Automatic detection of spelling variations in less-resourced dialect and languages

Yo Sato, Satoama Language Services

A language or dialect of a relatively small community may not have an established orthography. The writers thus exercise certain license, and as a result, one may find wide variations in spelling for the same word. Such variations may be problematic not just educationally or socially but also technologically, since although data are increasingly available as people come to use their own tongue in social media, it still may suffer from fragmentations. Establishing an orthography, however, would be a huge undertaking. In this work, therefore, we present computational techniques to deal with such variability, *not* by way of prescriptive 'normalisation', but by clustering, i.e, to discover automatically which variants belong to the same word, without commitment to their correctness. Our method, which proceeds in an unsupervised manner, would provide a solution to the fragmentation problem and could function as a preliminary step towards normalising criteria.

Unsupervised methods need some prior assumptions, and our method relies on two, both plausible in many small language/dialect communities: that the set of available phonemes is established, and that the writers borrow the established orthographies of neighbouring languages.

As a machine learning method we use the sequence-to-sequence architecture[1] in a 'zero-shot' transfer learning paradigm[2]. We first get the machine to learn, given the pairs of words and their corresponding phoneme sequences in the neighbouring languages, their association patterns. We then add the data from the target dialect/language, as well as its phoneme set, and make it learn a new association between them[3]. The homonym differentiation is made on the basis of word embedding similarities[4].

To demonstrate the performance of our method, we use Twitter corpora collected from two linguistic communities, Swiss German and Limburgish.

References

[1] Cho, Kyunghyun, D Bahdanau and Y Bengio (2014). Neural machine translation by jointly learning to align and translate, *Proceedings of the 3rd International Conference of Learning Representations*.
[2] Radford, Alec, J Wu, R Child, D Luan, D Amodei and I Sutskever (2019). Language models are unsupervised multi-task learners.
[3] Tursun, Osman and R. Cakici (2017). Noisy Uyghur text normalization, *Proceedings of the 3rd Workshop on User-generated Text*
[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv: 1810.04805