REFTEX - A CONTEXT-BASED TRANSLATION AID

Poul Søren Kjærsgaard
University of Odense
Campusvej 55
DK-5230 Odense M

ABSTRACT

The system presented in this paper produces bilingual passages of text from an original (source) text and one (or more) of its translated versions.

The source text passage includes words or word compounds which a translator wants to retrieve for the current translating of another text. The target text passage is the equivalent version of the source text passage. On the basis of a comparison of the contexts of these words in the concorded passage and his own text, the translator has to decide on the utility of the translation proposed in the target text passage.

The program might become a component of translator's work bench.

Introduction

Computers can contribute to translation either automatically or as an aid to the human translator (machine-aided translation). The latter represents a large spectrum of different approaches as to the degree of human intervention in the translation process and to the method(s). Some systems are semi-automatic in the sense that they only ask for human intervention for the resolution of ambiguities (Melby, 1981). Other systems are designed to relieve the human translator of some tedious aspects (such as dictionary look-up) of the translation work, either interactively via a terminal or by batch processing overnight.As to method(s), most systems are based on dictionary look-ups - sometimes combined with automatic insertion of the retrieved equivalents (McNaught, Somers, 1979).

This paper will describe an alternative method, REFTEX. A major difference between REFTEX and most other machine-aided translation systems that I know of is that REFTEX emphasises the context, whereas other systems rely on bilingual dictionaries containing translations (sometimes uncommented) and possibly definitions or explanatory remarks.

The system was first implemented on a CDC mainframe installation, but has now been converted to an IBM XT-microcomputer. The primary scope of the program is to provide a supplemental aid for human translators.

The principles of REFTEX

The name of the system, REFTEX, is an acronym for reference text. Its main characteristics can be summarised as follows: The system is meant to be used when the translator comes across some word or word compound that cannot be looked up in a dictionary or the translations of which do not seem relevant in the context of the actual translation. The translator can then have recourse to texts that have already been translated, in order to try to retrieve the wanted word(s) and its/ their translation(s). Such texts exist in an original (source language) version and one or more translated (target language) versions. In REFTEX, such texts are designated reference texts. During execution of the program, the program will access passages (concordances) of the original text that contain the word and the equivalent passages of (one of) the translated versions. The translator will then decide if the translation contained in the target language version is useful in the actual translation.

It is an interactive, screen-oriented system that can be used by a translator during the translation process. In the present version, the text to be translated and its translation are supposed to exist independently on paper, but nothing prevents the implementation of an integrated version using windows (cf. last section).

REFTEX can thus be conceived of as a computerised combination of bilingual concordances used in philology (usually on ancient texts) and the manual use of translated text as an aid for the translator. But in contrast to traditional concordance making, the project does not aim at producing a finished product of the works of an author, but at supplying the translator with an ad hoc tool.

The REFTEX system

REFTEX has been implemented as a program package of two independent programs: ARBORAL and REFTEX.

The former uses one or more slightly pre-edited reference texts as input and transforms each into an equivalent data structure that contains both the original information (thus permitting a reconstruction of the original text) and some new information which facilitates the searching of words in the text and the concordance making.

The data structure is organised as two records. The first one contains a node or an index for each different word of the text together with some satellite information: absolute word frequencies and pointers to the first occurrence of the word. The second record is a list structure containing a reference for each individual word of the reference text to its position in the first record, and pointers to possibly following occurrences of the word and to the beginning of the paragraph (concordance) that contains the word.

Once the finished data structure has been established, the program writes it on a file, from where it can be accessed by the main program REFTEX.

The pre-editing of the reference text that was mentioned above consists of the insertion in the source text of period markers (the number sign: #) together with a number that uneqivocally identifies each passage. A passage normally consists of one period, possibly two. Then, parallel period markers and numbers are inserted into the target text(s) to ensure the retrieval of parallel extracts (concordances) of the source and target texts. If this pre-editing were not carried out, it would not be possible to extract parallel passages, if the source and target languages involved are structurally different in respect to modes of expression. And even for closely related languages such as the Scandinavian languages, this would probably be the case.

REFTEX is the part of the program package that will be used by the translator during the process of translation.

Program execution starts by asking the translator to key in names of the pair of reference texts he/she wants to use for solving the problems of the actual translation. The program then asks for the first key word to be searched in the reference text, whose equivalents the translator wants to know. If the reference source text contains that word, the program will print out the passage containing the first occurrence of the word together with the equivalent passage of the target language version. On the basis of his world knowledge (pragmatics) and knowledge of the two languages involved, the translator now has to decide whether the source language passage is sufficiently similar to the context of the actual translation to permit reusing the translation contained in the target language passage. The decision of course depends on the quality of the translated reference text and relies on the translator's ability to detect possible errors.

If the first bilingual concordance does not contain an acceptable translation, the translator can "scroll" to the following occurrence(s), until he finds an adequate translation or the reference text is exhausted. If either the word does not exist in the reference text or it does not have appropriate translations, it will be saved in a special array for non-retrieved words and can be searched in another reference text, after the translator has finished the list of words or expressions that he wants to look up. In case that words have been saved in this array, the program will ask for another pair of reference texts. Supposing that they are available, the program will try to retrieve passages containing the words that were saved.

