Multilingual Computer-based Communication and Language Processing: Lithuanian case

Bronius Tamulynas

Kaunas University of Technology, Department of Computer Networking,
Studentu 50, Kaunas LT-3028, Lithuania
Phone: +370 7 300369, fax: +370 7 300352
e-mail: bronius@pit.ktu.lt

Abstract. The article gives short overview of language processing technology in Lithuania and focuses on development the computer-based translation problem from English to Lithuanian. Common machine translation models are discussed, conceptual hierarchical model for computer-based translation system from English into Lithuanian is proposed. It is based on hierarchical blackboard architecture and includes virtual dictionary and several knowledge sources. It is shown, that such model and special set of knowledge sources with grammatical components may reduce the translation problem and improve its quality.

1 Introduction

Integration process of Lithuania into the EU and NATO stimulates development and maintaining of national culture as well as language, because documents of these organisations must be translated into all languages of its members. So, only modern and oriented to contemporary communication needs language can survive. No small language has yet avoided influence of larger languages. The Lithuanian language is no exeption because Lithuania geographically is located among the Slavonic nations, and has always been closely linked economicaly and politically with these nations. Therefore the most important goal of specialists of Lithuanian philology in the age of globalization is to keep the native language as authentic as posible. Specialists of linguistics, interested in practical use of the language, are searching for new ways to embody this purpose.

On April 19, 2000 the Government approved of support program for the Lithuanian language in information technology use for the years 2000-2006. It is expected that this would help the Lithuanian language to secure its position in computer software equipment. The program was prepared to help the Lithuanian language to get into the sphere of computers, otherwise it is under the threat of extinction. The absolute majority of personal computers used in Lithuania today have programs, which communicating with the user in English only. The program is oriented to support main language processing problems according to the information society program and growing needs of communication beetween the EU countries. One of the main tasks is to investigate computer-based translation possibility from the EU languages and back (English, French, Germany etc.). The government program aims at creating computer

software that would automatically translate texts of special documents (economical, informatics, politic, low, business) into Lithuanian.
The first step of this work was made in 1997 by creating Computer Based Lithuanian Language Learning System which satisfies needs of various users (school-children, students etc.) [1]. The Computer Based Lithuanian Language Learning System was considered as a part of the general Intelligent Tutoring System, which included the following modules: subject oriented tutoring modules, subject oriented data base modules (vocabularies, specific subject information), tasks and lessons making data base modules, and knowledge based students modules.
Currently, Lithuanian language processing research area consists of four main topics: speech recognition http://www.likit.lt; general and special electronic dictionaries (Lithuanian-English-Lithuanian, Lithuanian-Germany-Lithuanian as well as wide scale of Lithuanian language processing tools, http://www.led.lt and http://www.tilde.lt respectively, Lithuanian-English http://www.fotonija.lt; free on-line dictionary of computing http://foldoc.doc.ic.ac.uk/foldoc/index.html); 100 million words text corpora http://donelaitis.vdu.lt. Research on Computer-Based Translation (CBT) was initiated at KTU two years ago.
The article gives short overview of language processing technology in Lithuania and focuses on developing computer-based translation from English to Lithuanian. Common machine translation models are discussed, the conceptual hierarchical model of computer-based English-Lithuanian translation system is proposed. It is shown that such model and special set of knowledge sources with grammatical components may reduce the translation problem and improve its quality.

2 Computer-based translations into the Lithuanian

Computer-based translation systems are normally classified in terms of their basic translation strategy. Many early CBT systems as well as recent translation software for PC employ *direct* translation strategy. The next idea facilitating translation involves analysis of the source input into a *transfer* structure, which abstracts away many grammar details of the source language. *Interlingua* systems transforms source sentences into language-neutral representation from which target language is generated. Obviously, variations in a basic strategy are possible [2]: transfer system may incorporate different levels of abstraction in its representation, or hybrid of *interlingua* and *transfer* elements as well as combinations of the basic *direct* and *transfer* strategies and *corpus-based* techniques may be used.
Multilingual text processing includes a variety of technical issues and uses great amount of computational linguistics techniques. Analyzing and generating word forms is a very important step in the processing of natural languages. Nevertheless, translation process can be decomposed into independent parts so, that we can easily recognize multilevel structure of these parts. We suppose that general CBT module corresponds *blackboard* problem solving architecture, which arose from the **Hearsay II** speech understanding system [3]. In this basic model blackboard (BB) system is composed of tree main components: the *blackboard*, a set of *knowledge sources (KS)* and *control* mechanism.

In case of CBT architecture the blackboard is a global database that contains the source and target texts or some kind of hypotheses, linguistic rules etc. It could be structured as a hierarchy of levels, or particular classes of rules. The set of KS embodies the problem-solving knowledge, examines the state of the blackboard and modifies existing rules or content of BB database. In spite that, KS being independent and self-activating, the agenda-based *control* mechanism is needed to apply knowledge and to focus on the search process. Elements of *control* are: BB monitoring, the agenda, the scheduler and the focus-of-control database backward.

