[*MT Summit III*, Washington, DC, USA, July 1991]

The Translator's Workbench: An Environment for Multi-Lingual Text Processing and Translation

M. Kugler, G. Heyer, R. Kese, B. von Kleist-Retzow, G. Winkelmann*
TA Triumph-Adler AG
TA Forschung EF
Fuerther Strasse 212
8500 Nuernberg
e-mail: marianne@triumph-adler.de

Abstract

The Translator's Workbench provides the user with a set of computer-based tools for speeding up the translation process and facilitate multilingual text processing and technical writing. The tools include dictionaries, spelling, grammar, punctuation and style checkers, text processing utilities, remote access to a fully automatic machine translation system and to terminological data bases, an on-line termbank, and a translation memory in an integrated framework covering several European languages.

1 Introduction

Due to the increase of international contacts, multilingual text processing and translation are getting more and more important. While the costs in computer time and memory are very high for fully automatic translation systems and their performance is still a problem, we have decided to develop an integrated package for assistance, not replacement of translators. The TRANSLATOR'S WORKBENCH, (ESPRIT project 2315) was designed to give assistance to professional translators and secretaries, scientists and engineers working in technical fields in performing multi-lingual text processing and handling large volumes of documentation in one or more languages. At several public demonstrations the public interest in an integrated system providing these facilities was very high.

The Translator's Workbench provides the user with an integrated set of computer-based tools for speeding up the translation process and facilitate multi-lingual text processing and technical writing with respect to sophisticated language checking and help for developing, retrieving, and updating terminology. The tools include dictionaries, spelling, grammar, punctuation and style checkers, text processing utilities, remote access to a fully automatic machine translation system and to terminological data bases, an on-line termbank and termbank building tools, and a translation memory in an integrated framework covering the languages English, German, Spanish, and to some extent Greek. Work is in progress to add the languages French and Italian in order to cover the main European languages.

2 State of Implementation

After the first two years of the three-year-project, a partial prototype of the overall workbench as well as prototypical realizations of stand-alone modules have been implemented based on FrameMaker and X Windows under UNIX on a SUN 3/80 workstation. Work on a MS Windows prototype with reduced functionality is in progress. The running prototype integrates a multi-lingual editor, sophisticated language checking facilities, a phrasal translation memory, remote access to the METAL machine translation system and the remote term-bank EURODICAUTOM, and an on-line term-bank with ca. 4000 terms on automotive engineering. The user profile has been tailored to the user's needs, as they were established by a detailed user requirements study. The interchange of documents to and from METAL is format-preserving and adheres to the international ISO standard for ODA/OD1F (ISO 8613).[1]

3 User Requirements

An extensive study of the requirements of professional (free-lance and in-house) translators was made during the first phase of the project. It included close observation of six translators at work and in-depth interviews with ten translators.

The result of this user requirements study "showed that there is a need for computer-based tools in professional translation circles. These tools include 'smart' editors, terminology data banks, and remote access to machine translation systems. The 'smart' editor will provide conventional word processing facilities, support the analysis of documents in both the source and target languages, and incorporate syntactic and stylistic analysis tools."[2]

*The work described in this paper is funded by the European Community, as ESPRIT project 2315.

[1] J. Delgado, F. Jordan and M, Medina: Access to automatic translation machines using X.400 messaging and ODA documents. 10th International Conference on Computer Communication, 5-9 November 1990, New Delhi.

[2] H. Fulford, M. Hoege & K. Ahmad: User Requirements Study. Mercedes-Benz AG, University of Surrey 1990.

4 The Tools

4.1 Language Checking

A conventional word-based spell checker is available for all the languages of the project. A new development is the context-sensitive spell checker for German exceptional cases, many of which can neither be handled by a word-based spelling checker nor by a grammar checker. This extended spell check treats special capitalization and concatenation problems (e.g. *in bezug auf* vs. *mit Bezug auf, radfahren* vs. *Auto fahren).* German and Spanish grammar checking and style checking is done by accessing the parser of the automatic machine translation system METAL. For the PC version, an ATN-based phrase parser with heuristics on sentence-level discovers errors in agreement (case, person, number, gender within phrases and number between subject and verb phrase).[3]

4.2 Term-Bank and Term-Bank Building Tools

The corpus-based Machine Assisted Terminology Elicitation (MATE) toolkit developed by the University of Surrey[4] enables translators and terminologists to access term-banks, text corpora and facilities for customized dictionary production, and thus enables the translator to store, retrieve and update terms. The term-bank developed within the project currently contains almost 4000 terms in the domain of automotive engineering in the languages German, English, and Spanish.[5]

The term bank available covers the whole range from definition, grammar, usage, collocation, and equivalents of a term in question up to encyclopedia, hierarchy and word family.

