Definition Prediction Based on Etymology

Noah Gardner[1]

[1]College of Computing and Software Engineering, Kennesaw State University

9/24/2021

1 Abstract

There are many different methods to extract the defition of a word given information about it besides the definition. Sometimes, the definition can be extracted from the context of the word, but the context is not always available. In every case, however, the word itself is available, and we if we can find the etymology used to build the word, we have a good hint of the definition of the word. Automically generated definitions could lower the cost of crowd-sourced annotations. This research proposes a method to extract the definition of a word by parsing the etymology of the word.

2 Introduction

Wiktionary[1] is an online dictionary that provides details about many words, including etymology, definition, pronounciation, and examples. Wiktionary also supports many different languages. Some research has used data dumped from Wikitionary to support NLP research such as etymology prediction and word sense disambiguation.

The etymology of a word is a tree structure that describes the word's origin. The tree is a series of nodes, each of which has a parent and a child. The root node is the word itself. The child nodes are the words that are used to build the word.

[1] wiktionary.org

3 Literature Review

1 Computational Etymology and Word Emergence [1]

Wu et al. describe a comprehensive Wikitionary parser that allows them to predict the etymology of a word across multiple languages. They use a LSTM model for three different experimental settings that explore the relationship between words and their etymology. They also show their model is capabable of predicting the parent language of a word given it's relationship.

2 Augmenting semantic lexicons using word embeddings and transfer learning [2]

Al-Shaabi et al. describe a method to augment the semantic lexicons of a language using word embeddings and transfer learning. They argue that lexicon-based models for sentiment analysis systems are more interpretable and easier to use than contextual models. However, it can be challenging to add sentiment information to new words in the lexicon. They propose two models that are to predict sentiment scores using word embeddings and transfer learning. Their methods show human-level performance on a dataset of annotated sentiment scores.

3 Deep contextualized word representations [3]

Peters et al. describe a deep representation model for words that can be used across NLP tasks such as sentiment analysis and textual entailment. They argue that high quality representations are difficult to learn, and should ideally model characteristics of the word and how it's uses can vary in different contexts. They propose a model that uses both forward and backward contextualized word embeddings. Their methods obtain benchmark results on main NLP tasks and can be incorporated into many modern NLP systems.

4 Methodology

4.1 Dataset

The dataset used for this project will come from *Yawipa Data Extract*[2] dataset that parsed information on each word from Wiktionary, including definitions and etymology [1].

[2] cs.jhu.edu/ winston/yawipa-data.html

4.2 Model

Given a supportive dataset and modern NLP algorithms, it may be possible to train a model to predict the definition of a word given the etymology of the word. The definitons will also come from the Wikitionary dataset but are also available from different datasets and API's.

If it is impossible or proves too difficult to generate the definition of a word, then at least this project will attempt to provide statistical correlations between etymology and definitions.

References

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

[2] Thayer Alshaabi, Colin Van Oort, Mikaela Fudolig, Michael V. Arnold, Christopher M. Danforth, and Peter Sheridan Dodds. Augmenting semantic lexicons using word embeddings and transfer learning. *arXiv:2109.09010 [physics]*, September 2021. arXiv: 2109.09010.

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