Research areas: computational linguistics, field linguistics
Abstract category: 15 minute

A tool for semi-automatic creation of questionnaires for field work or machine translation

Problem: Resources like the Comrie and Smith (1977; *Lingua*) *Questionnaire* for linguistic description or Payne's (1977) *Describing Morphosyntax* can serve as invaluable aids to field linguistsm since they enumerate a range of different grammatical phenomena which need to be examined when doing documentary work on underdescribed languages. However, an important aspect of elicitation which is not well supported at present is the construction of sentences in the working language which, when translated into the language being documented, are likely to reveal its grammatical properties. This paper describes a novel, machine-based method for dynamically creating a set of elicitation sentences for particular grammatical phenomena, which is sensitive to already discovered properties of a given language. This system has applications in field linguistics and machine translation.

Background: Avenue (Probst et al. 2002) is a novel machine translation system that automatically learns translation rules from data that is provided by informants. A central component of Avenue is a questionnaire called an Elicitation Corpus (EC). The EC is presented to informants via an interface called the Elicitation Tool (see figure).

The Method: This paper discusses a method for automating the creation of elicitation corpora for Avenue which can be extended to support the creation of corpora for field linguists. The first step is for a linguist to enter a set of features and values in a database. The features cover a range of functional parameters, e.g., temporal relations, utterance types, and argument roles. The second step is for the linguist to specify how different features can be combined (for example, combinations of person and number can be combined with different argument roles). Finally, features are automatically configured into feature structures from which elicitation sentences can be generated (in English or Spanish). The automatic generator is a unification-based system similar in spirit to Lexical Functional Grammar.

Crucially, the corpus generator can be made aware of known grammatical properties of a language, which could prove relevant for the construction of the EC. This is particularly important when the elicitation process is administered by the elicitation tool, which should not try to elicit grammatical phenomena that have been established as not being grammatically marked in a given language.

Implications and research issues: The EC system provides an example where the methods of computational linguistics can be directly applied to field linguistics. It was designed as part of Avenue to assist in automating the production of machine translation systems for new languages, especially indigenous languages, which do not typically have the large electronic corpora or trained computational linguists that are normally required to build an MT system. However, by incorporating a relatively high degree of linguistic sophistication into the system, it can be readily adapted for use by field linguists.

Probst, K., Levin, Peterson, Lavie, and Carbonell (2002) "MT for Minority Languages Using Elicitation-Based Learning of Syntactic Transfer Rules". *Machine Translation*, 17:4, 245-270.

Word count: 483