Get It Where You Can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation

Mary S. Neff, Brigitte Bläser, Jean-Marc Langé, Hubert Lehmann, Isabel Zapata Domínguez[1]

The acquisition and maintenance of lexical entries for MT can be a costly piece of work. While estimates are in the order of half an hour for creating one lexicon entry, popular wisdom holds that around 40,000 entries is a minimum for a working MT application that is not restricted to a very well defined sublanguage. Automated methods of lexicon acquisition are of interest to any large-scale MT project.

Our work in lexicon building is part of LMT, Logic-based Machine Translation, an international effort with multiple language pairs that shares a common framework and philosophy. Source analysis is based on Slot Grammar; translation is transfer-based.

We have extracted automatically lexical data from three sources: machine-readable dictionaries (monolingual and bilingual), existing terminology data bases, and text corpora; developed strategies for merging data from the various sources; and designed a data base repository for maintenance and optimal reuse of automatically acquired and manually created or revised bilingual lexical information.

The linguistic basis of Slot Grammar is the concept of heads and modifiers, i.e. heads have slots, and modifiers are the fillers of these slots. A single LMT entry with one or more slot frames contains (1) source syntactic and semantic information (among other things, part of speech, slots, inherent features) required for source language parsing; (2) correspondence in the target language of both source word and source slot fillers, along with a map for source-target structural divergence, if any; and (3) morphological information about the target word, (4) syntactic, semantic, or domain criteria for target term selection. All of the above four data types have been instantiated automatically from lexical data bases, which were automatically created from machine-readable dictionaries using a dictionary parsing program and dictionary grammar in a process aimed at recovering all the lexical information, both explicit and implicit, in the machine-readable resources. Some of the data types have been instantiated from hand-built terminology data bases using a simple transducer. Finally, some were instantiated from aligned text corpora, using statistical methods to extract bilingual terminolgy.

The different processes of automatic lexical acquisition we have used are not error-free; nor do they yield entries that are complete and detailed enough for using asis in LMT. LMT's lexicon precedence mechanism allows us to use these entries for words missing from the main, human made/revised lexicon; however for better translations, it is often necessary to rework the extracted entries manually. The TransLexis data base and user interface enforce consistency and integrity of data while supporting multi-user access to a multilingual database for multiple MT language pairs in which morphological, syntactic, and semantic information for words and expressions of each language is described only once and all bilingual mappings are explicit. Lexical information extracted from online materials by automatic means is imported into TransLexis to be used asis or fleshed out manually on a word frequency or as-needed basis. Runtime access of LMT to the database is slow, due to the complex structure of the database; there is an export program for producing a complete lexicon in LMT format.

We have around 26,000 entries extracted from the Collins English-French dictionary; 34,000 from Collins English-German; an English-German terminology data base yielded about 25,000 terms; an English-French about 15,000. The statistical, corpus-based methods are still in development. In one experiment, we obtained good word pairs in 65% of the cases (10% are wrong, and 25% yield no result because of strict filtering criteria). This result is quite good considering the rudimentary methods that were used, and we are very confident that refinements where uninflected forms and linguistic information are used will give us much better results.

[1] Neff: IBM T. J. Watson Research, P. O. Box 704, Yorktown Heights, NY, 10598, NEFF@watson.ibm.com; Bläser and Lehmann: IBM Germany, Scientific Center, Institute for Knowledge Based Systems, Postfach 105068, D-6900 Heidelberg, Germany, ALSCHWEE@dhdibm1.bitnet and LEH@dhdibm1.bitnet; Langé: IBM France, Scientific Center, Paris; Zapata: Centro de Investigación UAM-IBM, Santa Hortensia 26-28, 28002 Madrid, Spain, ISABEL@emdcci11.bitnet.