An additional feature of REFTEX is a semi-automatic routine that enables the program to retrieve inflected forms of a word, for instance feminine and/or plural forms as in the Spanish word español - española, españoles, españolas. The routine solely relies on formal characteristics of words (such as word endings) and not on semantic or other markers that would imply some sort of "understanding" of the word (as is the case in many grammars). For the time being, the routine has been implemented for regular nouns, adjectives, verbs and participles in French and Spanish.

Computational concordance making

Given that the REFTEX-approach relies on a bilingual concordance, this section will briefly introduce two of the problems this causes: word-form diffusion and homoform-insensitivity. The former problem reflects the wish to group together different inflected forms of the same word. The solution proposed in REFTEX is to depart from the primary form and consequently generate inflected forms automatically, when regular and manually, when irregular.

The latter problem reflects the homograph or polysemy problem. To solve this problem completely, one would need either a sort of tagging (requiring extensive pre-editing) or some semantic analyzer. Neither of these solutions has been chosen in the REFTEX-approach. A "pragmatic" solution, based on the immediate context, has been developed, thus reducing the amount of superfluous information or "noise".

An example will illustrate its function: The French word "application" has multiple meanings, and may in some texts be quite frequent. If the key word to be looked up is the compound preposition "en application de", the word takes on yet another meaning. In order to narrow the search field, REFTEX permits the translator to look for the word "application" together with "en" and "de".

In this way, a lot of, though not all, irrelevant information will be excluded.

Methodological considerations

The use of bilingual concordances implies that REFTEX can be characterised as a context-oriented translation aid in opposition to the dictionary-oriented approach that most machine-aided systems rely on.

These two approaches both possess weaknesses. The problem of a context-oriented approach can be restated as the question of how reliable the translation of the reference source text is, whereas the problem of a dictionary-oriented approach may be the difficulties of defining precisely the words of a language (cf. Wittgenstein). In fact, the difference between the two approaches comes down to the question of whether words possess an independent meaning, defined at the "langue"-level or their meaning is influenced by the actual contextual use of the words, the "parole"-level.

The difference between the two approaches may be illustrated by a well-known example from the MT-literature: the English verb "to know", which is rendered in many European languages by two different verbs. Does this verb have two distinct meanings which the lexicographer can account for or would it be preferable to let the translator decide the relevant equivalent on the basis of a series of bilingually concorded examples? A similar example would be the German word "Schlagsahne" which is rendered into Danish by two different words: piskefløde (cream) and flødeskum (whipped cream).

The strength of a bilingual dictionary approach is of course its ability in many cases to convey to the user a fairly good idea of the meaning of a word in another language.

The strength of an context-oriented approach is its ability to help deciding (just) which among a number of different proposals should be retained for the current translation. And, needless to say, in some situations, it will certainly be possible to combine the two approaches in order to make the best out of each.

The belief that the linguistic context contributes to determining the meaning of words is of course implied in the use of a context-oriented approach. Supposing that this holds true, another aspect of the approach is to determine whether the impact of the context is equally strong for any sub-vocabulary. In the negative, this would mean that a context-related approach would be less relevant in some cases.

No conclusive answer has been given to that question, but it seems fairly reasonable to suppose that the more specialised the vocabulary is the less the meaning of the word is influenced by the context. In such cases, the utility of the REFTEX approach may be the possibility to retrieve newly coined compounds that have not yet been lexicalised, or "loose" collocations that never appear in dictionaries.

Alternative applications

The primary scope of the program - as was stated in the introduction - is to provide a supplemental aid for human translators. In that respect, it could probably become an integrated part of a translator's work bench or amanuensis (Kay, 1980), enabling the translator to carry out all parts (translation, dictionary and reference text look-ups, text processing) of the translation process. This part of the project has not been completed.

A context-oriented approach may also be an appropriate tool for lexicographers and other researchers because it can provide the "raw material" for syntactic investigations as well. The system might thus prove useful for making "translation rules", i.e. rules stating how to translate syntactic phenomena from one language into another.

Relevant literature

Arthern, Peter: Machine Translation and computerized Terminology Systems; a Translator's viewpoint pp. 77-109 in Snell(ed.): Translating and the Computer. North Holland. Den Haag 1979.

Carestia-Greenfield, Carestia et Serain, Daniel: La traduction assistée par ordinateur: Des banques de terminologie aux systèmes interactifs de traduction. Paris 1976.

Kay, Martin: The Proper Place of Men and Machines in Language Translation. Xerox. Palo Alto/Cal. 1980.

McNaught, John and Somers, H.L.: The Translator as a Computer User. UMIST. Manchester 1979.

Melby, Alan K.: Translators and Machines - Can They Cooperate? in L'informatique au service de la traduction. Numéro spécial de META 26.1. Montreal 1981.