

Automatically Extracting Bilingual Dictionaries from Language Data

PROBLEM STATEMENT: Comprehensive bilingual dictionaries are crucial elements of applications from machine translation to pedagogy to language documentation. Building these dictionaries manually, however, requires a great deal of human time and effort, which is expensive and -- especially in the case of under-documented languages -- often scarce. Automating this task would allow for more rapid language documentation and would support the development of multilingual NLP systems for more diverse language sets. A discussion of this issue would address potential and existing applications, challenges, and resources for automatically extracting bilingual lexicons from (potentially limited) language data.

APPLICATIONS: A system that could automatically extract word-to-word mappings between language pairs could be used by documentary linguists for more rapidly generating glosses and word lists, and by computational linguists for generating lexicons for machine translation. Other applications may exist that could inform the development of an automatic lexicon-extraction system.

CHALLENGES: Some challenges for developing an automatic-lexicon-extraction system include resource scarcity, methods for discovering cross-language word alignments, and methods for evaluating system output. How can word-alignments be learned automatically from text? Can semi-supervised methods be developed that bootstrap from known word alignments or language patterns? How reliable would the output of such a system be, and how could its reliability be tested? Short of a fully-automated solution, a system that required some human input for verification or bootstrapping could still relieve a lot of the initial work load.

RESOURCES: Potential resources for extracting bilingual word pairs are parallel and comparable corpora, as well as existing bilingual word lists. Parallel corpora are language data for which there are direct translations into other languages, and while these are conceivably the most useful type of data for lexicon extraction, they are not always available for under-documented languages. Examples of common parallel corpora for less-documented languages are Bible translations or translations of narratives gathered in the field. Comparable corpora are texts that are not necessarily direct translations of each other, but cover similar topics; an example are news stories in different languages that discuss similar events. Extracting word-alignments from comparable corpora may not be as easy as from parallel corpora, but comparable corpora are often more abundant and easier to obtain. Other language data, such as existing glosses and lists of word pairs, may be used to bootstrap word-pair discoveries from these types of resources.