Machine Translation for Latvian

Inguna Skadiņa
Tilde,
Institute of Mathematics and Computer Science, University of Latvia
Riga, Latvia
Inguna.Skadina@tilde.lv

1. Introduction

Idea of a translation machine is older than the first computer. Today for many people global communication is part of everyday life, in which translation tools can facilitate communication and help to overcome language barriers.

Similar to other languages with rather small number of speakers Machine Translation (MT) system for Latvian is required by users who are not translators, but interested in getting a draft translation of the text from a language, which they do not know very well or do not know at all. Perhaps the user needs is one of the reasons why during last 50 years there have been several attempts to develop MT system for Latvian.

The first MT system for Latvian was developed in the beginning of 60-ies at the Institute of Electronics (Гобземис А.Ё., Горобец В.Г., Юрик В.А., Якубайтис Т.А., 1961). The system was a typical product of that time: it used word to word translation strategy to translate scientific texts from Russian into Latvian. However, since the system provided translation between two highly inflected languages, good morphological analysis tools were developed as part of the system.

In mid 90-ies, after more than 30 years of silence in this field, Artificial Intelligence Laboratory of Institute of Mathematics and Computer Science (IMCS) has started work on an MT system for Latvian. LATRA model (I. Greitāne, 1997) is an interlingua system based on ideas of SWETRA system (Sigurd, 1988, 1990). Similarly to SWETRA, the first version of LATRA was developed for translation of weather reports and stock market texts. At present, this system has been modified for more general text types.

In 1996 IMCS joined the Universal Neworking Language (UNL) project (Hiroshi U., Zhu M., Della Senta T., 1999). The aim of the UNL project is to overcome the language barrier by storing information into UNL and converting it into human language. Through the project basic converting rules that allow translation from UNL into Latvian are developed.

At the same time direct MT system for aviation documentation was proposed at the Riga Aviation Institute (Ореховский В.А., Мишнев Б.Ф., 1995).

Besides research prototypes company Tilde aims at development of a commercial MT system for the Latvian market. Recent developments (advanced electronic dictionaries, morphological processing tools and limited syntactic parser) are the first steps towards a commercial product.

2. LATRA: first interlingua MT prototype for Latvian

The Artificial Intelligence Laboratory of IMCS has started research on MT in 1993 when the first Latvian morphology tools for personal computers were developed. It has been supported by the Latvian Council of Sciences through several projects: “Limited Model of Automated Machine Translation System for Latvian” (1993-1996) and “Development of Probabilistic Methods for Automated Disambiguation of Natural Language Texts and Applications for Machine Translation” (1997-1999).

LATRA is implemented in Prolog programming language. The main constituents of LATRA are lexicon, morphology rules, syntax rules and rules for work with the functional representation (interlingua).

The lexicon in LATRA is implemented as a set of predicates. For rapid access the lexicon is divided into different sections identified by different predicates. For each lexical

item morphological, syntactic and semantic features are stored. In addition the 'meaning' of the word in so-called ‘Machenese’ English is stored into lexicon.

Implementation of LATRA parser was strated with the research of typical Latvian sentence structures. Although stock market texts have rather simple structure, other common syntactic structures were included in the parsing constituent.

Originally SWETRA system was designed for languages with fixed word order, while Latvian language has a rather free word order. Therefore Latvian parser uses predicate grammar introduced for Russian and Polish parsers in SWETRA (Gawrońska B., 1993).

Since LATRA is implemented in Prolog, phrase structure rules used for parsing can are used in generation process also.

The interlingua format used in LATRA and SWETRA is called functional representation. It consists of nine functional concepts (constituents): a subject, a predicate, two objects, four adverbials and a coordinator. Word meanings in the functional representation are given in so-called ‘Normalized Machinese English’ (Sigurd, 1994). Figure 1 shows translation from Latvian into English through functional representation.

```
Ceturtdienas apgrozība bija 4 miljoni.

[subj(s(s(m(thursday,prop),[]),m(trading,sg),[])),pred(m(m(be,past),[
])),obj(s(4,m(million,pl),[])),obj([]),advl([]),advl([]),advl([]),adv
l([]),co(s(s(m(thursday,prop),[]),m(trading,sg),[]),[.])]

Thursday’s trading was 4 million.
```

Figure 1. Latvian-English translation through functional representation.

