CoToHiLi at SIGTYP 2023: Ensemble Models for Cognate and Derivative Words Detection

Liviu P. Dinu[1,2], **Ioan-Bogdan Iordache**[1], **Ana Sabina Uban**[1,2]
[1]Human Language Technology Research Center, University of Bucharest, Bucharest, Romania,
[2]Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
ldinu@fmi.unibuc.ro, iordache.bogdan1998@gmail.com, auban@fmi.unibuc.ro

Abstract

The identification of cognates and derivatives is a fundamental process in historical linguistics, on which any further research is based. In this paper we present our contribution to the SIGTYP 2023 Shared Task on cognate and derivative detection. We propose a multi-lingual solution based on features extracted from the alignment of the orthographic and phonetic representations of the words.

1 Introduction and Related Work

In this paper we describe our participation in the SIGTYP 2023 Shared Task on cognate and derivative detection.

As both the cornerstone of historical linguistics and a starting point of historical enquiry, automatic detection of cognates and derivatives provides access to a wide range of areas in social sciences (Campbell, 1998; Mallory and Adams, 2006; Mailhammer, 2015). Concrete examples of the usefulness of accurate prediction of cognates and cognate chains were previously mentioned in the works of Atkinson et al. (2005), Alekseyenko et al. (2012), and Dunn (2015) through linguistic phylogeny, which in turn can be applied to back tracing linguistic relatedness (Ng et al., 2010). Linguistic contact can also be inferred from such predictions (Epps, 2014), and this in turn can provide a better understanding and insight into the interaction of ancient communities (Mallory and Adams, 2006; Heggarty, 2015). While looking for similar patterns that regulate the cognitive mechanisms involved in semantic change, an extended view on cognate chains can be used as a basis for the identification of meaning divergence (Dworkin, 2006). The study of language acquisition (Huckin and Coady, 1999) as well as the challenging problem of removing false friends in machine translation (Uban and Dinu, 2020) would both benefit from an accurate understanding on the cognate pairings between any two related languages.

Today there is a vast volume of linguistic data that is yet to be analysed from a historical perspective (List et al., 2017). This illustrates the paramount importance of looking into automatic methods and algorithms that can accurately detect cognates and derivatives for both highly resourced and lowly resourced languages.

Recent years have seen a proliferation of techniques for automated detection of cognate pairs (Frunza and Inkpen, 2008; Ciobanu and Dinu, 2014; Jäger et al., 2017; Rama et al., 2018; Fourrier and Sagot, 2022). A lot of these techniques employ feature extraction from various orthographic and phonetic alignments used for training shallow machine learning algorithms in the supervised setting, or used along with clustering methods for the unsupervised approaches (Simard et al., 1992; Koehn and Knight, 2000; Inkpen et al., 2005; Mulloni and Pekar, 2006; Bergsma and Kondrak, 2007; Navlea and Todirascu, 2011; List, 2012; Ciobanu and Dinu, 2014; Jäger et al., 2017; St Arnaud et al., 2017; Cristea et al., 2021). Ciobanu and Dinu (2014) reported results on cognate detection for several Romance language pairs, in which cognate and non-cognate pairs are distinguished via features extracted from orthographic alignments that are used for training Support Vector Machines, with accuracies reaching as high as 87%.

Deep learning models for cognate detection and other similar tasks were mentioned in fewer studies. Siamese convolutional neural networks trained on character sequences for either the orthographic, or the phonetic representations of the words, and augmented with handcrafted features were shown to perform well when tested on cognate prediction for three language families, out of which the most proeminent one being the Austronesian family (Rama, 2016). Also, for borrowing detection Miller et al. (2020) employed deep learning architectures based on recurrent neural networks.

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 137–142

1.1 SIGTYP 2023 Task and Data

The SIGTYP 2023 competition includes two subtasks: supervised and unsupervised classification of word pairs into three different classes: cognates, derivatives, and neither. The dataset included 232,482 annotated word pairs in 34 languages, where each word pair was annotated with a language for each word, and with one of the three categories based on the relationship between the pair. The data was annotated based on Wiktionary.

2 Automatic Cognate Detection Experiments

2.1 Methodology

The models we experimented with were all multilingual, in the sense that we trained them on the whole dataset without any split with respect to the languages of the classified word pairs. We trained classical machine learning algorithms using various sets of handcrafted features. In order to improve overall performance, we also looked into training ensemble models using the best scoring algorithms.

2.2 Features

The models were trained using combinations of three types of features:

- graphic features, extracted from aligning the graphic form of the words in a pair
- phonetic features, extracted from a similar alignment, but for the phonetic transcriptions
- language features, represented as one-hot encodings for which pair of languages the words in an input pair come from.

