Computational approaches for protoword reconstruction

Alina Maria Cristea,[1] Anca Dinu,[1,3] Liviu P. Dinu,[1,2]
Simona Georgescu,[1,3] Ana Sabina Uban,[1,2] Laurențiu Zoicaș[1,3]

[1] Human Languages Technologies Research Center, University of Bucharest
[2] Faculty of Mathematics and Computer Science, University of Bucharest
[3] Faculty of Foreign Languages and Literatures, University of Bucharest

{alina.cristea,ldinu,auban}@fmi.unibuc.ro,
{anca.dinu,simona.georgescu,laurentiu.zoicas}@lls.unibuc.ro

Abstract

Given an input word, the task is to automatically produce its protoword. Having modern words in multiple sister languages, our goal is to automatically produce the Latin word from which they evolved. For example, having the modern word forms *capră* (Ro), *capra* (It), *chèvre* (Fr), *cabra* (Es), *cabra* (Pt), the goal is to automatically produce their common Latin ancestor, *capra*. We intend to apply the proposed methodology for cognate sets where the common Latin etymon is unattested. In our previous work [2,3,4], we developed a machine learning method for automatically producing related words based on sequence alignment and sequence labelling. We focused on reconstructing protowords, but we also addressed two related sub-problems, producing modern word forms and producing cognates. To align pairs of words, we employed the Needleman Wunsch [7] global alignment algorithm, which has been successfully used in natural language processing and computational biology. We used words as input sequences and a basic substitution matrix, which gives equal scores to all substitutions, disregarding diacritics (e.g., we ensure that *e* and *é* were matched). For the machine learning part we used sequence labelling, an approach that has proved useful in generating transliterations [1, 6]. We used a Conditional Random Fields model which uses the alignment of the related words as input in order to learn patterns and to infer the form of the related words [5]. We further employed ensemble-based aggregation for combining and re-ranking Latin protoword reconstruction from multiple source languages, in order to improve performance. We trained a model for each modern language and combined their results, in order to obtain Latin protowords with higher accuracy. Each model produced, for each input word, an n-best list of possible protowords. To combine the outputs of the model, we investigated multiple fusion methods based on the ranks in the n-best lists and the probability estimates provided by the individual models for each possible production. Given a cognate set, we combined the previously obtained n-best lists to compute a joint n-best list that leveraged information from all modern languages. We applied our method to multiple data sets, showing that our approach improves on previous results, also having the advantage of requiring less input data, which is essential in historical linguistics, where resources are generally scarce. In this paper, we intend to develop our previous work in several directions.

Phonetic versus graphic. Of the Romance languages, only contemporary French presents differences – often considerable – between the graphic and the phonetic form of the huge majority of words. Contemporary French cognates will have to be approached in their phonetic form. Furthermore, we intend to compare them with their older versions (Old French used a predominantly phonetic spelling). An important support is the recourse to the words with French and Latin origin from the English vocabulary (representing about 60% of the total English vocabulary).

From phonetics to morphology. The computational approach currently operates with the lemmatized forms of words (the nominative nouns, the masculine adjectives, the infinitive of verbs). We are considering extending this approach to the entirety of the paradigms. In this way, for Romanian we could find out how, for example, the strong or weak conjugations of the verbs were selected, how certain nouns passed from one gender to another, and so on.

Towards semantics. Finally, we plan to prepare the stage in which the analysis will also tackle the semantic dimension of the vocabulary.

Keywords

Protowords, Romance languages, sequence labelling.

Acknowledgements

This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108, CoToHiLi, within PNCDI III.

References

[1] Ammar, Waleed, Chris Dyer, and Noah A.Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proceedings of the 4th Named Entity Workshop*, pages 66–70.
[2] Ciobanu, Alina Maria and Liviu P. Dinu. 2019. Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, vol. 45, No. 4, pages 667–704.
[3] Ciobanu, Alina Maria and Liviu P. Dinu. 2018. Ab Initio: Latin Proto-word Reconstruction. In *Proceedings of COLING 2018*, pages 1604–1614.
[4] Ciobanu, Alina Maria, Liviu P. Dinu, and Laurentiu Zoicas. 2020**.** Automatic Reconstruction of Missing Romanian Cognates and Unattested Latin Words. In *Proceedings of LREC 2020*, pages 3226–3231.
[5] Dinu, Liviu P. and Alina Maria Ciobanu. 2017. Romanian word production: An orthographic approach based on sequence labeling. In *Proceedings of CICLing 2017, Revised Selected Papers, Part I*, pages 591–603.
[6] Ganesh, Surya, Sreeharsha Yella, Prasad Pingali, and Vasudeva Varma. 2008. Statistical transliteration for cross language information retrieval using HMM alignment model and CRF. In *Proceedings of CLIA 2008*, pages 42–47.
[7] Needleman, Saul B. 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*, vol. 48, No. 3, pages 443–453.