Etymological Embeddings for Contextless Definition Modeling

Noah Gardner
Kennesaw State University
College of Computing and Software Engineering
ngardn10@students.kennesaw.edu

Abstract

Definition modeling is the problem of estimating the probability of an output definition given an input word embedding. There exist some methods of creating word embeddings concatenated that take the input word concatenated with the context of the word. However, the context is not always available. Additionally, progress has been made in research for *etymology modeling* where the etymology of a word can be estimated from the input word embedding. In this paper, we propose a definition modeling method that uses etymological information.

1 Introduction

Word embeddings are vector representations of words that allow us to use words as inputs to machine learning models for natural language processing tasks (Mikolov et al., 2013). There are many word embedding methods that achieve state-of-the-art performance on NLP problems, such as sentiment analysis. Addtionally, contextualized word embeddings have been shown to improve performance with models such as ELMo and BERT (Peters et al., 2018; Devlin et al., 2019).

Dictionary definitions can yield value for sentiment aware models, however, crowdsourced annoations are costly. The task of definition modeling was proposed to address this problem. The goal of definition modeling is to estimate the probability of an output definition given an input word embedding (Noraset et al., 2016).

Wiktionary[1] is a free online dictionary that provides details about many words, including definitions, etymologies, pronounciation, and examples. Some research has used the large data dumped from Wikitionary to support NLP research such as etymology modeling and word sense disambiguation (Wu and Yarowsky, 2020).

[1] wiktionary.org

The etymology of a word is a tree structure that describes the word's origin. Although contextualized embeddings show improvement in NLP tasks, the context of a word is not always available. With advances in etymology modeling, if we know the source language of a word, we can predict the etymology of the word. Using this observation, we propose a word embedding with etymological information for the task of definition modeling. This work intends to show improvement for contextless definition modeling, although it may be used in conjuction with a contextualized embedding for even better performance.

2 Related Work

Definition modeling was intially described by Noraset et al. (2016). Their research is based on a recurrent neural network language model (Mikolov et al., 2010) with a modified recurrent unit. They use the word to be defined placed at the beginning of the definition so the model will see the word only on the first step.

Chang and Chen (2019) explore contextualized embedding for definition modeling. They reformulate the problem of definition modeling from text generation to text classification. Their results show state-of-the-art performance on the task of definition modeling.

Washio et al. (2019) proposed a method for context-based definition modeling that considers the semantic relations between both the word to be defined and the words in the definition. They apply semantic information to both the definition encoder and decoder.

Barba et al. (2021) introduce exemplification modeling, an adjacent problem to definition modeling that uses a definition embedding to generate possible example sentences. They use a sequence-to-sequence based approach and show near human-

level annotation performance. Their problem is similar in that they use the definition as context to create example sentences.

3 Overview

This section will provide an overview of the different works required to solve the problem of definition modeling with embeddings.

3.1 Raw Dataset

In this subsection, we investigate the parsed wikitionary dump from Wu and Yarowsky (2020) and discuss the relevant aspects of the dataset.

3.1.1 Definition

Word	Definition
free	$(lb\|en\|social)$ Unconstrained.
free	Not imprisoned or enslaved.
free	Generous; liberal.

Table 1: Parsed wikitionary example definitions for the english word *free*.

The definition dataset includes information on the source language, the word to be defined, the part of speech, and the defintion of the word. The definition of a word can also contain a specific context in which the definition is used. Figure 1 shows some example definitions from the dataset for the word *free*.

3.1.2 Etymology

Similar to the definition dataset, the etymology dataset includes information on the source language, the word to be analyzed, and the etymology of the word. The etymology is a tree structure that describes the word's origin, including roots from other languages. Figure 1 shows an example etymology from the dataset for the word *free*.

free [(eng—free)(root)(eng—ine-pro—*preyH-)]

Figure 1: Parsed wikitionary example etymology for the english word *free*.

3.2 Embeddings

In this subsection, we will investigate how we can create definition embeddings for definition modeling.

3.2.1 Universal Sentence Encoding

In order to create defintions, we need to obtain the embedding of the definition we wish to predict. We also must obtain embeddings of the input word and context. By using a pre-trained universal sentence encoder, we can obtain embeddings for both the inputs and the expected outputs.

3.2.2 Defintion Embedding Prediction

Once we have the input embeddings and expected output embeddings, we can train a model to predict the output embedding. In general, the problem of definition modelling is a text generation problem. We may also solve the problem as a classification task. That is, predict the closest definition in the dataset to the predicted embedding using a distance metric such as cosine similarity, the proposed approach of Chang and Chen (2019).

In their experiments, after creating a contextualized embedding as well as a word embedding, both are used as inputs to a convolutional neural network. The resulting model is able to predict a definition embeddinggs for the given input. This approach allows them to avoid some of the pitfalls of text generation, such as problems associated with greedy search and generating grammatically correct sentences.

With the predicted definition embedding, we compare the predicted embedding to definition embeddings in the dataset. Using an approach similar to the k-nearest neighbor algorithm, we may find k definitions that are most similar to the predicted embedding.

4 Methodology

4.1 Dataset

From the definition and etymology datasets, we use only the words with the english source language. For the defintions, we remove self-referential defintions and utilize a maximum of three definitions per word. Although the context-based definitions may provide some benefit for a context-based model, we remove the context from the definitions. Additionally, the definition of some words are completely context (such as alternate spellings) and are also removed. Fue to background limitations, rather than use the entire etymology tree to generate our etymological embeddings, we use only the first etymology for each word if there exist multiple. Finally, we ignore words tagged with *proper noun*.

Dataset statistics are shown in Figure 2 after the described steps are applied.

