Hierarchical Language Clustering using Sentence Embeddings

Tristan McDermott
02012319

Quang Tran
01991459

Alvin Yu
02039560

Abstract

In this project, we used the Europarl parallel corpus to determine the "distance" between pairs of languages based on the average distance between equivalent sentences' embeddings, and then used those computed distances to hierarchically cluster the languages represented in the dataset in an attempt to recreate the Indo-European language family tree. We found that the tree we generated does somewhat accurately recreate the gold standard tree, but other methods that take syntax into account can do it better.

1 Introduction

In the past, there have been at least 2 papers where the authors aimed to reconstruct a known linguistic family tree using computational methods: in (Serva and Petroni, 2008), they used average Levenshtein distance between equivalent words, and in (Rabinovich et al., 2017), they used the confusion matrix generated by a classifier trained to guess the original language of sentences translated into English. Both of these papers came up with their own notion of "distance" between languages, and then used a matrix of pairwise distances to hierarchically cluster them and create a family tree. Our notion of distance is generated by taking pairs of corresponding sentences between 2 languages, passing them through a sentence embedding model, and then finding the average distances between these corresponding embeddings. The dataset of sentences we used is the same one used in (Rabinovich et al., 2017).

The particular set of languages that we made a tree for have already been studied in great detail by linguists and so their family tree is already very well-known and uncontroversial, making the fact that we constructed a tree for them on its own nothing special. What we aimed to find out with this project is, do multilingual sentence embedding models capture and represent known relationships between languages? If they do, do they capture genetic similarity, or something else like geographic, syntactic, or lexical similarity?

2 Method

We started with a dataset of parallel sentences translated into various languages. As we will further describe in the next section, we used the Europarl dataset. We filtered out all the sentences in the dataset's 3 non-Indo-European languages, as well as Greek, and then further filtered out all the sentences that were not available in the 17 remaining languages. For all pairs of languages, we defined the distance between them to be the average distance between corresponding sentences' embedding vectors. The embeddings were generated by Google's LaBSE model (Feng et al., 2022). They were of dimension 768 and normalized (all embeddings produced by LaBSE are normalized by default). We used angular distance ($\arccos(A \cdot B) \in [0, \pi]$) as our notion of distance between vectors. The calculated distances between languages were then put into a 17×17 matrix, which was used to hierarchically cluster the languages and create a tree. This was done via a function provided by scipy. We used the unweighted average linkage method, which was also used in (Serva and Petroni, 2008). Our gold standard tree came from (Rabinovich et al., 2017), but that was a pruned version of the tree from (Serva and Petroni, 2008).

3 Data

We used the Europarl parallel corpus, which consists of European Parliament proceedings trans-

lated into 21 official languages of the European Union. Of these 21 languages, 18 are Indo-European, and the 3 that aren't are Uralic (Estonian, Finnish, and Hungarian). Given that our aim was to create a phylogenetic tree, we only focused on the Indo-European languages in this project, minus Greek since the paper we were comparing our tree to didn't have it. The specific release of the Europarl dataset we used comes from (Transformers, 2025), which has aligned sentences for all pairs consisting of English and the other 20 languages. Not every sentence is translated into every language: Romanian only has 387,000 sentences while French and Dutch both have slightly north of 2 million. English has around 2.4 million. The number of sentences translated into all 18 languages we're considering is around 200,000. The average sentence length also varies by language, from as low as 19.7 words in Latvian to 28.2 in Spanish.

4 Results

4.1 Evaluation

In comparison with the gold tree, our tree was quite different. We evaluated the differences between our tree and the gold tree through taking the sum of squares of the difference in the number of edges between each language pair in the gold tree and our tree. Using this metric, our tree has a distance of 2,144 from the gold tree, while the tree from (Rabinovich et al., 2017) has a distance of only 1,268. See figure 2 for the calculated distance matrix between languages, which were used to create our tree. Notice how high the distances are for English, even between it and other Germanic languages. We made another distance matrix for the number of edges between each pair of languages for both the constructed tree from (Rabinovich et al., 2017) and our tree vs. the gold tree in figures 3 and 4 respectively. The third distance matrix is a confusion matrix that shows the differences between the distance matrices.

The reason our tree was so different was because of the way the embeddings model was trained, which was to create semantic level embeddings. Meanwhile, the tree in (Rabinovich et al., 2017) used a different method, using structural, syntactic features rather than semantic ones.

4.2 Insights

In the EU Parliament, members speak in their native language. Their speech is translated into English, and then from English to all the listeners' native languages. In (Rabinovich et al., 2017), the authors talk about how translators' idiosyncrasies can impart certain characteristics onto their translations which, as they proved, can give you some information about what the source language of the translation is. They called this process "interference." We believe this may help explain why English is so distant from all the other languages: English sentences in our dataset only had to go through at most 1 translation, while sentences in any other language might be the product of up to 2 translations, giving them a double dose of common translation artifacts such as simplification, standardization, and explicitation (Rabinovich et al., 2017). All the Non-English target sentences would share these features, but the English sentences wouldn't, making English artificially different from all the other languages.

