Using a vector-based phonological representation for the cognate detection task

Ionov , Max (Goethe University in Frankfurt)

In this talk we present an ongoing work on searching phonologically similar words in related languages in and across dictionaries. This is a part of a larger project devoted to unraveling both synchronic and diachronic lexical connections in related and contact languages.

More specifically, we discuss problems of converting word representation found in different dictionaries to a single unified standard — a set of binary vectors built based on a PHOIBLE dataset (Moran et al. 2014).

The task of detecting phonologically similar words is a part of a well-established field of studying and modeling linguistic diversity and language evolution. Over the years its methods were applied in the fields of dialectometry (Heeringa et al. 2006) and historical linguistics (List & Moran 2013, List et al. 2017).

When dealing with a combination of different lexical resources, the first problem that arises is a problem of standartization of these resources. Apart from the problem of unifying formats of resources in general, it also means that all resources should follow the same orthography (for lexicographic research) or a scheme for phonetic transliteration (for research that involved phonology).

In case of low-resourced languages this can be a challenge, since there can exist multiple orthographies for one language, or different resources can use different transliteration schemes.

Our approach includes the conversion of our resources to an IPA representation in a scalable way which allows us to include new resources in new languages rapidly in several stages, where each stage increases the quality of transliteration. In our talk, we report the impact of increasing transliteration quality on our task.

Even though there are various well-known approaches for the task of detecting phonologically similar words, starting from simple yet popular minimum edit distance approach (Holman et al. 2011) to more sophisticated approaches like LexStat (List 2012), every approach has its weak spots: simple approaches lack linguistic insight, whereas complex approaches are computationally ineffective.

We present a new, computationally effective approach to this task in which each character of each word is represented as a binary vector of phonological features. We use the PHOIBLE dataset as an inventory of phonological features which makes this approach theory-neutral and reusable. We demonstrate that this approach shows reasonable quality and due to its high computational efficiency, it can be applied to compare a lot of lexical resources at the same time.

We further show that the impact on differences of two sounds in some features are more important for the task than the in some others. We investigate those differences and propose two ways of incorporating this into the algorithm: by using linguistic insight to set the importance of each feature, and by automatically extracting these importances from the data.

We show that by using those weights we can improve the quality of our method.

References

Heeringa, W., Kleiweg, P., Gooskens, C., and Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In Proceedings of the Workshop on Linguistic Distances, LD ’06, pages 51–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Holman, E. W., Brown, C. H., Wichmann, S., Mller, A., Velupillai, V., Hammarstrm, H., Sauppe, S., Jung, H., Bakker, D., Brown, P., Belyaev, O., Urban, M., Mailhammer, R., List, J.-M., and Egorov, D. (2011). Automated dating of the worlds language families based on lexical similarity. Current Anthropology, 52(6):841–875.

List, J.-M. (2012). Lexstat. automatic detection of cognates in multilingual wordlists. In Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources, pages 117– 125, Stroudsburg.

List, J.-M., Greenhill, S. J., and Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. PLOS ONE, 12(1):1–18.

List, J.-M. and Moran, S. (2013). An open source toolkit for quantitative histori- cal linguistics. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13–18, Sofia, Bulgaria. Association for Computational Linguistics.

Moran, S., McCloy, D., and Wright, R.,(2014). PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.