For the graphic features, we started by preprocessing the input words and removing the accents. The Needleman-Wunch algorithm for sequence alignment (Needleman and Wunsch, 1970) was successfully employed in previous studies (Ciobanu and Dinu, 2019) for aligning and extracting features from the graphic representation of word pairs, in order to classify such pairs as cognates or non-cognates. Using a similar approach we were able to extract n-grams around alignment mismatches (i.e. deletions, insertions, and substitutions). Another aspect we borrowed from previous studies is that for a given value of n, we extract all such i-grams that have the length $i \leq n$.

As for an example of graphic features extraction, we can look at the pair constituted of the German word "hoch" and the Swedish word "hög", annotated as cognates in the training dataset, and both meaning "tall". For the preprocessed pair `(hoch, hog)` we obtain the following alignment: `($hoch$, $hog-$)`, where `$` marks the start and the end of the alignments and `-` represents an insertion, or deletion (depending on the direction we are considering). For a chosen value of $n = 2$, the extracted features are: `c>g`, `h>-`, `oc>og`, `ch>g-`, and `h$>-$`.

For phonetic features, we employ the same method, but this time on the phonetic representation of the input words, where one could have been identified (if we did not identify the phonetic representation of at least one word in the input pair, we consider no phonetic features for this pair). To obtain the phonetic representations we used the eSpeak library[1], version 0.1.8.

All these features along with the encoding of the input languages are vectorized using the binary bag of words paradigm, and correspond to the input representation for the various Machine Learning models we trained.

2.3 Supervised classification: Ensemble Model

Using various combinations of the features described above, we experimented with training a few different multi-class classification algorithms: Support Vector Machine, Naive Bayes, and SGD Classifier. In order to compare the performance of the trained models (with various hyper-parameters) and their corresponding feature combinations, we computed F1 scores obtained from three-fold cross validation using the whole training dataset.

Out of these models we select the top performing ones and we then train a stacking ensemble classifier. We also experimented with the number of models selected and assessed the enseble performance using three-fold cross validation as well.

2.4 Unsupervised classification: Clustering model

For the clustering approach, we employed the whole set of features (graphic features, phonetic features, and language encodings) and fitted a KMeans algorithm with the number of clusters set to 3.

[1] https://github.com/espeak-ng/espeak-ng

Model and Hyper-Parameters	n	graphic	phonetic	language	F1	Acc
SGD Classifier, loss: "hinge"	3	yes	yes	yes	**0.793**	0.921
SGD Classifier, loss: "modified_huber"	3	yes	yes	yes	0.791	0.921
SGD Classifier, loss: "modified_huber"	2	yes	yes	yes	0.783	0.916
Linear SVM, $C = 0.1$	3	yes	yes	yes	0.782	**0.923**
SGD Classifier, loss: "modified_huber"	3	yes	no	yes	0.781	0.916
SGD Classifier, loss: "hinge"	2	yes	yes	yes	0.781	0.914
SGD Classifier, loss: "log_loss"	3	yes	yes	yes	0.780	0.913
SGD Classifier, loss: "perceptron"	3	yes	yes	yes	0.775	0.910
SGD Classifier, loss: "hinge"	3	yes	no	yes	0.775	0.911
Linear SVM, $C = 1$	3	yes	yes	yes	0.782	0.917

Table 1: Top ten best performing models with respect to macro F1 score for the supervised task. Best hyper-parameters and feature combinations are also reported in this table. n represents the size of the considered alignment n-grams for graphic and phonetic features. Evaluation was done using three-fold cross validation on the training data

.

2.5 Hyperparameters and experimental details

For selecting the best base models to be combined into the stacking ensemble for the supervised approach, and also for selecting the model for the unsupervised task, we trained various machine learning models using the *scikit-learn* Python library. The list of models and their parameters is the following (note that if not said otherwise, all other hyper-parameters are set to the defaults specified in the 1.2.0 version of the library):

- Linear Support Vector Machine (`LinearSVC`): $C \in \{0.1, 1, 10\}$
- Multinomial Naive Bayes
- SGD Classifier: loss $\in$ {hinge, log_loss, perceptron, squared_hinge, modified_huber}.

We evaluate each such model using all combinations of graphic, phonetic, and language encoding features, and using various values for the size of considered alignment n-grams ($n \in \{1, 2, 3\}$).

Lastly we select the top performing N models based on cross validation scores and train a `StackingClassifier` on the whole training set. Furthermore, we cross validate these ensembles as well in order to determine the best N.

3 Results

3.1 Supervised Task

We report metrics computed via three-fold cross validation performed using the provided training dataset. We report the macro F1 score (the metric used in the task description for evaluation purposes) and the classification accuracy. Table 1 contains the metrics computed for the top 10 performing classification models, along with their choice of hyper-parameters and features.

