IKTA5-146/02.

Development of an intelligent translation memory (Enhancing the qualiy and efficiency of machine-aided human translation by equipping translation memories with linguistic intelligence)

Project summary

Due to the extensive use of the Internet and the global exchange of textual information in general, there is an increasing demand for translation. In the early 21st century, the number and amount of texts created are rapidly growing, and these texts are very quickly spreading into foreign language environments. In the case of Hungary, the demand for translation is further increased by the approaching EU accession. During the last decade, it has been quite obvious that non-literary translations should be performed with the help of computational tools – in order to meet the deadlines with the translation tasks.

Developers are now extremely cautious when dealing with any form of machine translation, thanks to past and existing difficulties. As a result, they have begun to concentrate on helping certain subtasks of the translation process – rather than developing automatic machine translation solutions. There are now several tools on the market that facilitate the process of human translation.

In the present project proposal, we aim at enhancing what is believed to be the fundamental computational tool of translators – the translation memory (TM). The main task of a translation memory is to match sentences or parts of sentences in the current translation task to those stored in the TM's database. The difficulty here is that one cannot except many exact matches, so the program must also find sentences similar to the current one. If an exact or a proximity match is found, the translation memory should offer the corresponding translation stored in the database. There are a number of translation memory programs on the market (such as Trados TWB or ATRIL Déjà Vu, just to mention the most successful), no Hungarian products, however. By nature these tools are language independent, it is useless to talk about suiting them to the Hungarian language. However, the greatest drawback of these programs lies in their language independence: the proximity match is based on the mathematical processing of character codes in the sentences. (Note: in the terminology of existing TM's, a proximity match is usually called a fuzzy match, referring to the actual mathematical process used.) Character codes, however, do not infer any linguistic (syntactic or semantic) information in themselves. It can thus be said that the fuzzy match processes used in today's TM's are entirely inadequate in terms of assessing syntactic or semantic similarity. In turn, their implementation costs are low, and in some very special cases (like in many legal texts), they produce acceptable output. Thus the commercial developers are not interested in investing in alternative methods.

A linguistically intelligent translation memory – where not only character sequences, but abstract strings derived from linguistic analysis are compared – produces more exact and more adequate proximity matches, and is capable of recognizing more distant but linguistically adequate similarities as well. Such solutions, however, are language dependent because they need to be capable of parsing the sentences in the source language. At the earlier phases of the tool's life cycle, the range of available languages will be quite limited – our development uses English as the source and Hungarian as the target language, and will reverse the language pair if time permits. As a result, the market potential of such a tool is limited, compared to its language-independent counterparts, and it takes longer to cover the costs of development.

The main goal of this project proposal is developing a linguistically sound translation memory technology, and a prototype of a potential product built on this technology. During the 2 years of development, we will examine both the possibilities of using linguistic abstraction and the development of a more efficient mathematical (language-independent) proximity search solution. If both parts of the research are successful, we will implement a combination of both methods. The resulting technology will offer the advantages of the language independent translation memories (on a higher level), and will use linguistic matches if an appropriate linguistic database (lexicon) is present.

A great disadvantage of commercial translation memories is that they are usually delivered with an empty database. (Some producers and translation agencies are now tackling this problem.) In our project, we propose to develop an English-Hungarian parallel corpus of linguistic texts consisting of 1 million words per language. The corpus will have morpho-syntactic and shallow syntactic annotation, provided by MorphoLogic's existing linguistic tools, and corrected manually to some extent. This corpus will be used to examine linguistic similarity, and part of it will be used to build an initial linguistic database which makes the translation memory ready to recognize numerous phrases without requiring the user to enter large amounts of text. This reference corpus will be publicly available for research purposes.

The systems emerging from this project will initially offer *foreign*-Hungarian and Hungarian-*foreign* language pairs, one must consider that the proposal aims at creating a globally new technology. Implementing further language pairs and making the technology known abroad would enhance the image of Hungarian research and development.

The technology resulting from the project is ready to apply in English-Hungarian and Hungarian-English translation projects. After closing the two-year development process, it is possible to quickly add further languages to the system as the basic technology (the linguistic engines) will have a large language-independent part. We expect a dramatic increase in the overall efficiency of the human translators' work, and also an improvement in quality. As far as we know today, it is not only Hungary where the new technology will fill some gaps, but will be successful throughout the global translation industry.