3. Universal Networking Language

Development of Universal Networking Language was started in 1996 at the Institute of Advanced Studies of United Nations University.

In 1997 Artificial Intelligence Laboratory of IMCS was invited to join UNL project. That time 15 languages (Arabic, Chinese, English, French, German, Hindi, Indonesian, Italian, Japanese, Latvian, Mongolian, Portuguese, Russian, and Spanish) were involved in the development of language tools for communication through UNL.

UNL is an artificial language to express and exchange every kind of information in the form of a semantic network. It consists of the UNL Relations, the UNL Attributes, the Universal Words and the UNL Knowledge Base. The Universal Words make vocabulary of the UNL, the relations and attributes constitute the syntax of the UNL, and the UNL Knowledge Base constitutes the semantics of the UNL.

For Latvian language work was concentrated in two directions: development of a lexicon and development of conversion rules from UNL into Latvian.

In UNL concepts are represented by universal words (UW). A UW is made up of a character string (an English-language word) followed by a list of constraints. Constrains are used to define the concept through restriction of possible meaning of the English word.

In lexicon Universal Words (UW) are used to describe meanings or make correspondence between word in a natural language and UW. For the Latvian language, a general purpose lexicon with 5 000 words and a lexicon for sports domain, containing approximately 10 000 words, were developed.

Although the structure of Universal Word might be complicated, in most cases Universal Words are understandable to a lexicographer. Latvian lexical item contains a stem of the Latvian word in square brackets, Universal Word in quotes, and grammatical information in parenthesis (Milčonoka E., 2000). For instance lexical item for UW *austria(icl>country)* will be [Austrij] "austria(icl>country)" (N,FEM,4 DECL,SG).

UNL represents sentence information as a list of interrelated (semantic) labeled links, each between two of the concepts present in the sentence (Figure 2).

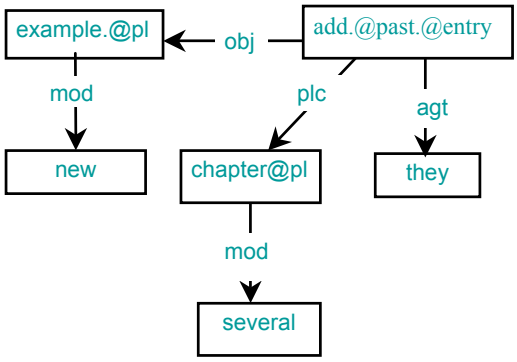

Figure 2. Sentence *They added new examples in several chapters* in UNL

The UNL system includes two tools - enconverter and deconverter. The encorverter converts natural language text into UNL, while deconverter converts UNL expression into natural language.

For Latvian language about 950 deconversion rules for translation of UNL sentence into natural language sentence are developed.

Latvian deconverter contains three rule types: syntax rules (420), agreement rules (150) and morphology rules (200). When a Latvian sentence is generated syntax rules are applied first. After a word is inserted, agreement rules are performed. Through agreement rules morphological features such as gender, number and case are inherited from the head constituent to the child constituents. Finally morphology rules generate the necessary word form.

4. Current situation and future perspectives

The work started through LATRA and UNL projects is being continued in the project "Automated synthesis of language independent text representation" funded by the Latvian Council of Science (2001-2004).

At present research is concentrated on semantic aspects of text analysis and development of a general purpose translation system.

The parsing and generation constituents are based on LATRA principles of constituent analysis, but more complicated semantic analysis and lexical descriptions are our current tasks. Finally we would like to move from SWETRA functional representation to UNL semantic representation.

At present Latvian lexicon has been improved by implementation of semantic types proposed in SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) project [Lenci et al. 2000].

