An Analysis of Cognate Identification Models on CogNet

Christian Konstantinov and Moeez Sheikh

1 Introduction

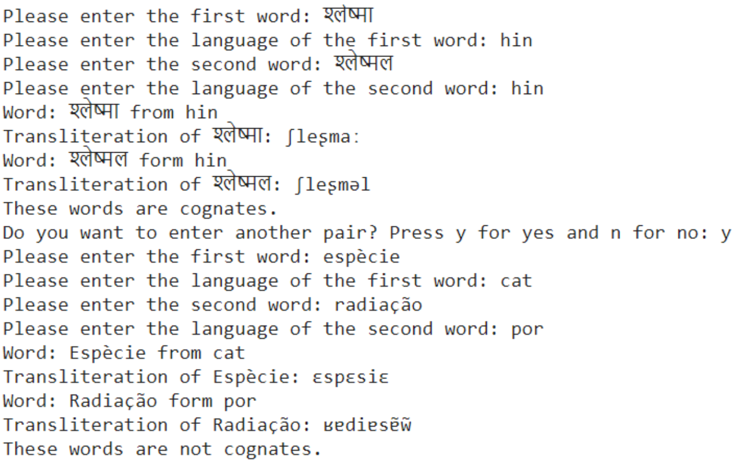
In this project, we explore various Natural Language Processing (NLP) concepts for identifying cognates across multiple languages. Cognates are defined as words that share a common language ancestor. In this paper, we will discuss our findings and the approach we took to create a model that accurately identifies cognates.

2 Goal

Our aim in this project was to explore various methods of identifying cognates and synthesize our findings to create a model that would perform well and produce accurate results. A classifier that identifies cognates accurately can become a part of a larger epistemology model. Such a model can help us in identifying how languages changed over the millenniums. It can also help us in finding the origin of words and preserving endangered languages.

2.1 User Interface

With the help of our program, a user will be able to input two words and their languages. In return, the program will output the transliteration of the two words and also tell the user if two words are cognates or not. A more programmatic way of interacting with our program would be to input a list of tuples, which would contain two words and their respective languages. The program will then output if the words are cognates or not, along with the words' transliteration.



```
Please enter the first word: श्लेष्मा
Please enter the language of the first word: hin
Please enter the second word: श्लेष्मल
Please enter the language of the second word: hin
Word: श्लेष्मा from hin
Transliteration of श्लेष्मा: ʃleʂmaː
Word: श्लेष्मल form hin
Transliteration of श्लेष्मल: ʃleʂməl
These words are cognates.
Do you want to enter another pair? Press y for yes and n for no: y
Please enter the first word: espècie
Please enter the language of the first word: cat
Please enter the second word: radiação
Please enter the language of the second word: por
Word: Espècie from cat
Transliteration of Espècie: ɛspɛsiɛ
Word: Radiação form por
Transliteration of Radiação: ʁediɐsɐ̃w̃
These words are not cognates.
```

Figure 1: An example of user interaction

3 Methods

There are many techniques for identifying cognates. Our methods primarily follow the work of [1] and [2]. We apply their proposed features on a variation of the CogNet 2.0 dataset [3]. We attempted to create a model that can be used on different languages, where the relationships between those languages may not be known. To this end, our model mainly utilizes orthographic and phonetic word features described below.

3.1 Orthographic Features

We use word distance features as described by [1].

1. Edit Distance.
2. Longest Common Sub-sequence Ratio (LCSR).
3. Dice Coefficient: Number of shared character bigrams / total number of bigrams.
4. Prefix Coefficient: Longest common prefix / length of the longer word.

3.2 Phonetic Features

1. IPA transliterations [4].
2. Multi-hot phoneme vector encodings [5].
3. SOUNDEX Code [6].
4. NYSIIS code [7].

Word distance measures are a common orthographic approach in cognate identification research [1], [2], [8]. The distances our model uses enable us to produce accurate results for words that belong to languages that share a common script. CogNet provides romanizations for these languages which are used for training [3]. If two words do not have available transliterations, then the model may make incorrect predictions. The phonetic features we utilize allow us to extend our comparisons to incorporate sonic similarities between words.

