CoToHiLi: computational tools for historical linguistics

CoToHiLi: herramientas computacionales para la lingüística histórica

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

[1] *University of Bucharest*

[2] *Research group: Human Language Technologies Research Center, University of Bucharest*

Abstract

This project represents a computational framework for historical linguistics. The general purpose of the CoToHiLi project is to integrate expert knowledge and computational power to address cognate identification, cognate-borrowing discrimination, Latin protoword reconstruction and semantic divergence. The goal of the project is twofold: 1) to automate certain parts of the traditional work-flow of the comparative method (such as the collection of data or the automatic alignment based on predefined or inferred rules), and 2) to bring new insights or avenues of investigation, which might not be easily accessible otherwise (e.g., the automatic identification of patterns and regularities in large amounts of data). The project will provide tools for the main Romance kernel group (French, Italian, Portuguese, Romanian, Spanish), as well as Latin. The methodologies and computational tools proposed could also serve as a basis for further development for other comparable language families, including less studied languages, with scarce resources available.

Keywords

Historical linguistics, cognates, semantic divergence.

1. Introduction

The general purpose of the CoToHiLi[1] project is to integrate expert knowledge and computational power to address the following topics: cognate identification, cognate-borrowing discrimination, Latin proto-word reconstruction and semantic divergence. Our project is focused on the Romance languages (French, Italian, Portuguese, Romanian, Spanish), and will provide tools for the main Romance kernel group and for Latin. The duration of the project is 3 years, starting from January 2021.

The research problems that we address are significant on multiple levels. From a scientific point of view, any advance in historical linguistics is of paramount cultural importance, being inherently connected with human history ("each word a history", cf. [1]). Longobardi (LanGeLin project, 2012-2018) explored the potential correlation of genetic and linguistic distances, starting from what he called Darwin's last challenge: "If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one" (see also [2]). Given that the socio-economical and cultural factors are some of the motivations for borrowing from one language to another [1, 3], the topic of this research project facilitates reconstructing certain aspects related to society and culture for groups of people speaking a given proto-language, and gaining insights into their past social interactions and into their social and cultural practices [3]. Moreover, establishing the direction and source of borrowing is important to our understanding of the social relations between the groups involved. From a technological perspective, as linguistic change is the most visible at the lexical and semantic level, computational tools can be designed to serve both aspects, for instance to automatically identify related words and to assess the semantic change. Even though historical lexicology has leveraged technological advances, and some pioneering work was initiated on various steps of the work-flow (cognate identification, proto-word reconstruction), historical semantics has not sufficiently benefited from the advances in computer

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain

alina.cristea@fmi.unibuc.ro (A. M. Cristea); anca.dinu@lls.unibuc.ro (A. Dinu); ldinu@fmi.unibuc.ro (L. P. Dinu); simona.georgescu@lls.unibuc.ro (S. Georgescu); auban@fmi.unibuc.ro (A. S. Uban); laurentiu.zoicas@lls.unibuc.ro (L. Zoicaș)

[1] This is the project's web page, where we will include our results and updates: https://nlp.unibuc.ro/projects/cotohili.html.

science. Yet, by drawing special attention to the semantic divergence occurring in pairs of cognates, we could both take a few steps forward towards a unitary theory of semantic change, and improve practical applications such as automatic translation systems or language e-learning systems, aware of false friends and related phenomena.

2. Objectives

The innovation of the project consists in integrating linguists' knowledge with new computational methods in a unified framework, to address important problems from historical linguistics, enabling experts to provide input and feedback throughout the whole development process, in the pre-processing, annotation, feature engineering, training and evaluation phases:

2.1. Identification of related words

We aim at going one step further than the current state-of-the-art methods by: a) proposing a more in-depth analysis, by identifying the direction of the borrowings and b) automatizing the whole process as a pipeline that, given a pair of input words, provides an automatic analysis regarding the relationship between them [4, 5, 6].

2.2. Latin proto-word reconstruction

To improve previous results, we intend to use more recent techniques [7], such as conditional random fields (CRF) for sequence labelling and deep learning, in particular character-level neural networks. The alignment technique, which stands at the foundation of our approach, will be improved by an heuristic for choosing the best alignment. We also address the challenging problem of multiple alignment (finding an alignment for more than two words), in order to be able to extract knowledge from cognate sets in multiple languages. Another promising line of research is to make use of the more recent Latin resources, such as The Latin Diachronic Database [8].

