Using Machine Readable Dictionaries for the Creation of Lexicons

David Farwell, Louise Guthrie, and Yorick Wilks
Computing Research Laboratory, New Mexico State University

Our position is not very contentious. Both monolingual and bilingual machine readable dictionaries (MRDs) have information related to words and their contexts. Lexicons for natural language processing (NLP) need information related to words and their contexts. Why not, then, use that information in an MRD which is relevant to the lexical specifications of an NLP system and which can be extracted automatically either to construct or partially construct the system's lexical items?

Arguments against such an effort include:

(1) the information in MRDs is wrong,

(2) the information in MRDs is not relevant,

(3) the relevant information in MRDs cannot be extracted automatically,

(4) relevant information can be extracted automatically but at too great a cost.

For each of these arguments, there is a reasonable counterargument.

With respect to (1), perhaps there is information in MRDs that is wrong. However, in most cases, a good deal of time and money has gone into making it right. More useful questions concern how often the information is incorrect (or correct) and whether the incorrect information is systematic. Even consistently assigned incorrect information can be exploited. Fortunately, with little effort, (1) is testable.

For a particular NLP system, the information that can be extracted automatically from a given MRD may well be irrelevant. However, what is or is not relevant is defined by the requirements of the NLP system's lexical specifications. In general, MRDs contain at least some basic lexical information (e.g., citation form, spelling forms, syntactic category) which most NLP systems require and often contain much more (e.g., hypernyms, synonyms, derived forms, collocations). The only way to address (2) is to look at the entries of the MRD and at the lexical requirements of the particular NLP system and see what can be used.

With respect to developing simple extraction procedures for many types of relevant lexical information, especially semantic and pragmatic information, (3) is certainly true. However, even today there exist procedures for extracting various types of relevant information from MRDs for use in the construction of lexical entries. It is only after identifying the requirements of an NLP system's lexicon, and how the corresponding information is expressed in the MRD, that this question can begin to be addressed.

Finally, MRDs have multiple systems of symbols, including natural language, that are used to analyze, categorize, define, or exemplify words. NLP lexicons commonly exploit systems of symbols (even natural language) for the same purposes. The correspondence between the systems of the MRD and of the NLP lexicon is sometimes very simple (no "problematical" choices during mapping) and direct (no inferencing rules), sometimes less simple (some "problematical" choices) and less direct (some inferencing), and sometimes very complex and rather indirect. Depending on various requirements of the application, (4) may well be practically true as well. But, again, it has actually been shown to be false under the appropriate circumstances.

In our own work, we have followed a five step approach to extracting information automatically from MRDs for use in constructing NLP system lexicons. These steps include:

(1) looking at the information requirements of the lexical entries of the NLP system,

(2) looking at what information is in the MRDs and how it is expressed,

(3) developing procedures for converting the relevant information provided by the MRDs into appropriate lexical formats for the NLP system,

(4) completing the specification of the lexical entries by hand,

(5) testing and debugging the lexicon within the application.

To expedite this process, especially steps 2 and 3, the CRL has developed an interface, Lexi-CAD/CAM, which allows the user to access and arbitrarily format different types of information in an MRD. We exemplify the approach using the MRD versions of the *Longman Dictionary of Contemporary English* and *Collins Spanish-English English-Spanish Dictionary* and the ULTRA machine translation system. We review the problems encountered (especially at step 3) and we provide some informal results (which at least partially support our claim that assertions 3 and 4 are false).