What we can conclude about sentence embedding models in general is that they can only be said to encode genetic relationships between languages to a small degree. They're able to reconstruct the 3 big subfamilies and relations between particularly close language pairs like Portuguese-Spanish and Czech-Slovak, but within the subfamilies the reconstructions are less accurate. In (Rabinovich et al., 2017), they represented sentences as part of speech trigrams, which are able to encode genetic relationships much better via syntax. Some syntactic constructions are used in certain language (sub)families more/less than others: Slavic languages don't have articles while Romance languages do; Germanic languages tend to use the perfect a lot. Sentence embedding models are focused much more on encoding semantic similarity between sentences than syntactic similarity, which doesn't give us nearly as much info about a language's typology. Since these models are still able to somewhat accurately reconstruct family trees, what we could say about this is that closely-related languages tend to express ideas in semantically similar ways to each other.

5 Conclusion

We explored the use of multilingual sentence embeddings to try to reconstruct the Indo-European family tree, using hierarchical clustering to pro-

duce a dendrogram. We found that our tree has partial similarities to the gold standard phylogenetic tree. This shows us that sentence embedding models can be used to construct phylogenetic language trees, since to a small degree, semantics reflects some level of genetic relationships. However, in order to truly construct a tree that has proper genetic relations, we need to also be able to capture syntax in the embeddings model. For future research, we could try using a model that converts syntax trees to embeddings and then apply the method from this paper to them. This way, syntactic information could be captured much more explicitly and should boost the accuracy of the generated tree.

6 Contribution Chart

Task	Student ID	Contribution
Proposal	02012319	Wrote the project proposal
Embeddings	02012319	Used the GPU server to calculate embeddings and distances
Tree visualization	01991459	Wrote the code to create and visualize our tree
Evaluation	02039560	Created evaluation matrices
Presentation	02039560	Created the presentation
Report	02012319	Wrote sections: Introduction, Method, Insights
Report	02039560	Wrote sections: Data, Evaluation, Conclusions

References

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations.

M. Serva and F. Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Sentence Transformers. 2025. sentence-transformers/parallel-sentences-europarl.

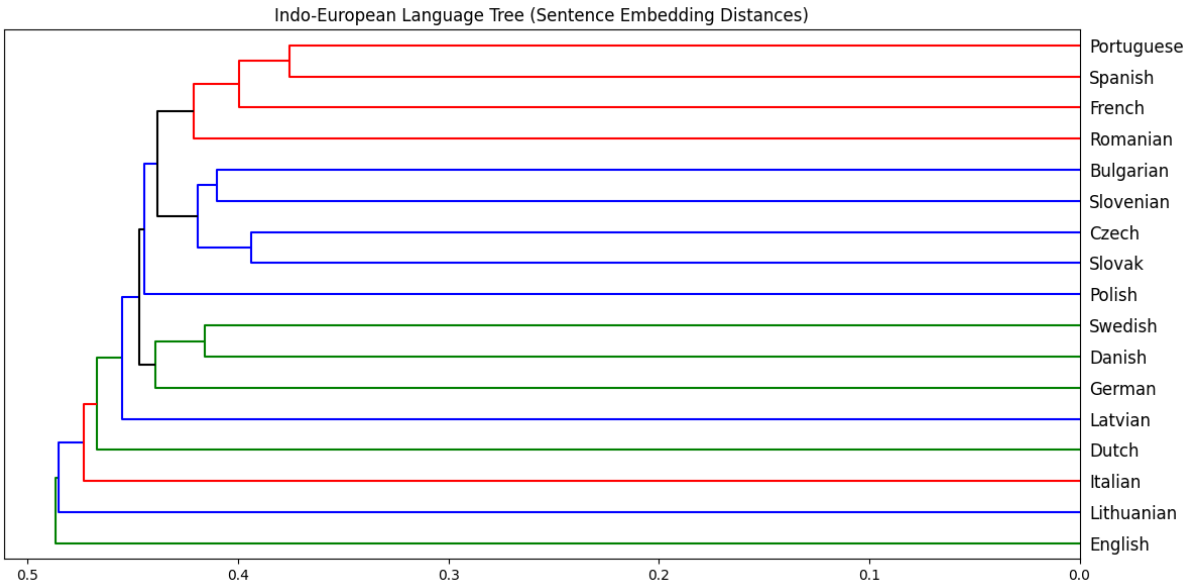

Figure 1: Our generated tree from the LaBSE embeddings model.