Type	Amount
Unique Words	**320855**
Average Definitions Per Word	**1.288**
Etymology (Average Length)	**41.725**
Definition (Average Length)	**51.711**

Table 2: Dataset statistics for the combined datasets.

4.2 Models

A simple sketch of the proposed approach can be found in Figure 2.

4.2.1 Encoder Model

BERT (Peters et al., 2018) is a commonly used encoder for the task of creating sentence embeddings due to it's ability to contextualize information with both forward and backward passes. However, we are working with limited resources and must choose a model with fewer parameters. Therefore, we have chosen the DeCLUTR model for unsupervised textual representation (Giorgi et al., 2020). Aside from being readily accessible in the huggingface library, it is a useful model due to its unsupervised nature.

4.2.2 Prediction Model

As described in the Overview section, we are using a classification-based approach for definition modelling. After our word, definition, and etymology embeddings are created, we use a simple dense neural network with 3 hidden layers to predict the closest definition to the predicted embedding. In order to train the model, we select a set of unique words from the dataset and use them as the input words. The leftover set of words is used for testing. In other words, any word that appears in the training set is does not appear in the testing set.

4.2.3 Definition Selection

Once we have predicted a definition embedding, we use a distance metric to find the closest 3 definitions to the predicted embedding. We use cosine similarity as our distance metric and use an approach similar to the k-nearest neighbor algorithm. Unfortunately, most of the metrics used in definition modelling cannot be applied to this approach as the metrics expect a generated embedding rather than the complete defintion from the dataset.

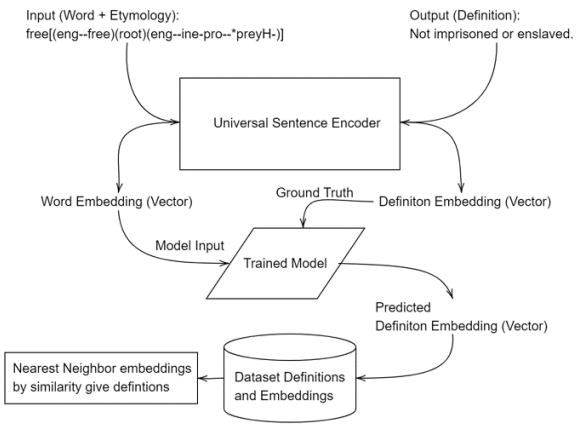

Figure 2: Overview of the proposed approach for definition modelling with a sample input.

```
Best words:
Word tediously Dist 0.8441480398178101
Word unpredictably Dist 0.8461641073226929
Word congruously Dist 0.8499659299850464
Word dryly Dist 0.8510894775390625
Word fabulously Dist 0.8524174094200134
Word ungeographic Dist 0.858171820640564
Word unplayfully Dist 0.8681503534317017
Word unenergetically Dist 0.869099497795105
Word unambivalently Dist 0.874753475189209
Word jeopardously Dist 0.8912408947944641
```

Figure 3: Best cosine similarity scores for the predicted embeddings.

5 Experimental Results

Due to the lack of quantitative metrics for a classification-based definition modelling approach, we simply show some of the best cosine similarity scores for predicted embeddings, shown in Figure 3. We also show some generated definitions in Figures 4 and 5.

```
Word and root etymology:
dryly [(eng|dryly)(suf)(eng|dry|ly)]
Definition:
In a dry manner.
Finding nearest neighbors...

Closest definitions:
In a despiteous manner.
In a cracked manner.
In a Johnsonian manner.
```

Figure 4: Generated definitions for the word *dryly*.

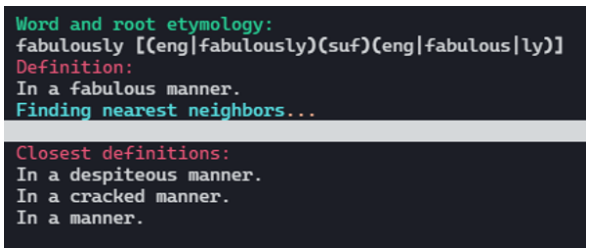

Figure 5: Generated definitions for the word *fabulously*.

6 Conclusion

In this paper, we provided a method of using etymological embeddings instead of contextualized embeddings for the problem of definition modelling. However, the problem of generating sentences from embeddings is a difficult topic. Additionally, due to limited understanding of the author, we use a simplified etymological tree rather than the entire structure. The experimental results show that the model scores best on adverbs ending in *ly*, and the best defintions are all sentences of the form "*In a _ manner.*". This shows either a problem with models used or our approach, as the model is stuck in a local minima of definitions that are technically similar but not necessarily correct to the expected definitions. Future work should address this issue by testing strong models, using the entire etymology tree, and using a more advanced approach to dataset setup. If these approaches can be tested, we believe etymological embeddings have the ability to be as useful as context in the task of definition modelling.

Acknowledgments

This work was supported by computational resources provided by the Kennesaw State University Department of Electrical and Computer Engineering.

References

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification Modeling: Can You Give Me an Example, Please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3779–3785, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Ting-Yun Chang and Yun-Nung Chen. 2019. What Does This Word Mean? Explaining Contextualized Embeddings with Natural Language Definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

John M. Giorgi, Osvald Nitski, Gary D. Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *CoRR*, abs/2006.03659.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Tomás Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2016. Definition Modeling: Learning to define word embeddings in natural language. *arXiv:1612.00394 [cs]*. ArXiv: 1612.00394.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365 [cs]*. ArXiv: 1802.05365.

Koki Washio, Satoshi Sekine, and Tsuneaki Kato. 2019. Bridging the Defined and the Defining: Exploiting Implicit Lexical Semantic Relations in Definition Modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3521–3527, Hong Kong, China. Association for Computational Linguistics.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.