Alan K. Melby

Brigham Young University
Dept. of Linguistics
Provo, Utah 84602
USA

ABSTRACT

One of the necessary tasks of a machine translation system is lexical transfer. In some cases there is a one-to-one mapping from source language word to target language word. What theoretical model is followed when there is a one-to-many mapping? Unfortunately, none of the linguistic models that have been used in machine translation include a lexical transfer component. In the absence of a theoretical model, this paper will suggest a new way to test lexical transfer systems. This test is being applied to an MT system under development. One possible conclusion may be that further effort should be expended developing models of lexical transfer.

1. An Early Approach to Lexical Transfer

Years before the machine translation community was burdened with guilt by the ALPAC report, David Hays, former chairman of the International Committee on Computational Linguistics, proposed a procedure for lexical transfer (Hays, 1963, pp 205-208). We will describe it, quoting pieces to preserve the original flavor.

1. a. File of Equivalent-Choice Data

"Most words...have uniform translations, but not all." "These exceptions to the general rule must be discovered and taken into account. The procedure is simple and straightforward. A file of equivalent-choice data...is required." This file is prepared using real text. When a word is encountered for the first time, one translation is selected and entered into a bilingual glossary. When the same word is encountered again, the human translator/editor attempts to use the translation already in the glossary. Additional translations are added only when the one(s) in the glossary are not acceptable. This procedure is supposed to avoid entering interchangeable alternatives that are only stylistic variations. The file of equivalent-choice data mentioned above is a record of how many times each translation was used.

1. b. Category

Once the equivalent-choice file has been compiled, the first step in analyzing it is to identify those words with two or more translations (i.e. equivalents). The next step is to identify whether the translation is governed by "grammatic category".

1. c. Function

If there are two or more translations within the same category, then the analyst looks at "grammatic function" (e.g. subject, object of preposition, etc).

1. d. Features

"If there is any kind of relation in which the word has two equivalents, the analyst continues by examining each word that governs the multiple-equivalent word." From the words that govern the word in question, the analyst derives "criterial classes" (i.e. features to mark on the governing words in the dictionary).

1. e. End of Procedure

But what if testing features on other words is not sufficient to determine the translation? Hays recognizes this possibility by noting "There is no certainty, of course, that the governors and dependents of an occurrence determine its translation, but is seems plausible that they will often do so." Even following this procedure of assigning features may be anything but "simple and straightforward", and (even worse) it is not sufficient.

2. Some Counterexamples

Obvious counter-examples to Hays' hope come to mind, such as "chip" in micro-electronics (=integrated circuit) and in gambling (=token to represent money), or "buck" in hunting (=male deer) and in slang (=dollar). Even older systems like SYSTRAN provide for this by prioritizing various domain-specific dictionaries.

The other situation where the Hays approach clearly breaks down is when the word is part of a fixed expression, such as "chip off the old block" (=like his father) or "pass the buck" (=avoid responsibility). All machine translation systems provide for this by including expression dictionaries that override word-for-word lexical transfer.

A variation on the expression dictionary is to key a lexical-structural transfer from a single word that affects surrounding words. For example, the adjective "hungry" stimulates a transfer going into French that changes "x is hungry" to "x has hunger".

Unfortunately, sometimes all the tricks listed above combined still do not suffice to identify an acceptable translation. The following two examples and some of the subsequent discussion are adapted from Melby (1985).

2. a. The "plate" Example

For example, consider the word "plate" (which can mean a household item on which food is placed to be eaten or a marker on the ground in the game of baseball) in the sentence

There was an egg on the plate

in the context of a discussion of a baseball game in which an angry fan threw a raw egg

and it landed on home plate.

Then consider the sentence

She threw the food on the plate

in the discussion of what a teenager did because she was in a hurry to fix her breakfast and get to school.

Either of these sentences could occur in a narrative of the life of a young person (so a special dictionary will not help), and there is apparently nothing in the syntactic or lexical environment of "plate" that determines the translation, and no idiomatic expressions are involved.

2. b. The "rock" example

Or consider the problem of "rock", which is a single lexical item in English with several specific translations in French ("pierre, roc, caillou", etc.), no one of which is as general as the English word "rock", which would be translated differently in each of the following sentences:

She found the rock on the beach and placed it in her pocket.

He climbed up and sat on the rock to get a better view.

While watching the parade, she got tired and sat down, not seeing the sharp rock, and screamed from the pain.

2. c. Opinions

Some claim that the above examples are far fetched and that the identification of lexical category, grammatical function, general subject matter, and fixed expressions is sufficient to develop lexical transfer algorithms that produce acceptable translations.

Others say that the dynamic nature of natural languages will often present a significant number of cases where lexical transfer is not handled adequately by standard techniques. Consider, for example, the search for lexical transfer criteria for the English preposition "through" going into French found in Bourquin-Launey (1984).

The facts are (1) that raw machine translation output, even after these many years of development, is seldom up to publication standards without post-editing; and (2) there has been little development of the models of lexical transfer beyond the stage described in section one (which was in place already in the mid 1960's). One reason is that the linguistic models that have been used do not include a lexical transfer component. This can be verified by looking at Hutchin's updated survey of linguistic models in machine translation (Hutchins, 1984).

To summarize the discussion to this point, the fact is that work in lexical models has been neglected for the past twenty years; the question is whether that neglect is justified.

