Etymological Embeddings for Contexless 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).

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-

[1] wiktionary.org

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

3 Overview

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

3.1 Definition

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

Figure 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*.

3.2 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 2 shows an example etymology from the dataset for the word *free*.

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

Figure 2: Parsed wikitionary example etymology for the english word *free*.

4 Methodology

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. For the etymologies, 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 3 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**

Figure 3: Dataset statistics for the combined datasets.

5 Experimental Results

6 Conclusion

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.

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.