The SIMPLE recommendations have distinguished between three components of the model: the set of semantic units, the ontology and the template. Our main interest is in ontology, which provides the Conceptual Core shared by all lexicons of SIMPLE. For Latvian 629 noun meanings and 508 verb meanings were encoded using semantic classes proposed in SIMPLE (Skadiņa I., 2003).

Semantic classes are part of SIMPLE semantic types, which can be simple or unified. The unified types, which are multidimensional types, are the next step for the Latvian semantic dictionary. SIMPLE templates of semantic classes will be used for this purpose.

5. Tendencies in the industry

MT is not only an interesting subject for research, it is also a popular product in the market. Although quality of MT output is still far from human translation it is widely used and becomes more and more popular since Internet becomes one of the ways of communication.

Bearing this in mind company Tilde aims to develop a translation system for Latvian users.

At present Tilde has a stable position in the market of HLT products not only in Latvia, but also in Lithuania (Vasiļjevs A., Greitāne I. 2001). Tildes product for Latvia *Tildes Birojs* has three generations. The latest version includes advanced translation dictionaries for the following languages: Latvian-English-Latvian, Latvian-Russian-Latvian, Latvian-German-Latvian. Tilde's electronic dictionaries are a composition of general and terminological dictionaries forming the most complete computer word-stock of Latvian words.

The electronic dictionary can be used in several ways: as a separate software package; as a translation tool for a marked word in the text, for instance, in Internet Explorer, and as part of translation software in the latest versions of MS Word.

The dictionary incorporates morphological analysis tools, allowing translation of any word form. This is an important issue for inflected languages – Latvian and Lithuanian – where base form can differ from one word form to another.

Tilde has chosen a step-by-step strategy where every new generation of a product is one step towards the translation system. While in the current product multilingual translation dictionaries and a limited parser are included, a phrase translation tool is planned in the next product version.

References

Gawrońska B. 1993. An MT Oriented Model of Aspect and Article Semantics. Lund: Lund University Press.

Greitāne I. 1997. Mašīntulkošanas sistēma LATRA. In *Latvijas Zinātņu akadēmijas vēstis,* (3), pp. 1-6.

Hiroshi U., Zhu M, Della Senta T. 1999. The UNL, A Gift for a Millennium, UNU Institute of Advanced Studies.

Lenci A., Busa F., Ruimy N., Gola E., Monachini M., Calzolari N., Zampolli A., et al. 2000. SIMPLE Linguistic Specifications (Project Deliverable D2.2).

Milčonoka E. 2000. Vārda raksturojums datorleksikonā. In *IX Starptautiskā baltistu kongresa "Baltu valodas laikmetu griežos" referātu tēzes*. pp. 209-211.

Sigurd B. 1988. Translating to and from Swedish by SWETRA - a multilanguage translation system. In: Maxwell,D., Schubert,K. & Witkam,A.P.M. (eds) *New Directions in Machine Translation*, Conference Proc. Budapest 18/19-8-1988. Budapest: John von Neuann Society for Computing Sciences/Dordrecht:Foris

Sigurd B., M.Eeg-Olofsson, B. Gawrońska-Werngren & P.Warter. 1990. Swetra - a multilanguage translation system. Lund: Institutionen för lingvistik.

Sigurd B. 1994. Swetra "Referent Grammar". In *Computerised Grammars for Analysis and Machine Translation*, pp. 7-56. Lund: Lund University press

Skadiņa I. 2003. Electronic Dictionaries and Multilingual Information Society. In *Terminology and Technology Transfer in the Multilingual Information Society*, Termnet Publisher, pp. 140-146.

Vasiļjevs A., Greitāne I. 2001. Baltic Challenges on the IT Frontier: Language and Culture. In: Baltic IT&T Review, No. 2 (21).

Гобземис А.Ё., Горобец В.Г., Юрик В.А., Якубайтис Т.А. 1961. О машинном переводе с русского языка на латышский. In *Автоматика и вычислительная техника*, pp. 149–164.

Ореховский В.А., Мишнев Б.Ф. 1995. Алгоритм идентификации слов естественных языков и его применение при обработке текстовых документов. In *Автоматика и вычислитительная техника*, No 3, pp. 36-47.