3.3 Models

We studied both statistical and machine learning models for our implementation. Our research brought us to the following models for our cognate identification task.

1. Multi Layer Perceptron Classifier (MLPC).
2. Conditional Random Field (CRF).
3. Siamese Convolutional Neural Networks (SCNN) [2].
4. Long Short-Term Memory (LSTM).

These models enabled us to learn the features that we described above. We implemented these models with a combination of different features[1]. Our results are discussed in 4.4.

[1] Due to time and budget constraints, we were unable to successfully implement the SCNN and LSTM.

4 Experiments

This section describes the data that we used for this project, metrics that evaluated our model's success rate, the baseline methods we compared against, and how well our final models performed compared to the baseline method.

4.1 Data

4.1.1 Baseline

The baseline model uses an aggregate dataset, which we compiled from CogNet v2.0 [3], and a dataset from the FAA 2020: Cognate Identification competition [9]. [3] Is a high-quality, large-scale, multilingual cognate database. It contains approximately 8.1 million cognates in 338 languages. The database was evaluated to score 94% on precision. [9] Provides false cognates in addition to true cognates; necessary data that [3] lacks.

4.1.2 Improvements

Unfortunately, [9] is a smaller dataset which needs further processing as some of the words transliterated to IPA script and some are not. For this reason, we did not use the data from the Cognate Identification shared task, opting to work solely with CogNet instead.

Our final model utilizes Epitran, a package for transliterating text into IPA text [4]. To prepare the data for feature extraction, we take only word pairs (along with their romanizations) in languages that Epitran supports. This reduces CogNet's 8.1 million word pairs to approximately 1.76 million. To create false cognates, the word pairs are then randomly scrambled. Any false negatives are then removed, leaving approximately 3.5 million word pairs in the dataset. The data is then split by 80, 10, and 10 percent for training, testing, and validation respectively.

4.2 Training

4.2.1 Baseline

Our model was compared against a baseline MLPC trained on our aggregate dataset implemented with scikit-learn. The PREFIX coefficient, Dice's similarity coefficient, and Least Common Subsequence Ratio (LCSR) as described in [1] is extracted from each cognate pair, then converted into a feature vector. The feature vector is input into the MLPC, with the target output being each corresponding class of true or false cognate. The model was trained for at most 500 epochs with a stopping criteria tolerance of 10^{-9} and a patience of 10 epochs. Because our aggregate dataset was sufficiently large, the Adam optimizer was used with a learning rate of 10^{-3}.

4.2.2 Improvements

Another problem with this approach was that we were not introducing enough features to our model. For instance, we were not using any features associated with the phonetic similarity of words. In our second experiment, the features we used included SOUNDEX and NYSIIS phonetic codes, along with the distances over word characters. We used an MLPC in our second experiment as well. We achieved far better scores because of the inclusion of these new features.

We also worked with a CRF model. This model produced higher scores using the same features that are described above. The training algorithm that we used in this model was Gradient descent using the L-BFGS method.

In our last experiment, we introduced another feature to our model. This feature was the LCSR over the IPA transliteration of the words. We used a similar MLPC model as described above with the inclusion of this new feature.

This experiment proved to be our best performing model. The details of the results that we achieved through these models are described in 4.4

Because we use phonetic encodings and LCSR of IPA transliterations, we call our final model the Phonetic MLPC. For the Phonetic MLPC, 3 hidden layers with a dimension size of 100 is used. Validation is performed with a tolerance of 10^{-4} and a patience of 10 epochs. The Adam optimizer with an adaptive learning rate is used.

4.3 Metrics

The performance of our cognate identification models is measured by accuracy, F_1 score, precision, and recall. We define accuracy similar to [2]: the number of correctly labeled word pairs (cognate or non-cognate) divided by the number of total word pairs. Furthermore, since the weight of precision and recall should be equal, the F_1 score is used.

