9 The MicroCat Machine Assisted Translation System

An Assessment Project at the University of Exeter

By Derek Lewis
Department of German, School of Modern Languages
University of Exeter

Introduction

In 1977 the Weidner Corporation produced an interactive Machine Assisted Translation system for providing fair quality translations which could be post-edited using an integrated word processor. The system came with its own core dictionary of general vocabulary, although the user could build subject-specific dictionaries as required. In the 1980s a scaled down version appeared, the MicroCAT. The MicroCAT became the first commercial MAT system for IBM PC and was promoted as a low-cost operational system for medium-sized organisations. No longer marketed or upgraded for post-8086 processors, the MicroCAT is technically obsolete, although its linguistic capabilities have not necessarily been superseded by comparable systems for stand-alone PCs. The following account focuses on the core of the MicroCAT, viz. its translation module and dictionaries, which were evaluated as part of a course on the Theory and Practice of Translation for final year modern languages students at Exeter University between 1989 and 1992. The language direction is English to German.

Types of error

The most obvious error in translation output occurs whenever a term is not in the system's dictionary, e.g. the English *sixth formers* is rendered as German *sechstes *Formers** (the asterisks delimit a form not found). The solution is to enter the multiple word form *sixth former* as *Abiturient* in the dictionary, especially if it is likely to occur frequently within the subject-domain.

Failure to translate is often preferable to mistranslations of multi-word terms, especially where the attributive member of a multi-word term is syntactically or semantically ambiguous, as in *ground, conviction* or *cover* below:

ENGLISH INPUT:	MICROCAT OUTPUT:	CORRECTED VERSION:
cover story	*Decke Geschichte*	*Titelgeschichte*
conviction date	*Uberzeugung Datum*	*Verurteilungsdatum*
ground coffee	*Erdkaffee*	*gemahlener Kaffee*
ground forces	*Erdkrafte*	*Bodenstreitkrafte*

Again, the solution is to enter the desired translations in the dictionary.

The MicroCAT recognises only one translation for a word of a particular syntactic category. Thus the English noun *bank* must be translated either as *Ufer* (river bank) or *Bank* (financial institution). There is no way of programming the system to choose between the two in the same text. MicroCAT's designers claim that this ensures consistency of terminology for subject-specific texts. Thus *worm* in engineering will always be *Schnecke* (literally: *snail*), never *Wurm*. Even in limited semantic domains, however, ambiguities can arise. Is a *conviction* in a text on English law likely to be *Verurteilung* or *Uberzeugung?* Both are surely possible.

The ambiguity problem is more serious with verbs, given their pivotal role in sentences. Examples of English verbs which are two- or even three-way ambiguous in German include *present* (*uberreichen, darstellen*), *file* (*feilen, einordnen*), *ground* (*niederlegen, Fahrerlaubnis entziehen, grunden*), *pitch* (*werfen/befestigen/stimmen*), not to mention the phrasal verbs such as *put over, take off, get through, look up*, etc. Experience shows that post-editing rapidly breaks down where such items are mistranslated. Although most ambiguities can be resolved by reference to linguistic context (e.g. the type of object of *file* or the position of the object of *look up* to distinguish *look a word up* and *look up the staircase*), there is no method of passing such information to the MicroCAT's processing rules.

MicroCAT output typically slips up on word order and subject-verb agreement, even for short subordinate clauses. Thus *while teachers may also find it helpful* is rendered as *wahrend Lehrer auch es hilfreich finden kann*, where *auch es* and *Lehrer ... kann* are deviant. As a rule such obvious errors are easily corrected in post-editing.

A more serious error is the assignment of the wrong syntactic category. An example is *Stromung Deutsch Praxis* for *current German practice*, where the adjective *current* is wrongly parsed as a noun. It takes an imaginative bilingual post-editor to pick this one up!

It is possible to identify areas of syntactic complexity where translation quality consistently breaks down to the point of incomprehensibility. Briefly, these are gerunds and participial constructions, embedded clauses, co-ordinate main clauses, main clauses introduced by an adverb, and noun-verb ambiguities. Some examples (1 to 5) are given below:

(1) EMBEDDED CLAUSES:

English: *If we hurry we will find him*
German: *Wenn wir uns wir beeilen, ihn finden werden*
Comment: The second *wir* should be located in the phrase after the comma.