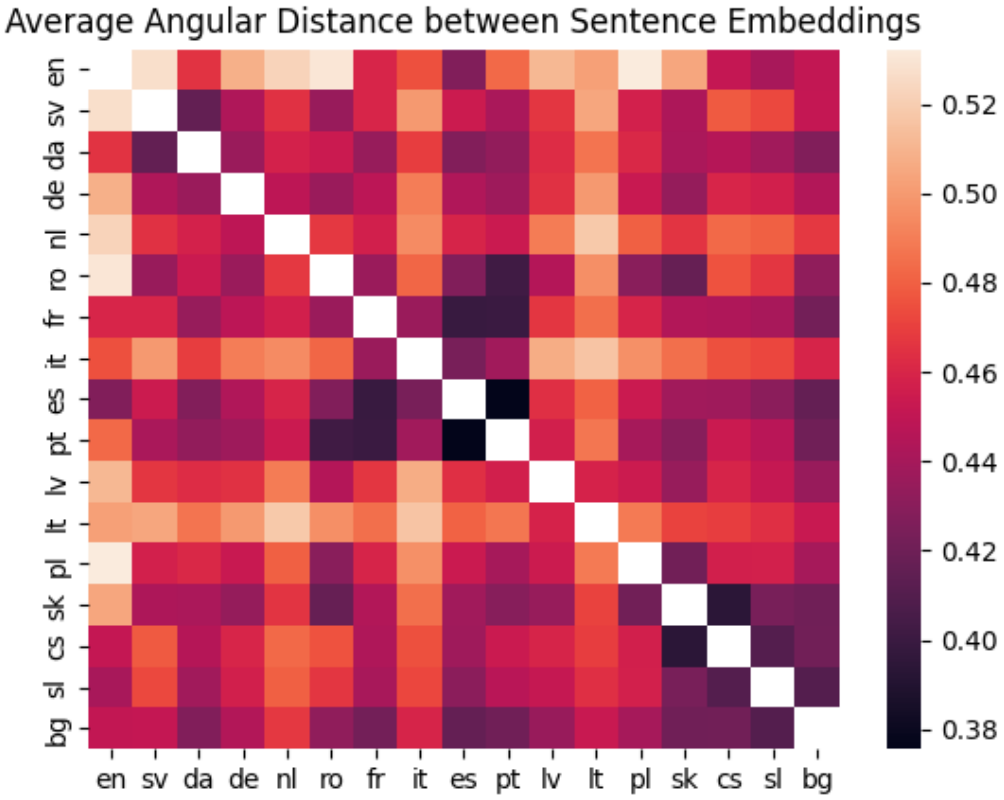

Figure 2: The distance matrix between languages, which was used to create our tree

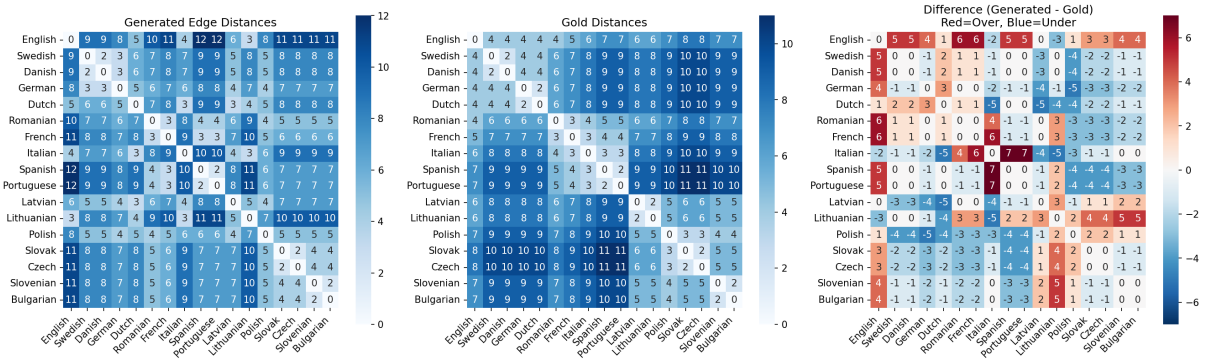

Figure 3: Distance matrix comparing the number of edges between languages for our tree vs the gold tree. The third matrix shows the differences between the distance matrix, where white means the same, and the intensity of the red or blue means higher or lower than the gold number of edges respectively.

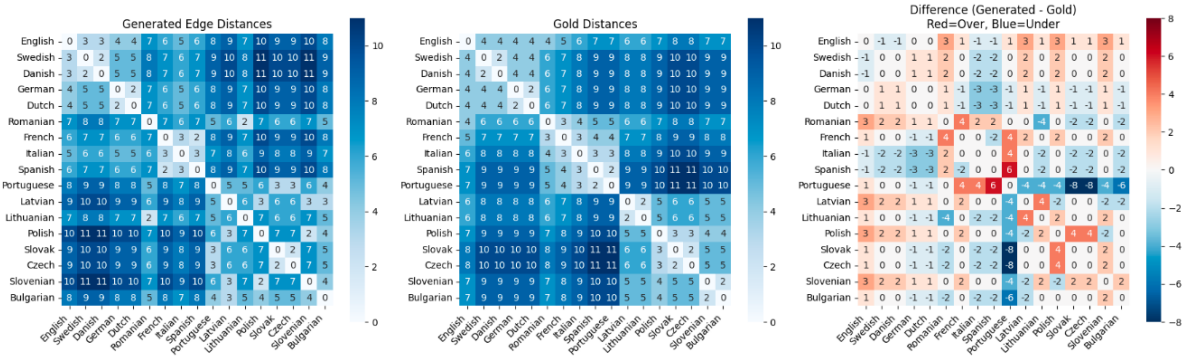

Figure 4: Same as figure 3, but comparing the tree from (Rabinovich et al., 2017) with the gold tree