Automatic Learning of Common Representations of Related Languages with No Prior Linguistic Knowledge

As part of the Digital Language Typology (DLT) project, we are interested in studying raw language data from languages for which resources and tools typically used for language analysis are not available. In the Discovery Group in the Department of Computer Science, we have been attempting to develop techniques to learn about connections between two languages from only raw, low-level data, without any prior knowledge about the languages concerned. Whilst some knowledge about, for example, historical connections with other languages or formal characteristics of a language are often known, what exactly we know and how we can use that knowledge depends heavily on the specific language in question. We have therefore addressed the question: how much can we learn about associations of commonalities between a pair of languages known to be quite closely related from fully automatic analysis, making as few assumptions about the languages as possible?

For example, Swedish and Danish are closely related and share many cognates and loanwords. They use quite a different vocabulary of sounds, but the relationships between words in one language and their corresponding cognates in the other are somewhat systematic. If we could use an algorithm or a statistical model to capture these relationships, the result could be useful in, among other things, assisting speakers of one language to learn the other. Whilst the applications to Swedish and Danish are limited, such tools could be of great value if applied, for example, to Finnish and Veps, which has only a couple of thousand speakers.

We have begun by building statistical models of *phonetic transcriptions* of language. In our preliminary work, we use an existing tool to produce the transcriptions automatically from text; in future, we plan to apply exactly the same analysis techniques to transcriptions extracted directly from speech data. For the purposes of assessing the usefulness of our models, we have run them on data from pairs of related languages that are well studied, such as Finnish and Estonian, or Swedish and Danish. This allows us to check whether the characteristics found by the models correspond to known connections between the languages before we apply the methods to less studied languages.

Our first step is to learn how to organise the sounds of the two languages in question, potentially discovering sounds in one language that correspond in their usage to different sounds in the other, or establishing, for sounds that exist in only one language, which sounds of the other they resemble most in their usage in the words of the language. We might, for example, discover that the Danish *a* sound is often used in contexts where Swedish uses ɔ (å), leading to a close correspondence between these two. To do this, we make an assumption to start with that any sound in Danish is equally likely to be related to any sound in Swedish. This allows for the possibility that the two languages might have undergone radically different sound changes and even, in the implausible extreme case, not use any of the same sounds. This means that an *da:a*, the *a* sound in Danish, is treated as distinct from *sv:a*.

Our model uses the idea of *distributional similarity*, as applied in Natural Language Processing to learn similarities between words – that words that tend to appear in similar linguistic contexts (e.g. other words in the same sentence) often have a closely related meaning. We apply the same idea to the phones in phonetic transcriptions: we characterize them by the the phones that tend to appear immediately before or after them. The main contribution, however, is that we apply this intuition in a circumstance where the contexts (surrounding phones) never overlap between the observations in the two different languages, since we treat the sounds as language specific.

We use a neural network which takes as input short sequences of phonemes and predicts whether the sequence is a real sequence from one of the languages, or a randomly chosen distractor. The part

of the network that makes the prediction is shared between the two languages and each (language-specific) phone is represented in a high-dimensional vector space. The hope is that, as the network is trained, seeing a mixture of sequences from the two different languages, it learns to share information in the prediction function and learn vectors for the phones such that, if a Danish phone and a Swedish phone tend to appear in similar contexts, they will be close in the vector space, even though the contexts are exclusively Danish and Swedish respectively (also represented by learned vectors).

The results show that the model is effective in learning vectors that correspond to many expectations about how the languages correspond to one another. In many cases, a phone gets placed in the vector space very close to the same phone in the other language. In other cases, clusters of phones are mixed up, such as a number of vowels in Danish and Swedish, suggesting a less consistent pattern of usage across the languages. The model produces similar results for other language pairs.

After these extremely promising results, we are now considering how to build on the learned representations to extract further information about the relationships between the languages. One possibility is to use the learned similarities to help identify related words and thereby learn more specific patterns of how related words are rendered differently in the two languages. We also plan to investigate the application of the same technique to speech data, without the need for human phonetic transcription or transcription systems with hand-specified rules. Ultimately, we aim to apply the models to data from much less studied languages.