The CBT process consists of several independent knowledge sources, which determine content of programming components respectively [3],[4]:

- initial source text analyses (parsing, syntactic and semantic analyses according to grammar atributes);
- virtual dictionaries for CBT (general, special, phrase oriented....);
- translation engine (direct, transfer or interlingua techniques);
- target language analyses (syntactic, semantic, problem oriented...);
- knowledge based source text and target text analyses and updating (multilingual specialized text corpora).

The analysis of translation process shows that translation problem is rather complicated task: plenty word and phrase translation modes, ambiguity and multimeaning, context sensitive and pragmatics, problem oriented semantic space etc. In any way, within communicative, functional and cognitive approaches it is possible to identify scope of the growing need for non-literary translation. From that point of view, proposed conceptual CBT model gives a chance to put into reality very important system features: flexibility, consistent renovation and upgrading possibilities. Many of CBT steps can be performed in parallels. According to this, CBT model paradigm specialized English translation system into Lithuanian is created. The main programming components, which implement declared functions of KS, are:

- virtual electronic dictionaries (general, phrase-based, multilingual text corpora...);
- syntactic sentence analyzer;
- computational morphology;
- semantic source and target mapping, evaluation of translation semantic adequacy;
- problem oriented knowledge extraction from special text corpora and its use for semantic space development.

Users or translation experts can be interpreted as knowledge sources if they take active role in the interactive CBT process. Control strategy in the CBT BB model architecture implies agenda-based KS action planning and supports feedback between database BB levels and KS as well as translation quality control and search complexity management. User or expert-translator can implement control function.

CBT uses virtual electronic English-Lithuanian dictionary. English word includes grammar attributes. Corresponding Lithuanian word is connected with its grammar attributes as well. Direct translation strategy with some transfer elements to compile syntactic sentence groups [6],[7],[8] for extracting more semantic knowledge from the source text and transfer it into the target text is used. Translation process implies KS interface translation environment interacting with Word, Web and Outlook Express.

Detailed analyses of *interlingua* principle show that quality of translation is not satisfactory due to many reasons [5]. However, in general, rather universal *interlingua*

frame is very perspective and enables to create large-scale language translation system for a small group of languages from many-to-one and back. It could be done by efforts of all these countries in the frame and support of the EU funding. As a partial solution, *interlingua* frame is very exiting to elaborate *intelligent* source text parsing to extract language knowledge components and apply to generate target text.

3 Conclusions

The article gives short overview of language processing technology in Lithuania and focuses on developing the computer-based translation from English to Lithuanian. Common machine translation models are discussed, conceptual hierarchical model for computer-based translation system is proposed. Detailed analysis of *interlingua* principle shows that universal *interlingua* frame is very promising and enables to create large-scale language translation system for small group of languages from many-to-one and back. All these countries in the frame and support of the EU funding could do it. According to the CBT model paradigm a system for specialized English translation into Lithuanian is created. Direct translation strategy with some transfer elements of syntactic sentence groups is used. It allows implementing better translation quality for more complicated sources.

Acknowledgements

The author is thankful to graduate students A.Sidorov and P.Pacevicius for the first version of CBT environment and virtual translation dictionary implementations. The analysis of use *interlingua* posibilities and elaboration of syntactic sentence groups were done by PhD student G.Misevicius and M.Žemaitis to whom I am grateful.

References

1. Baniulis, K., Tamulynas, B.: Flexible Learning in an Intelligent Tutoring Environment. In: Kommers, P., Dovgiallo, A., Petrushin, V., Brusilovsky, P. (eds): New Media and Telematic Technologies for Education in Eastern European Countries. Twente University Press, Enschede (1997) 395-409.
2. Trujillo, A. Translation Engines: Techniques for Machine Translation. Springer. 1999.
3. Carver, N., Lesser, V. The Evolution of Blackboard Control Architectures. In: Expert Systems with Applications. (1993).
4. Arnold, D.L., Balkan, L. and others. Machine Translation: An Introductory Guide. Colchester, (1993).
5. Misevicius, G. *Interlingua* principai kompiuterinio vertimo sistemose i lietuviu kalba. In: KTU konf. "Informacines technologijos 2002". Kaunas (2002) 311-316.
6. Tamulynas, B., Žemaitis, M. Sintaksiniu sakinio grupiu sudarymas ir ju panaudojimas kompiuterinio vertimo programose. KTU konf. "Informacines Technologijos 2002". Kaunas (2002) 317-320.
7. Dobrovolskis, B., Kniukšta, P. ir kt. Lietuviu kalbos žinynas, "Šviesa" , Kaunas (2000).
8. Piesarskas, B., Didysis anglu-lietuviu kalbu žodynas, Vilnius (1999).