It seems that it would be worthwhile to at least examine the nature of the problems in machine translation output. If it turns out that a significant portion of the problems are actually failures in lexical transfer, then further studies of lexical transfer in the computational linguistics community should be encouraged.

The following section describes an on-going effort to test the lexical transfer component of a machine translation system currently under development. The method allows a test of lexical transfer even before the entire system is operational, thus providing feedback as early in the life of the project as possible, hopefully allowing design changes to be made, if they are needed, before they become too costly. The method could also be adapted to studies in lexical transfer somewhat independent of a particular machine translation system.

3. A Method of Testing Lexical Transfer

3. a. Origin of the Testing Project

The BSO Company is a systems house in Utrecht, The Netherlands. The BYU-HRC is a center which promotes research in the College of Humanities at Brigham Young University (BYU), in particular providing support for research involving language and computers. In 1984, the author replied to a request for comments on the then proposed BSO machine translation project. That reply led to discussions which resulted in an agreement between BSO and the BYU-HRC to create a text and translation data base as a joint venture.

The specifications for the data base were that it would consist of paired texts in French and English, that it would include at least 500,000 words in each language, that the translations would be done by qualified professional translators, and that the text type would be straightforward modern English and French avoiding texts that are literary or intentionally ambiguous.

BSO supplied the source documents, which were mostly public reports on agriculture, social conditions, etc., published by the European Economic Community (EEC) or the United Nations (UN).

The documents were placed in machine readable form using the Kurzweil OCR device and transferred to a disk pack on an IBM 370/138.
Preliminary to the test, the data base was split into two parts. A smaller part (about 200,000 words in each language) is accessible to BSO for syntactic studies. The larger part (about 300,000 words in each language) will be used for the lexical test, under

control of BYU, and will not be accessible to BSO. The test will be executed in five steps.

3. b. Secret selection of sample texts

From the lexical part of the data base (300,000 words of English and 300,000 words of French), BYU has selected 4 sample texts. Each of these samples is a coherent stretch of text of approximately 300 word tokens. Unless otherwise indicated, we now refer to the English version of each sample text.

From the 4 texts there were about 600 non-function words. Adding some "misleaders" from other sections of the data base, a list of 800 words was sent to BSO.

3. c. Construction of trial lexicons

BSO has received the alphabetically sorted list of 800 English words. They do not know the context of any of the words, nor will they know whether a word is part of one of the secret sample texts or merely a "misleader".

BSO is now building an English to intermediate language (IL) lexicon and an IL to French lexicon. In the case of the BSO project, the intermediate representation is a formalized Esperanto. The English-IL lexicon will have more "exits" than input words (estimate -- 1200). So the IL-French lexicon will have about 1200 entries.

The English-IL and IL-French lexicons will each be separately tested at BSO. For this purpose, BSO will use English and IL trial input texts, especially written for this purpose by external consultants based on the 800 and 1200 word lists. In addition to the 800 words, BSO will add a few hundred function words (articles, pronouns, prepositions, etc.).

3. d. Translation using the trial lexicons.

After completion of both trial lexicons by BSO, they will be sent to the USA for overall testing by BYU. Two translation phases will be distinguished as part of the testing procedure.
First, monolingual students of English will translate the four sample texts (see A) into IL by mechanically following the rules contained in the entries of the English-IL trial lexicon. We emphasize that these students will have no knowledge of IL, which will help them apply the rules as mechanically as possible.

Second, a different group of students -- again English monolinguals -- will translate the IL-output of phase I to French, by mechanically applying the rules of the IL-French lexicon (without knowing IL or French). Since the test is directed at lexical word-choice, no complete sentences will be required for the French output.

The presence of an IL-version as intermediary between the two translation phases will require that the IL-output of phase I be converted into complete sentences before serving as input to phase II.

3. e. Evaluating the French output

As a final part of the whole procedure, BYU will carefully compare the French output of the above described DLT trial translation with the high-quality human translation of the text samples in the database. Of course, this comparison will concern only lexical elements, ignoring case endings, word order, etc.

Where the output differs from the translation in the data base, a French-English bilingual will decide whether the output is an acceptable alternative and if not will note the discrepancy for further study by BSO.

4. Conclusion

Once the study is completed (which will be years since the above procedure is only the first major phase), we should know more about the types of problems to be encountered in lexical transfer at later stages when the entire machine translation system is in place. The results should identify general problems in lexical transfer that are not specific to the BSO project. Based on the findings, we may recommend that further work in lexical transfer be pursued by the computational linguistics community.

REFERENCES

Bourquin-Launey, M-C. (1984) Traduction Automatique -- Aspects Europ/eens. Paris: ADEC, 99, boulevard Saint-Michel (5th arrondissement); January 1984.

Hays, David G. (1963) "Research Procedures in Machine Translation" in Natural Language and the Computer, edited by Paul L. Garvin, pp. 183-214. New York: McGraw-Hill; 1963.

Hutchins, J. (1984) "Methods of Linguistic Analysis in Machine Translation" in the proceedings of the International Conference on Machine Translation, Cranfield, England, February 1984.

Melby, Alan K. (1985) "A Text and Translation Data Base", a paper presented at the International Conference on Data Bases in the Humanities and Social Sciences, Grinnell College, 1985. (Submitted for publication in the proceedings.)

--end of text--