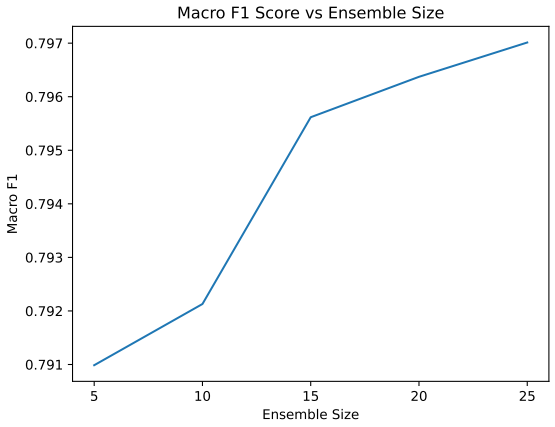

Figure 1: Computed macro F1 scores through three-fold cross validation for the supervised ensemble architectures trained using various numbers of base models.

We also tracked the performance of the ensemble architecture for various numbers of base models. As can be seen in figure 1, slight improvements are achieved when picking more models, although at some point this process shows diminishing returns and a longer time for training.

For the supervised submission, we chose the 25 models ensemble that displayed a 0.797 macro F1 score on the cross validation experiment, while for the unsupervised one, our KMeans model displayed a clustering score of 0.816.

4 Conclusions

In this paper we described our approaches for both the supervised and the unsupervised subtasks from the SIGTYP 2023 Shared Task on cognate and derivative detection. Our methods mostly rely on feature engineering powered by sequence alignments for both orthographic and phonetic transcriptions.

As we have seen from the results reported on the train labels, the combination of graphic and phonetic features seem to provide better performance than the models relying on one but not the other. One disadvantage is the lack of phonetic transcriptions for some of the low resource languages, which should be an important item in the long list of studies still needed for these type of languages.

Our submissions for the shared task yielded a macro F1 score of 0.83 for the supervised subtask, which was only 0.04 below the best reported result, and a 0.49 clustering accuracy for the unsupervised subtask, which was the best reported result and achieved a 30% improvement over the baseline.

For future work we are considering a qualitative analysis of the errors, in order to better understand on which language pairs our models were registering better results and where they struggled to provide accurate predictions.

Acknowledgments Research supported by the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, CoToHiLi project, number 108/2021, Romania.

References

Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337:957–960.

Quentin D. Atkinson, Russell D. Gray, Geoff K. Nicholls, and David J. Welch. 2005. From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103:193–219.

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 656–663.

Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of ACL 2014, Volume 2: Short Papers*, pages 99–105.

Alina Maria Ciobanu and Liviu P. Dinu. 2019. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai, and Ana Sabina Uban. 2021. Automatic discrimination between inherited and borrowed Latin words in Romance languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2845–2855, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Dunn. 2015. Language phylogenies. *The Routledge handbook of historical linguistics*, pages 190–211.

Steven N Dworkin. 2006. Recent developments in spanish (and romance) historical semantics. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*, pages 50–57.

Patience Epps. 2014. Historical linguistics and sociocultural reconstruction. In *The Routledge Handbook of Historical Linguistics*, pages 579–597. London: Routledge.

Clémentine Fourrier and Benoît Sagot. 2022. Probing multilingual cognate prediction models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3786–3801. Association for Computational Linguistics.

Oana Frunza and Diana Inkpen. 2008. Disambiguation of partial cognates. *Lang. Resour. Evaluation*, 42(3):325–333.

Paul Heggarty. 2015. Prehistory through language and archaeology. In *The Routledge Handbook of Historical Linguistics*, pages 598–626. Routledge.

Thomas Huckin and James Coady. 1999. Incidental vocabulary acquisition in a second language: A review. *Studies in second language acquisition*, 21(2):181–193.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *RANLP-2005, Bulgaria*, pages 251–257.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2000. Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm. In *Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, pages 711–715.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics Joint Workshop of LINGVIS and UNCLH*, pages 117–125.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, 12(1):1–18.

Robert Mailhammer. 2015. Etymology. In *The Routledge handbook of historical linguistics*, pages 441–459. Routledge.

James P Mallory and Douglas Q Adams. 2006. *The Oxford introduction to proto-Indo-European and the proto-Indo-European world*. Oxford University Press on Demand.

John E Miller, Tiago Tresoldi, Roberto Zariquiey, César A Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists. *Plos one*, 15(12):e0242709.

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2387–2390.

Mirabela Navlea and Amalia Todirascu. 2011. Using Cognates in a French-Romanian Lexical Alignment System: A Comparative Study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 247–253.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Ee-Lee Ng, Beatrice Chin, Alvin W. Yeo, and Bali Ranaivo-Malançon. 2010. Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *Int. J. of Asian Lang. Proc.*, 20(2):43–62.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying Cognate Sets Across Dictionaries of Related Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528.

Ana Sabina Uban and Liviu P. Dinu. 2020. Automatically building a multilingual lexicon of false friends with no supervision. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3001–3007. European Language Resources Association.