Exploration of neural methods on English etymological classification

Robin Huo
robin.huo@mail.utoronto.ca

Tim Logatsang
tim.logatsang@mail.utoronto.ca

Charles Wong
charleshm.wong@mail.utoronto.ca

April 19, 2024

1 Introduction

Historical linguistics deals with different aspects of language change, one of which being etymology, the reconstruction and origin of words and other linguistic items. Traditional approaches to this field mostly involve looking at historical texts and tracing word formation and borrowings. The use of data-driven techniques is thus a lesser explored domain of research. The central observation motivating our approach is that patterns in morphology emerge as a result of borrowings from any specific language into another. As such, it may be possible to identify such classes of vocabulary within a language using statistical modelling, specifically using neural networks.

In this paper, we assess the viability of using different neural network architectures in predicting the source language of borrowings into English. We compare CNN, RNN, and Encoder-only Transformer architectures. An important feature of our approach is that the inputs to all of these models are fixed-length character-level embeddings.

Section 2 of this paper further defines the problem space and some theoretical challenges. Section 3 discusses prior work done in this area. Sections 4 and 5 describes our methodology, and presents our experiemental results. Section 6 includes our conclusion and discussion of future work.

2 Problem definition

Etymology is the study of the origin of words or other linguistic items. The term *etymology* is also used to refer to a specific etymological origin. Generally speaking, words and linguistic structures are either borrowed or inherited. A linguistic item or construction is said to be borrowed if it is introduced into the language from an external language or variety, which provides the borrowed form. Borrowed words are also commonly called loanwords. For example, the English word *pasta* is borrowed from the Italian word *pasta*, which means pasta or noodles. Anything that is not borrowed is inherited. For example, the word *word* has existed in English, in this or an earlier form, for as long as historical linguists can presently discern.

Words of different origins may sometimes interact in different ways with the other systems of the language. For example, words of Latinate origin typically take the *in-* prefix when deriving a negated form, e.g., *in-validate*. On the other hand, native English words typically take the *un-* prefix instead, e.g., *un-lock*. Thus, knowing the etymology of the word can in some cases allow a speaker or system to more competently use language.

It is common for words of different sources to display characteristic formal properties reflecting their origins. Thus, it is possible in some cases to predict the origin of a word given its form, to an extent. It is this problem that we wish to explore using neural techniques. Specifically, given an English word as input, we wish to predict the origin of the word given a set of predefined etymological classes. We will use the etymology-db (Roher, 2023) dataset (more detail later) for etymological data.

3 Summary of prior work

While we are the first to apply such methods to etymological classification within the English language in particular, use of neural and deep learning methods to identify cases of lexical borrowing as well as the donor language has been explored in the past.

Mi et al. (2016) uses recurrent neural networks (RNNs) to classify loanwords in Uyghur as originating from one of Chinese, Russian, and Arabic, achieving a macro-averaged F1 score of 0.7950. Later, Mi, Yang, Wang, Zhou, and Jiang (2018) uses convolutional neural networks (CNNs) with long short-term memory on the same problem to achieve an average F1 score of 0.8040.

J. E. Miller et al. (2020) implements an RNN among other models in the task of predicting whether a word is borrowed or inherited across 41 different languages, with up to 2,000 words per language as training data, achieving an F1 score of 0.661. J. Miller, Pariasca, and Beltran Castañon (2021) applies a simplified transformer model to the same task, achieving a lower F1 score of 0.559, but a shorter training time than the baseline based on RNNs.

One of the authors previously submitted a project for CSC311, applying CNNs to achieve an F1 score of 0.83101 in classifying Korean nouns across nine different languages of origin.

4 Methodology

4.1 CNN

4.2 RNN

4.3 Transformer

5 Experiments and Results

6 Conclusion

References

Doherty, L., Panigrahi, S., & do Carmo, E. (2016). *ipa-dict - monolingual wordlists with pronunciation information in ipa.* Retrieved from https://github.com/spellcheck-ko/korean-dict-nikl/tree/master

Mi, C., Yang, Y., Wang, L., Zhou, X., & Jiang, T. (2018, May). A neural network based model for loanword identification in Uyghur. In N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1565

Mi, C., Yang, Y., Zhou, X., Wang, L., Li, X., & Jiang, T. (2016, October). Recurrent neural network based loanwords identification in Uyghur. In *Proceedings of the 30th pacific asia conference on language, information and computation: Oral papers* (pp. 209–217). Seoul, South Korea. Retrieved from https://aclanthology.org/Y16-2019

Miller, J., Pariasca, E., & Beltran Castañon, C. (2021, September). Neural borrowing detection with monolingual lexical models. In S. Djabri, D. Gimadi, T. Mihaylova, & I. Nikolova-Koleva (Eds.), *Proceedings of the student research workshop associated with ranlp 2021* (pp. 109–117). Online: INCOMA Ltd. Retrieved from https://aclanthology.org/2021.ranlp-srw.16

Miller, J. E., Tresoldi, T., Zariquiey, R., Beltrán Castañón, C. A., Morozova, N., & List, J.-M. (2020). Using lexical language models to detect borrowings in monolingual wordlists. *Plos one, 15*(12), e0242709.

Roher, D. (2023). *etymology-db.* Retrieved from https://github.com/droher/etymology-db