English: *If you do not work, I will not pay*
German: *Wenn Sie ich nicht arbeiten, nicht werde zahlen*
Comment: *ich* should be located in the phrase after the comma.

English:. *If you press the button, a light comes on*
German: *Wenn Sie den Knopf drucken, den das Licht auf kommt*
Comment: The second *den* does not belong anywhere.

(2) CO-ORDINATE MAIN CLAUSES:

English: *Do not cook eggs in the shell or they will explode*
German: *Kochen Sie Eier in der Schale nicht, den es explodieren wird, nicht*

(3) MAIN CLAUSES INTRODUCED BY AN ADVERB:

English: *But do not be surprised*
German: *Aber nicht ist uberrascht*

Deleting *but* solves the problem:

English: *Do not be surprised*
German: *Seien Sie uberrascht nicht*

(4) NOUN/VERB AMBIGUITIES:

These arise with identical plural noun and 3rd person singular verb forms (e.g. *wishes, passes*):

English: *The queen was thinking of her wishes*
German: *Die Konigin wunscht an sie dachte.*

A singular wish, on the other hand, is translated perfectly:

English: *The queen was thinking of her wish*
German: *Die Konigin dachte an ihren Wunsch*

(5) Lists of nouns can give odd results: occasionally all the nouns are omitted, leaving only a string of commas. In other respects it is evident that the system relies heavily on punctuation during parsing to separate constituents - pre-editing a complex sentence by adding extra punctuation can often improve performance.

In the following cases an experienced post-editor may anticipate the quirks of the system:

(6) The English simple past tense is consistently translated by the German present perfect instead of the imperfect (*I lived --> ich habe gewohnt*). The German imperfect is reserved for the past continuous (*I was waiting*).

(7) The English personal pronoun *one* and, more importantly, the elliptical form, are confused with the indefinite article:

English: *One has to pay the bills regularly* .
German: *Ein mu die Rechnungen regelma ig zahlen*

English: *We can not build one*
German: *Wir konnen uns Ein nicht bauen*

(8) There are gaps in the system's knowledge of irregular English morphology. Thus the MicroCAT does not know *teeth* as the plural of *tooth* so it will never normally recognise the plural form of this word. The user has no access to the English morphological or syntactic processing rules. Usually the dictionary update program asks him to enter more information about a German translation then its English source word.

(9) The various uses of English *no*, as in *There are no chairs, No, I cannot come, No less than 50, No more wine*, are generally rendered by a blanket *kein*.

The MicroCAT boasts a sophisticated procedure for entering idioms, which are typically multiple word entries with so-called "holes" for variable items (such as general word categories). There is no space here to discuss this topic further, but an example may help. To translate *I like her* by *sie gefallt mir*, the English subject becomes the German indirect object and the English direct becomes the German subject.

The dictionary entry for the verb is:

%1 *like* %2
%3 *gefallen* %4

The %n values denote "holes" which take any noun or pronoun and are cross-referenced, so %4 translates %1 and %3 translates %2. *Gefallen* is specified as taking an indirect object in the dative case. Results are as follows:

I like her --> Sie gefallt ich
She likes me --> Mich gefalle sie

The man likes the woman --> Die Frauen gefallt Mann

Men like women --> Frauen gefallen Mannern

The system changes the order of constituents but fails to mark case relationships correctly. In practice only the simplest of idioms are worth entering.

Conclusion

Most users concluded that the most comprehensible results were obtained for simple mono-clausal declarative sentences where English and German exhibited parallel syntactic and semantic structures - in other words where translation came closest to word-for-word substitution. An experienced post-editor could anticipate many features of the MicroCAT and compensate for its shortcomings. Often, however, the misplacement of constituents, the inability to handle ambiguity and the accumulation of errors required recourse to the source language text.

Many users noted the knock-on effect on comprehension of relatively minor errors early on in a sentence or text. Since the position and semantic load of individual phrases or items will vary from text to text, useful measurements of the comprehensibility of the output of a particular MAT system may have to be based on statistical assessments over large volumes of text rather than detailed studies of individual fragments.