4.3 Translation Memory

Due to the vast knowledge necessary to do a full-fledged automatic translation, only few systems exist that are open to a larger market. None of them can be used on a small machine with little memory and disk resources. The more primitive dictionary-systems that do simple word-to-word translation are not applicable in most cases due to the low quality of their output when translating text. Therefore, while providing remote access facilities to the automatic translation system METAL developed by Siemens, we have followed the intermediary approach of providing partial translation as an additional tool which can be used on personal computers. The learning translation memory developed by the Fraunhofer Gesellschaft[6] enables the user to quickly retrieve previous translations in a flexible way.

Technical texts and business correspondence texts are highly formalized and repetitive. Furthermore, when writing manuals etc., it is important to ensure that a technical term is translated the same way anywhere in the text, in order to provide a uniform and standardized surface.

4.3.1 How it Works

The translation memory is based primarily on statistical methods, it works with Markov models of trigrams. Word S3 in the sequence S1 S2 S3 is translated to T3 in the sequence T1 T2 T3 (S: source language, T: target language). Frequency information is stored in the database, so if several translations are available, the more frequent one is selected.

Prior to the translation a training phase has to be performed. This is done in the same text processing windows as the translation.

In the training phase a variable number of words in language A can be linked to a variable number of words in language B. The system asks the user whether to update the dictionaries with the unknown words and then presents two windows with symbolic links between words or word groups of the two languages. Only translations that have not been trained previously are asked for, the others are suggested automatically and have to be confirmed by the user, thus speeding up the training process.

Whenever a sequence of words turns up in a new text, the corresponding translation is retrieved. The larger the sequence matching the originally trained text, the better the grammatical structure of the target language can be reconstructed. If the text is very dissimilar from previously trained material, the worst case would be word-to-word translation with occasional untranslated words, depending on the size of the dictionary. This is obviously not desirable, so the best application for this tool is text with repetitive expressions.

The method is inherently language independent, thus the database can easily be trained for any pair of languages.

4.3.2 The Database

The performance of the system with untrained text did improve only slightly by using a large non-specialized dictionary (25000 word forms, German/English). Translating text in the range of computer manuals showed that the vocabulary is highly application specific. Therefore work concentrates on providing a technical dictionary (technical terms, economics) on the one hand, and training of phrases from business correspondence on the other. The best performance was reached when training the system on the first paragraphs of a new manual and then applying it on the subsequent translation of the remaining text.

As the number of possible sentence constructions is enormous, training of full sentences requires too much time and provides too little output. The 'phrase unit', however, seems to be easier to standardize and thus looks the appropriate size for storing translations. We are at present experimenting with easier decomposition of sentences into trainable phrases.

[3] G. Thurmair: Parsing for Grammar and Style Checking COL1NG 1990.

[4] K, Ahmad, A. Davies et al.: A Methodology for Building Multilingual Termbases and Special-Purpose Lexica. TWB Technical Report, University of Surrey 1990.

[5] M.Albl, K. Kohn, S. Pooth & R. Zabel: Specification of Terminological Knowledge for Translation Purposes. TWB Technical Report, Universitaet Heidelberg, Institut fuer Dolmetschen und Uebersetzen 1990.

[6] B. Keck: Theoretical Study of a Statistical Approach to Translation. TWB Technical Report Fraunhofer Institute IAO Stuttgart 1989.

5 Example of a Working Session

Using the language checker, the source text is prepared for translation by detecting ill-formed input and complicated constructions. Having corrected the source text, unknown terms can be looked up in the term-bank, which among others provides encyclopedic knowledge, the translation(s) of the term and information on its usage. If a term is not available from the on-line term-bank, a link to a remote term-bank (e.g. EURODICAUTOM) can be made and the information can be retrieved from there.

The text is then translated, which can be done either interactively using the statistical translation memory or in batch mode by sending the whole text or parts of it to METAL, the automatic machine translation system.

The translated text can then be post-processed using the language checker of the target language. Source and target text can be viewed simultaneously in order to enable easy comparison and control of the original and the translation.

When using the translation memory, part of the revised translation could be trained immediately to be available for future use.

All the tools mentioned are integrated into a common system with a unified user interface. Most of the tools can be called from within the editor, thus enabling the user to perform the tasks without leaving the editing environment. Formatting information is not destroyed by applying the tools, when doing the batch translation by METAL, format information is preserved using the ODA/ODIF conversion format.

At present the domain of application is in manufacturing, especially within the automobile industry.

6 Products

SNI is interested in a workstation version of the Translator's Workbench integrating several tools using the FrameMaker environment (under UNIX), while TA/Olivetti will launch a product aimed at the MS-DOS PC market, containing an editor, conversion to and from the main text processing systems, a translation memory, and language checking integrated under MS Windows, covering the languages English, German, French, Italian and Spanish.