

Computer aided translation – the cloud approach

Veno Pachovski¹, Eva Blazevska¹

¹ University American College Skopje, III Proleterska brigada br. 60,
1000 Skopje, Republic of Macedonia
{pacovski, blazevska}@uacs.edu.mk

Abstract. The need for automated translation grows very fast and so is the demand for creating tools to meet it.

The research presented represents a possible semi-automated solution i.e. a concept of cloud aided translation tool based on aligned corpora. Namely, the concept gives an opportunity to translate a text by keeping its structure (paragraph-wise) intact, using all other same language texts as context sources. The final goal is pairing the server and a user application in such a manner that the server which hosts the corpora (translation pairs of texts) is in communication with an application in which a source document is paired with the one user is working on.

Keywords: Natural language processing, translation, corpora

1 Introduction

The need for automated translation grows very fast and so is the demand for creating tools to meet it. Current situation in EU (60+ languages in Europe, 5 main languages and 23 of other member countries) and its commitment to multilingualism dictates further efforts and investments to be made. Within Horizon 2014-2020 which estimates more than 80 billion Euros, part of ICT and specifically for LT remains to be defined depending on the activity in the field. [1]

Macedonian language faces additional difficulties because of its relatively small population (less than 2 million), its specific grammar as well as alphabet, lacking corpora and basic language processing tools. All these mean that there is a real need for some original research. Also, any automatic solution requires rich context (the richer, the better), which means corpora upon which the translation system could base its decisions. This is an effort to do both.

1.1 The Motivation and Problem Approach

As a result of previous research and techniques and technologies involved ([2][3]), the approach adopted in this project is based on two important requirements.

The first one is that the translations are done paragraph-wise meaning that the translator is required to translate the text by paragraphs. Thus, the basic structure of the text is kept and even if (for various reasons) the sentences are not aligned (within the paragraph), at least the search for similar meaning for a word (or a phrase) i.e. the context (in future) is (will be) limited to that paragraph. Also, while the user is working on a paragraph, the search option is available, so that the user can search for possible translations of a particular word, or similar paragraphs, so that the sense of a word to be translated can be clarified.

The second, as a direct consequence of the first, is that the aligned pairs of translated texts result in an aligned corpora. Here, the working assumption is that as the corpora grow, the context will become richer and WSD when choosing an appropriate word (or phrase) will be easier. In time, the system may become (hopefully) a valuable platform for testing software and algorithms for automatic translations, considering that there will be paired paragraphs in various languages.

Also, considering that the system requires authentication of its users, it preserves the authorship, so the work can be copyright protected

Finally, inspired by the latest approaches to cloud storage (like DropBox), the idea appeared that a local application could be paired with the web server, so that a user can work on text in-directly, using a local application (which will again, in the background, use system resources).

2 Conclusion and future work

The system looks promising as a platform for future research. The first step is to make the web site popular among translating community (or least within University), so that the corpora can grow faster.

Next, the optimal protocol for communication between a desktop application and the web server (the cloud) should be devised, in order to follow the process of translation on the local machine, so that users can create and optimally coordinate their local resources with the cloud application.

References

1. European Commission : CORDIS : FP7 : ICT : Language technologies, <http://cordis.europa.eu/fp7/ict/language-technologies/>
2. Paskaleva, E., Pacovski V., Aligning the translations – a possible strategy for creation of aligned corpora (for South-Slavic languages). In: Proceedings of the Conference on Formal Approaches to South Slavic and Balkan Languages, pp. 113--117, (2006)
3. Nakov, P., Pacovski, V., Acquiring “False Friends” from Parallel Corpora: Application to South Slavonic Languages. In: Readings in Multilinguality. Selected Papers from Young Researchers in BIS-21++. Galia Angelova, Kiril Simov, Milena Slavcheva (Editors). Incoma Ltd. Shoumen, Bulgaria. (2006)