Project Baseline

Christian Konstantinov and Moeez Sheikh

1 Data

We will be using data from two different sources. The first one is Cognet v2.0 [KG], and the other one is a dataset from FAA 2020: Cognate Identification competition on Kaggle.com [IEL]. Cognet v2.0 is a large-scale high-quality cognate database. It has about 8.1 million cognates in 338 languages. The quality of this database was manually evaluated to be 94% precise. The reason why we had to use a second database was that Cognet does not have any negative cases. We need to test our data on an equally balanced dataset that has both negative and positive cases of cognates. This is where the FAA 2020 competition's dataset comes in. This dataset has about 113,9907 entries of negative and positive values. We combined the two datasets and divided the data into equally distributed sets of negative and positive values. From this combined data, we extracted a dataset for training and another one for testing. Our new datasets have 395,286 positive cases and 352,691 negative cases of cognates in various languages.

2 Metrics

The metrics that we need to measure for our model are accuracy, precision, and recall. For this reason, we will be using the F1 score metric as our measure of success.

3 Baseline Method That we will be Comparing Against

4 How Well Does our Baseline Method Perform?

References

[IEL] Indo-European Lexical Cognacy Database (IELex). *FAA 2020: Cognate Identification*. URL: https://www.kaggle.com/c/faa-2020-cognate-identification/. (accessed: 03.16.2021).

[KG] Gabor Bella Khuyagbaatar Batsuren and Fausto Giunchiglia. *CogNet: A large-scale cognate database, Proceedings of The 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.* URL: https://aclweb.org/anthology/papers/P/P19/P19-1302/. (accessed: 03.16.2021).