TRANSLEARN

Interactive Corpus-based Translation Drafting Tool

Stelios Piperidis
Institute for Language and Speech Processing
22 Margari Street, 115 25 Athens, Greece
email : Stelios.Piperidis@eurokom.ie

Abstract

This paper describes the research and development activities carried out during the first phases of the TRANSLEARN project. The aim is to build a translation memory tool and the appropriate work environment. The development of the prototype tool for the envisaged system proves the application's usefulness in the translation process of international multilingual organisations as well as in the localisation-internationalisation process of international enterprises.

Introduction

Translation work is very frequently characterised by two parameters: repetition and high demand on quality. This is particularly true for translation of technical and administrative documentation. The aim of this project is to tackle this problem by providing a computational environment, in more practical terms a toolbox that will :

- rid translators of the repetitive part of their work by reusing existing human translations and learning from them
- enhance quality and consistency of translation by being able to integrate ancillary translation tools.

TRANSLEARN's "end product" will be a prototype translation memory tool and the appropriate translation work environment for machine assisted translation in multilingual professional environments like translation departments of international organisations and enterprises wishing to have their products localised in the shortest possible time. In addition, multilingual corpora processing tools and resources developed during the project will constitute project side-products.

Achievements to date

TRANSLEARN's descriptive goal is to develop a machine translation aid tool dedicated, as previously said to managing repetition phenomena in the translation of specific types of text. This technology has been described in the literature as example-based machine translation or in practical terms, translation memory technology.

Its methodological goal is to employ sophisticated pattern matching techniques, involving linguistic and numerical processing, in order to identify the longest coherent part of source language text that is identical or similar to an input to-be-translated-text and retrieve from the memory the corresponding target language text.

The R&D activities of the project revolve around three major axes :

- multilingual parallel corpora, i.e. texts in different languages one being the translation of the other,
- alignment of parallel texts, i.e. establishment of correspondences between elements (paragraphs, sentences) of parallel texts,
- text matching mechanisms, i.e. mechanisms that can identify pieces of to-be-translated-text that have already been translated (or sharing a great similarity with pieces that have been translated) and stored in a database; subsequently, the proposed target language equivalent(s) are presented to the user in a ranking order based upon a scoring mechanism.

In the reported phases, work has concentrated on assessment of computational approaches to example-based machine translation, specification of user requirements, outline of system architecture, multilingual corpus acquisition, analysis and alignment as well as the development of increasingly sophisticated text matching mechanisms.

TRANSLEARN has collected and investigated a large body of parallel ASCII texts, between 5 and 6 million words, for each language, English, French, Portuguese and Greek. The corpus has been extracted from the CELEX database, the European Union's (EU) documentation system on EU law. The characteristics of the administrative sublanguage span the whole corpus while technical/financial sublanguage is used depending on the subject matter of each text. The corpus texts are of regulatory type with slight variations while the structure of almost all texts is the same. The corpus by itself validates the usefulness of the project by the high percentage of frequently recurring pieces of text that need not be retranslated since one can reuse existing human translations. In parallel, samples of texts extracted from car operation and maintenance manuals have been studied revalidating the usefulness of the approach.

A sample corpus representative of all sectors and all types of texts has been analysed in all four language versions. Normalisation and segmentation exercises have taken place so that the corpora are free from information that cannot be exploited by the project and are rendered in a form suitable for further processing. Tools have also been developed to extract statistical information from the corpus and identify a list of the most frequently used formulaic expressions.

A portion of the corpus has been lemmatised and tagged at part-of-speech (pos) level. Lemmatisation is performed by access to a morphological dictionary. The tagsets used are compatible with the TEI and NERC guidelines, catering at the same time for the peculiarities of each language. Lemmatisation and tagging return for each word of the text the combination <lemma, pos>. If multiple such combinations are valid for a word then all possible combinations are output. Combinations of more than one <lemma, pos> tuples are then grouped together to form a morphologically ambiguous class and these ambiguity classes are treated as tags of their own. Lemma and pos tag information is later utilised in the text matching process in order to determine identical or similar sentences and subsequently rank their similarity.

The corpora have also been aligned so that each paragraph and sentence in the French, Portuguese and Greek version is linked to the corresponding paragraph and sentence of the English version. The alignment software that was developed, based on techniques considering mainly statistical information, computed 97% correct alignments while methods for improvements and increasing robustness are currently being explored. Experiments for alignment below the level of sentence have also been made yielding promising results. The new methods combine the power of statistical modelling and surface linguistic information in order to establish correspondences between phrases/clauses across multilingual texts. The

alignment software will be used not only for translation data preparation but it will constitute an integral utility of the TRANSLEARN environment so that if the future user has translated texts available (s)he will be able to align them, store them and reuse them.

