና አጋፋሪ ሆና ድግሱን ሲታሞቀው ውለለች ። የቡና ዋና ጥቅም ሚቃቃት ነው ። አቶራጩን መንገድ ትዶ በዋናው መንገድ ማጣ ። 0.5347
መሪ ፣ ሀላፊ ፣ : አለቃ የሆቴሉ ዋና ሀላፊ ጥራለኝ ። ዋና የሌለው ጦር ቶሎ ይሸነፋል ። 0.3580
የሰሕር ላይ ስፖርት: ይሁን እንጂ ውሃ ዋና በአንድ ዘመን ብልጭ ብሎ የጠፋ ስፖርት ሆኗል ። ዋና መዋኘት ስለቻለ ከባህሩ ወጣ ። ከብት የቁም ዋና ስለሚዋኝ ውሀ አይበላውም እያሉ የኛ አገር ሰ
ዎች ሲጫወቱ ብዙ ጊዜ ሰምቻለሁ ። 0.3449
```

Based on the result of our experiment, for the target sentence" እያንዳንዱ ዋናሐሳብ ራሱን በቻለ አንድ አንቀፅ ውስጥ ሰፍሯል ።" the correct sense of the ambiguous word ዋና(**wana**) is ጭብጥ(**Chibt - main point**), as it has higher similarity with the target sentence (0.5702) compared to the other senses, which are አይነተኛ(Aynetegna - principal) and መሪ/ሀላፊ(Meri/Halafi -leader) with similarity scores of 0.5347 and 0.3580 respectively.

Example 2: The sentence: "ረብሻው ከተረጋጋ በኋላ የተወሰኑ ታክሲዎች መንገድ ላይመታየት በመጀመራቸው ህዝቡ ተደሰተ" is also disambiguated as follow.

```
Target sentence: ረብሻው ከተረጋጋ በኋላ የተወሰኑ ታክሲዎች መንገድ ላይ መታየት በመጀመራቸው ህዝቡ ተደሰተ

ጎዳና: በአንድ ወቅት መንገድ ጻር ለተሰበሰቡ ሰዎች ፊልሙን ማሳየት ጀመርን ። ከዚህ ውሳኔ በኋላ የተወሰኑ ታክሲዎች መንገድ ላይ መታየት ጀምረዋል ። መንገዱ ስለሚያስፈራ ሸኘኝ ። 0.8098
አካሄድ: ትክክለኛና ቀትየቁ ጥያቄ ግን በሰላማዊ መንገድ መቅረብ ነበረበት ። ሰው በሁለት መንገድ ከተፋቱ ይማራል ። በትክክለኛው መንገድ መንገድ አለብህ ። እስኪ በትክክለኛው መንገድ ቁትሩን ደምረ
ው ። 0.4688
አስተሳሰብ: ከሆነ ይህንን መጽሔት ከሰጡህ ሰዎች ጋር የምትመሳሰልበት መንገድ አለ ። በምን መንገድ ልታስረዳኝ ትችላለህ ። 0.4052
ብልሀት ፣ ዘዴ: ይህን ለማድረግ የሚያስችለው አንደኛው መንገድ የስብስቴ ሥራችን ነው ። አራተኛው የውጭ ምንዛሪ ማግኛ መንገድ የመንግሥት ታቀላ የሚባለው ነው ። ኢየሱስ ከርስቶስ እውነተኛ ነፃ
ነት ማግኘት የሚቻልበትን ቀላል መንገድ ጠቁምናል ። 0.3305
አሰራር: በሰላማዊ መንገድ ነፃና ዴሞክራሲያዊ ምርጫ መኖር አለበት ። 0.3136
ሁኔታ: በዚህ መንገድ የአምላክን ሕግ መጣሳችው ኃጢአተኞች እንዲሆኑ አደረጋቸው ። የዘካርያስ የመናገር ችሎታ በተአምራዊ መንገድ ተመለሰለት ። 0.2925
```

Based on the result of our experiment, for the target sentence “ረብሻው ከተረጋጋበኋላ የተወሰኑ ታክሲዎች መንገድ ላይ መታየት በመጀመራቸው ህዝቡ ተደሰተ።” the correct sense of the ambiguous word መንገድis ጎዳና(**Godana – street**), as it has higher similarity with the target sentence (0.8098) compared to the other senses, which are አካሄድ(Akahiad - approach), አስተሳሰብ(Astesaseb - thinking), አሰራር(Aserar - procedure), and ሁኔታ(Huneta - situation) with similarity scores of 0.4688, 0.4052, 0.3305,0.3136 and 0.2925 respectively.

6 Conclusion

This study has developed an Amharic word sense disambiguation model by using a transfer learning approach. The process of identifying the correct meaning based on its context is known as word sense disambiguation. WSD is improving the performance of different NLP applications like machine translation so, to advance NLP research WSD is important. In addition, WSD will be abasis to build Amharic WordNet. These issues motivated us to conduct this research.

As far as we know, there is no standard sense-tagged Amharic text dataset for Amharic WSD task. So, we have collected 10k sentences from Amharic news, Amharic dictionary, Amharic Quran, Amharic bible, and Amharic textbooks. For the Amharic WSD task, we have collected 800 ambiguous words from different sources such as Amharic dictionaries, Amharic textbooks, and Abissinica online dictionary. A total of 33,297 sentences are used to finetune the AmRoBERTa model for the transfer learning.

