Dialectometry with topic models

Olli Kuparinen & Yves Scherrer

Dialectometrical analyses are most often based on dialect atlases compiled systematically over the last 150 years. Although such data have many benefits, such as good geographical coverage, the atlases tend to be old and only exist for some languages. Dialect corpora collected in semi-directed interviews have thus become important data sources more recently. In this study, a topic model approach to discover differences between dialects from interview data is presented as an alternative to analyses based on dialect atlases.

Topic models are often used to find latent semantic structure in a collection of text documents. Co-occurring words (e.g., *dog, bone, fetch*) in multiple documents are assumed to constitute a topic (dogs). Because the interest has been in semantic similarity, the model has been used on normalized and lemmatized language data to prevent modeling the same word in its different forms.

For dialects, the interest lies in these differing forms (i.e., structural differences). This means the model can be used on phonetically transcribed data directly, and it will find components that correspond to different dialectal features or combinations of them. It is then easy to do traditional dialectometrical analysis, and see which components are used where. Topic models have been used on dialect data before, but the analysis has focused on lexical variation in social media (Eisenstein et al. 2010) or the linguistic features have been searched for before the modeling (Kuparinen et al. 2021). The approach presented here uses transcribed data directly, without pre-processing steps.

The approach is tested on corpora from three languages: Finnish, Norwegian and Swiss German. The datasets include Samples of Spoken Finnish (Institute for the Languages in Finland 2014), Norwegian Dialect Corpus (Johannessen et al. 2009) and Archimob Corpus (Scherrer et al. 2019), all of which include interview transcriptions from several locations in their respective areas. The modeling is tested on different levels of the transcriptions: complete words, character n-grams (sequences of characters) and automatically segmented words. The model used in the study is non-negative matrix factorization (NMF; Lee & Seung 1999). It returns two distributions: one presenting the components over features (which features are used in which components) and one presenting the documents over components (which components are used in which documents).

The results are very promising and show that such a model can find important dialectal differences directly from interview transcriptions. When modeling based on complete words, the model finds differences in frequent words, such as personal pronouns (*minä, mää, mie* ‘I’ in Finnish) or negation (*itte, ikkje, ittje* ‘not’ in Norwegian). When using the character level as input, the model discovers phonological differences, such as diphthong opening or reduction (*m**iä**s* ‘man’, ***ae**na* ‘always’) in Finnish and l-vocalization in Swiss German (a**u**so ‘so’). The automatic segmentation of data is used to combine both levels and find important words as well as important character sequences. The languages differ somewhat, with Finnish and Norwegian producing clearer divisions based on complete words, and Swiss German based on character-level differences.

References

Eisenstein, J., B. O’Connor, N. A. Smith & E. P. Xing. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*: 1277–1287.

Institute for the Languages of Finland (2014). *Samples of Spoken Finnish* [speech corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2020112937

Johannessen, J. B., J. Priestley, K. Hagen, T. A. Åfarli & Ø. A. Vangsnes. (2009). The Nordic dialect Corpus - an Advanced Research Tool. In Jokinen, K. and E. Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series* Volume 4.

Kuparinen, O., J. Peltonen, L. Mustanoja, U. Leino & J. Santaharju. (2021). Lects in Helsinki Finnish. A probabilistic component modeling approach. *Language Variation and Change*, 33(1): 1-26

Lee, D. D. & H. S. Seung (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755): 788-791.

Scherrer, Y., T. Samardžić & E. Glaser (2019). Digitising Swiss German - How to process and study a polycentric spoken language. In: *Language Resources and Evaluation* 53(4).