4.4 Results

The results of our experiments evaluated using the metrics described in 4.3 are shown in Table 1. The Phonetic MLPC reaches a validation score of 95.7257% after 19 epochs. It outperforms the CRF model on the CogNet dataset in accuracy, F_1 score, precision, and recall.

Model	**Accuracy**	**F_1**	**Precision**	**Recall**
Baseline (aggregate data)	74.34	69.87	92.10	56.28
CRF	92.30	92.45	90.66	94.32
Phonetic MLPC	95.71	95.66	96.97	94.38

Table 1: *Accuracy, F_1 score, precision, and recall of the tested models.*

5 Outputs

Even though our model gives us quite accurate results, it still fails for some words. Some of the word pairs that our model fails to classify correctly are as follows.

False Positives	**False Negatives**
äquator | islàmic	konditor | confitero
langganan | kosaikhana	telemetria | télémesure
terambah | abah	flor | fjura
leptodactylus | leccinum	morozilnik | xolodylnyk
informaticien | imatge	terjulur | mencocol

Table 2: *Faulty Results.*

For the words *terambah* and *abah* with an LCSR of 0.5 and an edit distance of 4, we can see that our model falsely identifies these words as cognates. The words *telemetria* and *télémesure* highlight the model's likely inability to match the characters *e* and *é*. [8] Use a substitution matrix that ignores diacritics, ensuring that characters such as those aforementioned will match.

Another way we can produce better results is by incorporating word embeddings in our models. This may also improve precision on data that includes cognates such as Hindi's *chakra* and English's *wheel* [2].

6 Conclusion

In this paper, we analyzed the performance of different models for identifying cognate pairs in large scale datasets, and described our process to achieving better results with modern NLP tools. Our models do not need additional language features, as they are constructed to work with orthographic distance features alone. With the addition of phonetic transliterations and encodings, we were able to further improve performance. Our models, however, may show worse performance on languages where transliterations are not available, as we noticed in our baseline.

The task of cognate identification is a problem that is worth researching further. Semantics and language features in addition to orthographic and phonetic features could help classify cognates in future research. A cognate classifier could also be used to train generative models for cognate discovery, such as a discriminator in an adversarial model. For these reasons, we believe that this task is worth researching further.

References

[1] G. Kondrak, "Identification of cognates and recurrent sound correspondences in word lists," *TAL*, vol. 50, pp. 201–235, Jan. 2009.

[2] T. Rama, "Siamese convolutional networks for cognate identification," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 1018–1027. [Online]. Available: `https://www.aclweb.org/anthology/C16-1097`.

[3] K. Batsuren, G. Bella, and F. Giunchiglia, "CogNet: A large-scale cognate database," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3136–3145. DOI: `10.18653/v1/P19-1302`. [Online]. Available: `https://www.aclweb.org/anthology/P19-1302`.

[4] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. (chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 2018, ISBN: 979-10-95546-00-9.

[5] D. R. Mortensen, P. Littell, A. Bharadwaj, K. Goyal, C. Dyer, and L. Levin, "PanPhon: A resource for mapping IPA segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 3475–3484. [Online]. Available: `https://www.aclweb.org/anthology/C16-1328`.

[6] (). "Soundex system," [Online]. Available: `https://www.archives.gov/research/census/soundex`.

[7] P. E. Black. (Mar. 2019), [Online]. Available: `https://www.nist.gov/dads/HTML/nysiis.html`.

[8] A. Cristea and L. Dinu, "Automatic detection of cognates using orthographic alignment," vol. 2, Jun. 2014, pp. 99–105. DOI: `10.3115/v1/P14-2017`.

[9] I.-E. L. C. D. (IELex). (). "Faa 2020: Cognate identification," [Online]. Available: `https://www.kaggle.com/c/faa-2020-cognate-identification/`. (accessed: 03.16.2021).