Transfer Learning and Word Sense Disambiguation for Low-resource Languages, the Case of Amharic

Neima Mossa Ahmed  
Bahir Dar University,  
Bahir Dar, Ethiopia  
neimamussa32@gmail.com

Million Meshesha  
Addis Ababa university,  
Addis Ababa, Ethiopia  
meshe84@yahoo.com

Abstract

Word sense disambiguation plays an important role, in increasing the performance of NLP applications such as information retrieval, question answering, and machine translation. The manual disambiguation process by humans is tedious, prone to errors, expensive, and time-consuming. Low-resource languages such as Amharic do not have WordNet which makes the WSD task challenging. Previous works try to identify and develop Amharic word sense disambiguation using different approaches but still, there are no easy and automatic WSD tools. For this study, we collect the WSD dataset and employ contextual embedding, namely AmRoBERTa to automatically disambiguate words to their proper sense. We employ CNN, Bi-LSTM, and BERT for the classification task and attain 90%, 88%, and 93% accuracy respectively. For the WSD task, we have employed two experiments. When we use the masking technique of the retrained 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.

1 Introduction

Natural language processing (NLP) is a field of artificial intelligence that assists computers in understanding, interpreting, and manipulating human language. 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, and others.

Most of the words in natural languages are polysemic (Hassen, 2015). Amharic is one of the languages that have many words with multiple meanings. WSD is a central concern and a hard challenge in NLP, intending to determine the exact sense of an ambiguous word in a particular context (Huang et al., 2019). 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 (Wiedemann et al., 2019). 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). In this work, we have exploited recently released pre-trained contextual models (Yimam et al., 2021) to build automatic WSD for Amharic.

2 Related Work

Some works address Amharic WSD using handcrafted rules. The research by Hassen (2015) developed a knowledge-based approach for Amharic WSD based on WordNet to extract knowledge from word definitions and relations among words and senses. In this research, they prepared Amharic WordNet manually and selected 2000 words including ambiguous words. They have conducted two experiments, evaluating the effect of Amharic WordNet with and without a morphological analyzer, achieving an accuracy of 57.5 % and 80%, respectively. The study by Mulugeta (2019) has developed an Amharic WSD system that uses the Amharic WordNet hierarchy as a knowledge base. They use context to gloss overlap augmented semantic space approach. Experimental result shows that context-to-gloss followed by augmented semantic space has achieved the highest recall 87% and 79% for three target words at word and sentence level respectively. However, as an issue faced by many low-resource language re-
searchers, almost all researchers do not publish the resources. As far as our knowledge is concerned, our resources (dataset, classification models, and annotation guidelines) are the first resource to be publicly released to advance Amharic WSD research.

3 Experimental Setup

3.1 Dataset Description

As far as we know, there is no standard sense-tagged Amharic dataset for WSD (Mulugeta, 2019). For classification, we have collected 800 ambiguous words and 10K sentences from Amharic news, Amharic dictionary, Amharic bible, and Amharic textbooks. For the disambiguation, a total of 33,297 sentences are used to finetune the AmRoBERTa model (transfer learning).

3.2 Data set Annotation

We have done two different annotations. The first annotation is to know whether the data set contains all the selected relationships of an ambiguous word or not. Therefore, we have selected three Amharic language and Literature Department experts to annotate the data. The second annotation determines the sense of the word. As shown in figure 1 for this task, we have used the WebAnno annotation tool to annotate the ambiguous word in the sentence. We have selected two annotators and one curator of Amharic language native speakers. The annotators annotate the sense of the word in the sentence by using the WebAnno annotation tool.

3.3 Word sense Disambiguation

AmRoBERTa fine-tuning: We fine-tuned the AmRoBERTa model using 33,297 sentences and 800 ambiguous words. Our experiment is conducted using an epoch of 200 and the batch_size is 64 in 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. 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 proper fine-tuning, we assume 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: ከእንታት አስት ያለው መትጋት ከምንም በላይ መሰረታዊ ነጥብ ነው። From this sentence the ambiguous word አስት (lik) is disambiguate as follow. Based on our experimental result shown in Figure 2, the model masks the ambiguous word አስት (lik) then the top 4 meanings of the masked word are predicted.

Word Sense Disambiguation with Flair embedding technique: For this experiment, we have used the fine-tuned retrained 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. 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 ዓይነ (Wana), ለበት (Menged), ከለ (Sale), ከሳ (Akal), ዺተ (Waga), ይር (Gena), ካወ (Qena), ሆወ (haq), ለያ (Hayil), and አስት (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 and the sense is already annotated by the annotators. Then we use the FLAIR document
embedding with the fine-tuned 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. So based on the given sentence in the gloss, the model disambiguate the target word into its correct sense. Example: The sentence: "እያንዳንዱ ዋና ሃሳብ ራሱን በቻለ አንቀፅ ውስጥ ሰፋሯል።" is disambiguate as follow.

Based on the result of our experiment, as shown in figure 3 for the target sentence "እያንዳንዱ ዋና ሃሳብ ራሱን በቻለ አንቀፅ ውስጥ ሰፋሯል።" the correct sense of the ambiguous word ዋና (wana) is ይቶስት (Chibt - main point), as it has higher similarity with the target sentence (0.5702) compared to the other senses, which are እየነተኛ (Aynetegna - principal) and መሪ/ሀላፊ (Meri/Halafi -leader) with similarity scores of 0.5347 and 0.3580 respectively.

4 Discussion

We have set a few parameters during model configuration that are suitable for managing the feature of our models, such as a dense layer, dropout layer, number of neurons in each dense, learning rate, and activation function. For each of the three models, we run three experiments based on the selected hyperparameters, such as CNN, BiLSTM, and BERT. These three models score classification with a performance of 90%, 88%, and 93% respectively. For the disambiguation model, we have employed the AmRoBERTa model, where the models attain an accuracy of 70% and 71% using the masking and target-gloss similarity approaches. For future work, we recommended considering syntactic, phonological, and referential ambiguity. We also recommended an Amharic WordNet as it is important for the evaluation.

5 Conclusion and Recommendation

This study has developed an Amharic WSD model by using a transfer learning approach. To develop the disambiguation model we have used 33,297 sentences. For classification, we have experimented with CNN, BiLSTM, and BERT algorithms. According to the results, CNN, Bi-LSTM, and BERT obtained 90%, 88%, and 93% accuracy respectively. For the disambiguation model, we have employed the AmRoBERTa model, where the models attain an accuracy of 70% and 71% using the masking and target-gloss similarity approaches. For future work, we recommended considering syntactic, phonological, and referential ambiguity. We also recommended an Amharic WordNet as it is important for the evaluation.

References


