ELECTRONIC LEXICONS
FOR ENGLISH-JAPANESE MACHINE TRANSLATION

Alexis Kauffmann
LATL - Université de Genève

A RULE-BASED MULTILINGUAL MACHINE TRANSLATION SYSTEM

Its-2 is a rule-based MT system (see Wehrli, Nerima, 2008). It translates text sentences. It uses the multilingual linguistic syntactic parser FIPS for syntactic parsing of the source sentence.
Then it applies syntactic and lexicalised translation rules.
Finally, it creates the target sentences using generation rules. The rules are coded in language-specific transfer and generation modules, programmed in Component Pascal.
Its-2 uses monolingual and bilingual electronic lexicons.
In order to obtain a good quality of translation, these lexicons must meet three criterions:

- -good coverage of general vocabulary; this can be obtained creating large lexicons with at least about 50000 lexemes (words in their canonical form) and enough syntactic and semantic information.
- -reliability;
- -adequacy for MT.

JAPANESE MONOLINGUAL LEXICON

A first version was created in 2007, using data from the French-Japanese bilingual electronic dictionary "Fr-Edict" and entries entered by hand. It contained about 15000 lexemes.
A larger, new version has been developped since 2008. Data was automatically inserted using CJK Dictionary Institute's Japanese lexical database. This process was ordered in four steps:

- -automatic classification (using Perl scripts). The lexicon has been ordered following a classification closed to the ones of other monolingual lexicons used by Fips and Its-2;
- -morphological generation (using Component Pascal procedures);
- -insertion (with SQL/Access); the canonical forms were inserted into the "lexeme" table and all the conjugated ones into the "word" table;
- -corrections and improvements.

The monolingual database contains now 207615 lexemes and 5725727 conjuguated forms.

JAPANESE-ENGLISH BILINGUAL LEXICON

It has been developed since 2008, using data from CJK Dictionary Institute's English-Japanese and Japanese-English databases.
It stores one-to-one bilingual correspondences between Japanese and English lexemes or collocations.
First, they were automatically inserted and then they got ordered.
The bilingual database contains now 117354 bilingual correspondences.

ONGOING IMPROVEMENTS

Lexicons:

- -creation of a Japanese collocation database (collocations are pairs or group of words that are usually associated, like "make a deal","好きです"... see Seretan, 2007);
- -specification of syntactic and semantic properties;
- -description of Verb argumental structures, using information from Japanese a case frame database which was developed at Tokyo University (see Kawahara, 2006);
- -insertion of proper nouns, taken from the EnamDict electronic dictionary.

MT System:

- -translation of simple sentences;
- -translation of sentences with asymetric bilingual correspondences.

"love" => "好きです"("suki desu")
English verb Japanese (adjective + predicate) collocation

FUTURE WORK

Lexicons:

- -improvement of the collocation database;
- -creation of a new French-Japanese lexicon, by transitivity process.

MT System:

- -translation of complex sentences;
- -improvement of the French-Japanese MT module;
- -creation of a possible choice for the user to select the right Japanese politeness level.