In our study, we have compared different models to select the most suitable model for WSD classification. To select the best fit model, we have conducted different experiments. For the classification task, we have experimented with CNN, BiLSTM, and BERT algorithms with 2 dense layers and a sigmoid activation function. According to the results, CNN, Bi-LSTM, and BERT obtained 90%, 88%, and 93% accuracy respectively. Based on our findings, the model based on BERT has achieved the vesting result.

As AmRoBERTa is a general-purpose pre-trained language model, we have fine-tuned it with 33,294 sentences and 800 ambiguous words. Finally, the AmRoBERTa model has been applied and when we use the masking technique to find the correct sense, it attains 70% accuracy. We have also employed the FLAIR document embedding framework to embed the target sentences and glosses separately. We then compute the similarity of the target sentence with the glosses embedding. The gloss with the higher score disambiguates the target sentence. Our model was able to achieve an accuracy score of 71%.

References

Abate, S.T., Menzel, W.: Syllable-based speech recognition for Amharic. In: Proceedings Of the 5th Workshop On Important Unresolved Matters, Pages Prague, Czech Republic, pp. 33–40 (2007). https://doi.org/10.3115/1654576.1654583

Alian, M., Awajan, A., Al-Kouz, A.: Arabic word sense disambiguation using Wikipedia. Researchgate **12**(1), 61–66 (2016). https://doi.org/10.21700/Ijcis.2016.108

Assemu, S.: Unsupervised machine learning approachfor word sense disambiguation to Amharic words. Addis Ababa, Ethiopia. Masters thesis Addis Ababa University, Ethiopia (2011)

Bouhriz, N., Habib, E., Lahmar, B.: Word sense disambiguation approach for Arabic text. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(4), 381–385 (2016)
Chawla, A., Biemann, C., Wiedemann, G., Remus, S.: Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. Arxive (2019)
El-Razzaz, M., Fakhr, M.W., Maghraby, F.A.: Arabic gloss WSD using BERT. Appl. Sci. **11**(6), 2567 (2021)
Getaneh, M.: Amharic wordnet construction using word embedding. Masters thesis Addis Ababa University, Addis Ababa, Ethiopia (2020)
Gezmu, A.M., Nürnberger, A., Seyoum, B.E.: Portable spelling corrector for a less-resourced language: Amharic. In: LREC 2018 - 11th International Conference on Language Resources and Evaluation, pp. 4127–4132 (2019)
Hassen, S.: Word sense dismbiguation using WordNet. Addis Ababa, Ethiopia. Masters thesis Addis Ababa University, Ethiopia (2015)
Huang, L., Sun, C., Qiu, X.: GlossBERT: BERT for word sense disambiguation with gloss knowledge, pp. 3509–3514 (2019)
Kassie, T.: Word sense disambiguation for Amharic text retrival: a case study legal documents. Addis Ababa, Ethiopia. Masters thesis Addis Ababa University, Ethiopia (2009)
Mekonnen, S.: Word sense disambiguation for Amharic text: a machine learning approach. Addis Ababa, Ethiopia. Masters thesis Addis Ababa University, Ethiopia (2010)
Mindaye, T., Sahlemariam, M., Kassie, T.: The need for Amharic WordNet. In: Global WordNet Conference, GWC 2010 (2010)
Mulugeta, M.: Word sense disambiguation for Amharic sentences using wordNet hierarchy. Masters thesis, Bahir Dar University, Bahir Dar, Ethiopia (2019)
Pal, A.R., Saha, D.: Word sense disambiguation: a survey. Researchgate **5**(3), 1–16 (2015)
Reta, B.: Application of parts-of-speech tagged corpus to improve the performance of word sense disambiguation: The Case of Amharic. Masters thesis Addis Ababa University, Addis Ababa, Ethiopia (2015)
Salawu, A., Aseres, A.: Language policy, ideologies, power and the Ethiopian media. University of South Africa (Tunisia) **41**(1), 71–89 (2015). https://doi.org/10.1080/02500167.2015.1018288
Senay, D.: Automatic Amharic word sense disambiguation model at sentence level by deep learning approach. Masters thesis Bahir Dar University, Bahir Dar, Ethiopia (2021)
Siraj, D.: A generic approach towards all words Amharic word sense disambiguation. Masters thesis Adis Ababa Univrtsity, Adis Ababa, Ethiopia (2017)
Tadesse: word sense disambiguation for wolaita language using machine learning approach. Masters thesis Adama University, Adama Ethiopia (2021)
Tesema, W., Tesfaye, D., Kibebew, T.: Towards The sense disambiguation of Afan Oromo words using hybrid approach (Unsupervised Machine Learning And Rule Based). Ethiop. J. Educ. Sci. **12**(1), 61–77 (2016)
Wassie, G., Ramesh, B., Teferra, S., Meshesha, M.: A word sense disambiguation model for Amharic words using semi-supervised learning paradigm. Researchgate **3**(3), 147 (2014). https://doi.org/10.4314/Star.V3i3.25
Yimam, S.M., Ayele, A.A., Venkatesh, G., Gashaw, I., Biemann, C.: Introducing various semantic models for Amharic: experimentation and evaluation with multiple tasks and datasets. Future Internet **13**(11), 275 (2021). https://doi.org/10.3390/Fi13110275
Kebede, H., Tsgie, F., Alemu, F., Azene, M.: Ethiopian Languages Research Center. Addis Ababa, Artistic Publication. Addis Ababa University, Ethiopia, Amharic Dictionary (1993)