

Toward Automated Annotation: Machine Learning for Language Documentation

Linguists documenting endangered languages work under the immense time pressure inherent in the endangered status of the languages. One result of this time pressure is that the first concern of the documentary linguist is precisely to record language data while there are still communities of speakers to work with. Consequently, a major focus of technological developments for documentary linguistics to date has been on tools which facilitate the transformation of spoken language data into easily-accessible archives of sound and video recordings.

The work of *analyzing* endangered languages is subject to a second sort of time pressure: they take too much time. Grammatical annotation and interlinearization of transcribed language data are time-consuming, labor-intensive pursuits requiring linguistic training and expertise. In addition, much of the time involved in annotation is devoted to tedious, redundant tasks. It is time for us to devote serious effort to improving tools both for annotation and for the theoretical analysis which relies on this annotation.

Recent developments in Natural Language Processing (NLP) have a key role to play in decreasing the human time and effort required for the annotation phase of the documentation/description process. It is not unreasonable to expect that we could develop a suite of tools which will give documentary linguists a straightforward, user-friendly way to take advantage of these developments.

The advantages I have in mind come from machine learning methods in NLP. Over the last two decades, machine learning has revolutionized computational linguistics. Statistical methods and machine learning techniques have enabled researchers to build robust and accurate tools for linguistic analysis more rapidly than was previously possible. One clear advantage of machine learning approaches is that, given enough data, very good results can be achieved by a small number of people working over a relatively short period of time. An obvious disadvantage of these methods is the need for annotated data on which to train the system. Data scarcity is precisely one of the issues hindering computational work on less-studied languages.

This is where active learning techniques come in. The term ‘active learning’ in the context of machine learning refers to a bootstrapping process by which a learning agent interacts with a human informant to produce progressively better models of the phenomenon under discussion. In learning morphological segmentation, for example, the agent would first be trained on a small amount of data. From this data the agent produces a preliminary model of the morphology of the language. The underlying mechanism for producing the model is (roughly) statistical analysis of the training data followed by selection of the analyses judged to be more probable than other possible analyses. In an active learning approach, the system assigns a confidence measure to each segmentation it produces, and we select the segmentations with the lowest confidence measures to be returned to a human informant for correction, as these are the maximally-informative examples for refinement of the model. The system is then retrained on corrected data, and the process is repeated in order to produce the best possible model. Active learning is a viable technique for dealing with problems of data scarcity.

I believe that this and other techniques from computational linguistics have great potential to reduce the time required to get from transcribed data to grammatical analysis. The faster the data can be annotated with lexical, morphological, and part-of-speech information, the sooner the work of linguistic analysis can begin.