2.3. Diachronic semantic divergence

Semantic change is a continuous and complex process ([1] presents no less than 11 types of semantic change), which has been recently studied in the context of distributional semantics theory. Vectorial representations of word meaning (word embeddings) have been used for tracking semantic shifts across different time periods, especially for English. Our aim is to track the semantic change of words in Latin and across multi-languages, in the Romance language family, for the first time, with the substantial purpose of looking for common patterns characterizing the overall semantic divergence cases. Additionally, we intend to explore the statistical properties of the word embedding vectorial spaces [9, 10].

3. Impact

The methodologies and computational tools we propose could extend their applicability not only to various linguistic branches in the Indo-European family, but also to less studied languages or linguistic families. Such advances could provide new answers in historical and social sciences, given that lexical and semantic change is a key source of clues regarding both the dynamics of cultural interactions between groups in the past, and the technological innovations and exchanges that have taken place across space and time [3]. Moreover, the semantic data provided by the CoToHiLi project could be of great help for the cognitive sciences and neurosciences [11], to the extent that they can offer a new perspective on our brain mechanisms.

As for the socio-economic impact of the CoToHiLi project, in the context of the increasing number of attempts to create automatic tools designed for linguistic comprehension, our computational devices could support Romance intercomprehension by bringing into light the common linguistic features, as well as the semantic relations between the Romance cognates or borrowings. Such an advance can prove its usefulness in the constant efforts to improve the automatic translation systems.

4. Methodology

For the first two objectives, our methodology is focused on two main aspects: creating clean datasets and developing computational methods for achieving the proposed research tasks. For the Romance languages there are already some existing resources (for cognates, for borrowings and for proto-word reconstruction), but they are scattered, incomplete, or with uncertain availability (cf. [12, 13]). Thus, datasets do not have to be built from scratch, but the data need to be harmonized, verified and enhanced where necessary, in order to become a benchmark in the domain. By using computational tools, corroborated by the direct intervention of classical linguists, we have already built a significant part of the database, representing the starting point for

the computational methods that are being developed. We have continued with the alignment of word pairs. Given the lack of an unanimously accepted alignment method [13, 14], we confront a semi-automatic manner of choosing the alignment with the knowledge of classical linguists, in order to establish an heuristic capable of making the best choice. From the alignment, we extract features for machine learning models. We improve current existing computational methods with linguistic features provided by experts. We develop a machine-learning classifiers (using support vector machines), sequential models (using CRF and neural networks) and ensemble techniques. Moreover, we experiment with new ensemble techniques, to improve the overall performance by combining results from multiple sister languages. We are currently working with the orthographic form of the words, while for Romanian, Spanish and Italian we are planning to also use the phonetic transcription.

For the third objective, in order to identify semantic shifts across time periods as well as languages, we leverage vector space representations of meaning, or word embeddings, relying on traditional models such as word2vec and FastText [15, 16], as well as experimenting with state-of-the-art language models such as BERT [17]. The method consists of building vectorial semantic representations for the words in each of the target languages, based on the multilingual corpora, and then obtaining a shared multilingual semantic space. This will allow us to compute semantic distances between cognates as well as analyze the statistical and the linguistic properties of words whose meanings have diverged. The available corpora are unequal from one language to another; for instance, the Royal Spanish Academy provides an exhaustive diachronic corpus of its language, whereas for Romanian we only have access to a scarce data-base, composed of a fairly limited number of old texts. In order to ensure the accuracy of our analysis, in this stage of the project, we use mainly lexicographic resources, as well as data-bases built for the contemporary stage of each language (such as multilingual Wikipedia[2]).

5. Current Results

For the first two objectives, we have started building datasets of cognates and borrowed words for the Romance languages [18]. This first step relies on dictionaries that contain etymological information (e.g., for Romanian we use 13 dictionaries available in digital format). We have proposed a new method automatically discriminating between inherited and borrowed Latin words. We have introduced a new dataset and investigated the case of Romance languages - where words directly inherited from Latin coexist with words borrowed from Latin -, and explored whether automatic discrimination between them was possible. An initial trial was to automatically predict whether a word was inherited or borrowed by simply taking into account its intrinsic structure, given that borrowed words are presumably less eroded than inherited ones, subject to historical sound shifts. We then took a step farther and employed n-gram character features extracted from the word-etymon pairs and from their alignment, which led to considerably better results [6].