Translation data are stored in a database currently being developed. The selected platform is the ORACLE RDBMS in a client (Windows) - server(UNIX) architecture.

The core of the system is its text matching tool. Having processed the corpus in the manner mentioned above, and aligned it so that the system knows for each database sentence in a source language A the corresponding database sentence in a target language B, the matching tool can search for database sentences of language A that are identical or only similar to an input sentence (in source language A) and retrieves the equivalent sentence or sentences, if more than one exist in the target language. Two candidate matching algorithms have been proposed. Both algorithms require almost the same level of linguistic processing, i.e. lemmatisation and tagging at pos level. Modifications in order to simplify the algorithms and gain in efficiency and processing time have been considered. The approach that seems to be most promising is based on computations of common elements (words, lemmas, tags) between an input sentence and a database sentence and computation of consecutive elements in them. The possibility of considering whether consecutive elements make up a linguistically meaningful segment is a research issue. In addition, research is being made to devise algorithms for computing matches between a segment of a source language sentence and a segment of a target language sentence.

The matching tool first searches for exact matches between the input and the database sentences. In doing so, it does not take into account non-linguistic information of the sentences like dates and numbers so that real exact matches are not missed due to very minor differences. If no exact match is found, the matching tool searches for database sentences that are similar to the input, i.e. for fuzzy matches. In doing so, the tool considers surface linguistic data (words, lemmas and tags) in order to search for similar sentences and identify their common parts. The parts of the database and input sentences that are different are highlighted so that the user knows where to intervene in the proposed translation, thus saving his/her time. In addition, the system computes a similarity score between the compared sentences, based on the importance of the differences between them. The tool is externally configurable in that it can accept a minimum value for the similarity score in case of fuzzy matches. The modifications that the user may make to the proposed translations are then stored in the system for future use, thus enabling the system to learn new translations. A prototype implementation of the matching tool is already available. It has been tested on the CELEX data (approx. 3000 sentences), and has yielded promising results.

The approach adopted to text matching as well as the experimental results on text alignment will be presented at the COLING 94 and ACL 94 respectively.

In its first year of running, TRANSLEARN has cooperated and received feedback from completed and on-going research projects like the ET 10/63 project "Statistical Text-based complements for EUROTRA" and Translator's Workbench (ESPRIT) with which it shared interests in text alignment and multilingual corpora storage techniques. The project is also represented in the LRE-Information Retrieval Coordination Action. Currently, it is in close contact with the NERC2, EAGLES and RELATOR projects.

Interest has been expressed from private enterprises as well, especially with respect to the project's multilingual corpora processing activity while contacts with potential end-users have been established.

Forthcoming phases of work

The next steps in project development consist in improving the design of the textual database and ensuring the faultless communication between the different components (linguistic processors, matching tool). A rudimentary user interface will be developed to serve the prototype demonstration activities. The first version of the system prototype is expected to be ready by mid-June 1994.

In parallel, the linguistic processors will be enhanced by being able to tag groups of words according to the category to which they belong as a group. Such information is expected to lead to computing more meaningful fuzzy matches and yield better approximations to the similarity score. Improved versions of the linguistic processors are expected to be available in late September 1994.

The text matching tool is exhaustively tested in its current version using the CELEX data. It is currently being extended in order to consider surface linguistic information beyond wordform level. Other text matching algorithms are currently investigated as well.

The second version of the system prototype, incorporating the improved modules and more data resources is expected to be available in the beginning of 1995.

Available reports and deliverables

The following deliverables and reports have been issued and can be made available to interested parties :

Deliverables :

[1] Specifications of components, September 93 **

[2] Report on the linguistic analysis of TRANSLEARN corpora, September 93

[3] Selected Text Corpora, September 93

[4] Tagged and aligned corpora, May 94**

[5] Techniques adopted for text alignment, May 94**

[6] Database Design Principles, May 94**

Publications

Cranias L., Papageorgiou H., Piperidis S. "A matching technique in Example-Based Machine Translation", COLING 94 (to be published)

Papageorgiou H., Cranias S., Piperidis S. "Automatic Alignment in Parallel Corpora", ACL 94 (to be published)

** Although these reports are of restricted status, parts of them could be made available to broad public upon request.