For the third objective, a first step has been taken with the investigation of the semantic divergence of cognate pairs in English and Romance languages. To this end, we introduced a new curated dataset of cognates in all pairs of those languages. We described the types of errors that occurred during the automated cognate identification process and manually corrected them. Additionally, we labeled the English cognates according to their etymology, separating them into two groups: old borrowings and recent borrowings. On this curated dataset, we analysed word properties such as frequency and polysemy, and the distribution of similarity scores between cognate sets in different languages. We automatically identified different clusters of English cognates, setting a new direction of research in cognates, borrowings and possibly false friends analysis in related languages [10, 19].

6. Conclusions

Drawn within a computational framework, the CoToHiLi project addresses key concerns of historical linguistics centered on the Romance languages, such as cognate identification, cognate-borrowing discrimination, Latin protoword reconstruction and semantic divergence, towards which we have taken a few steps forward by performing various experiments. At this stage of the project, we analyze only the main five Romance languages (French, Italian, Portuguese, Romanian, Spanish), but as we advance we intend to include other Romance idioms as well. We predict that the methodologies and computational tools proposed will also serve as a basis for further development for other comparable language families, including less studied languages, with scarce resources available.

[2] https://github.com/facebookresearch/MUSE

Acknowledgments

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

References

[1] L. Campbell, Historical Linguistics. An Introduction, MIT Press, 1998.

[2] N. Ritt, Selfish Sounds and Linguistic Evolution. A Darwinian Approach to Language Change, Cambridge University Press, 2004.

[3] P. Epps, Historical linguistics and sociocultural reconstruction, in: The Routledge Handbook of Historical Linguistics, London: Routledge, 2014, pp. 579–597.

[4] A. M. Ciobanu, L. P. Dinu, Automatic detection of cognates using orthographic alignment, in: Proceedings of ACL 2014, Volume 2, 2014, pp. 99–105.

[5] A. M. Ciobanu, L. P. Dinu, Automatic discrimination between cognates and borrowings, in: Proceedings of ACL 2015, 2015, pp. 431–437.

[6] A. M. Cristea, L. P. Dinu, S. Georgescu, M. Mihai, A. S. Uban, Automatic discrimination between inherited and borrowed latin words in romance languages, in: Findings of EMNLP 2021, 2021, pp. 2845–2855.

[7] A. M. Ciobanu, L. P. Dinu, Ab initio: Automatic Latin proto-word reconstruction, in: Proceedings of COLING 2018, 2018, pp. 1604–1614.

[8] T. Spinelli, The latin diachronic database: a new digital tool for the study of latin, in: Recent Advances in Digital Humanities: Romance Language Applications, Peter Lang, 2022, forthcoming.

[9] A.-S. Uban, A. M. Ciobanu, L. P. Dinu, Cross-lingual laws of semantic change, Computational approaches to semantic change 6 (2021) 219.

[10] A. S. Uban, A. Cristea, A. Dinu, L. P. Dinu, S. Georgescu, L. Zoicas, Tracking semantic change in cognate sets for English and Romance languages, in: Proceedings of LChange 2021, 2021, pp. 64–74.

[11] D. Poeppel, D. Embick, Defining the relation between linguistics and neuroscience, in: The Routledge Handbook of Historical Linguistics. Twenty-First Century Psycholinguistics, Four Cornerstones, New York, Routledge, 2017, pp. 103–118.

[12] A. Bouchard-Côté, D. Hall, T. L. Griffiths, D. Klein, Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change, PNAS 110 (2013) 4224–4229.

[13] A. M. Ciobanu, L. P. Dinu, Automatic identification and production of related words for historical linguistics, Computational Linguistics 45 (2019) 667–704.

[14] G. Kondrak, A new algorithm for the alignment of phonetic sequences, in: Proceedings of ANLP 2000, 2000, pp. 288–295.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of NIPS 2013, 2013, pp. 3111—-3119.

[16] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, TACL 5 (2016) 135–146.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL 2019, 2019, pp. 4171–4186.

[18] A. M. Cristea, A. Dinu, L. P. Dinu, S. Georgescu, A. S. Uban, L. Zoicas, Towards an Etymological Map of Romanian, in: Proceedings of RANLP 2021, 2021, pp. 315–323.

[19] A. S. Uban, L. P. Dinu, Automatically building a multilingual lexicon of false friends with no supervision, in: Proceedings of LREC 2020, 2020, pp. 3001–3007.