UNIVERSITY OF LATVIA

RAIVIS SKADIŅŠ

Combined Use of Rule-Based and Corpus-Based Methods in Machine Translation

Doctoral Thesis

RIGA 2011

LATVIJAS UNIVERSITĀTE
UNIVERSITY OF LATVIA

UNIVERSITY OF LATVIA
FACULTY OF COMPUTING

Raivis Skadiņš

Combined Use of Rule-Based and Corpus-Based Methods in Machine Translation

DOCTORAL THESIS

FOR PH.D. (DR. SC. COMP.) ACADEMIC DEGREE

FIELD: COMPUTER SCIENCE

SECTION: SOFTWARE AND SYSTEMS ENGINEERING

Scientific Advisor

Dr. Phys.
Andrejs Spektors

RIGA, 2011

This work has been supported by the European Social Fund within the project «Support for Doctoral Studies at University of Latvia».

Advisor:

Dr. Phys. Andrejs Spektors
Institute of Mathematics and Computer Science, University of Latvia

Referees:

The defence of the thesis will take place in an open session of the Council for Promotion in Computer Science, University of Latvia, _____, in the Institute of Mathematics and Computer Science, the University of Latvia (Rīga, Raiņa bulv. 29, room 413).

The thesis and its summary are available at the Library of the University of Latvia (Kalpaka bulv. 4, Rīga).

Head of the Council J.Bārzdiņš

ABSTRACT

Machine Translation (MT) systems are built using different methods (knowledge-based and corpus-based). Knowledge-based MT translates text using human created rules. Corpus-based MT uses models which are automatically built from translation examples. Both methods have their advantages and disadvantages. This work aims to find a combined method to improve the MT quality combining both methods.

An applicability of the methods for Latvian (a small, morphologically rich, under-resourced language) is researched. The existing MT methods have been analyzed and several combined methods have been proposed. Methods have been implemented and evaluated using an automatic and human evaluation. The factored statistical MT with a rule-based morphological analyzer is proposed to be the most promising. The practical application of methods is described.

Keywords: Machine Translation (MT), Rule-based MT, Statistical MT, Combined approach

ANOTĀCIJA

Mašīntulkošanas (MT) sistēmas tiek būvētas izmantojot dažādas metodes (zināšanās un korpusā bāzētas). Zināšanās bāzēta MT tulko tekstu, izmantojot cilvēka rakstītus likumus. Korpusā bāzēta MT izmanto no tulkojumu piemēriem automātiski izgūtus modeļus. Abām metodēm ir gan priekšrocības, gan trūkumi. Šajā darbā tiek meklēta kombināta metode MT kvalitātes uzlabošanai, kombinējot abas metodes.

Darbā tiek pētīta metožu piemērotība latviešu valodai, kas ir maza, morfoloģiski bagāta valoda ar ierobežotiem resursiem. Tiek analizētas esošās metodes un tiek piedāvātas vairākas kombinētās metodes. Metodes ir realizētas un novērtētas, izmantojot gan automātiskas, gan cilvēka novērtēšanas metodes. Faktorēta statistiskā MT ar zināšanās balstītu morfoloģisko analizatoru ir piedāvāta kā perspektīvākā. Darbā aprakstīts arī metodes praktiskais pielietojums.

Atslēgas vārdi: mašīntulkošana (MT), zināšanās balstīta MT, korpusā balstīta MT, kombinēta metode

ACKNOWLEDGEMENTS

I wish to thank my supervisor Dr. Andrejs Spektors who gave me the freedom to explore many paths of research, encouraged me and gave me guidance along the way.

This work has been supported by the European Social Fund within the project "Support for Doctoral Studies at University of Latvia". The thesis is based on the research and development done in several projects at Tilde, including the European Union Seventh Framework Programme (FP7/2007-2013) project ACCURAT (grant agreement no 248347), the ICT Policy Support Programme (ICT PSP) project LetsMT! (grant agreement no250456) and the European Union EUREKA's Eurostars Programme project SOLIM (grant agreement no E! 4365).

Tilde has been a great place for me to work and develop myself as a researcher. I have received a lot of encouragement, inspiration, feedback, and insights that helped me with my work.

I would like to thank Daiga Deksne, Kārlis Goba, Linda Goldberga, Tatiana Gornostay, Māris Puriņš, Inguna Skadiņa, Valters Šics, Jörg Tiedemann, and Andrejs Vasiļjevs who have been great colleagues in my research and are co-authors of my main scientific papers. I acknowledge their valuable contribution in the research and development work described in this thesis work. I would like also to thank colleagues from partner organizations in the European Union projects, especially people from University of Edinburgh, Uppsala University, SemLab BV, and AGM-lab.

On a personal note, I would like to warmly thank my wife and family for their endurance and support.

CONTENTS

1 INTRODUCTION

1.1 Research area

The research area in the focus of this research is the Machine Translation (MT). The machine translation originated in 1949, when Warren Weaver proposed applying cryptographic and statistical techniques from the field of communication theory to deal with the problem of text translation.

"When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode'" (Weaver, 1955)[1]

However, early efforts in this direction brought doubts about whether high quality, automatic, machine translation was possible in principle. In 1966 the ALPAC (Automatic Language Processing Advisory Committee) report concluded that machine translation could not overcome the low cost of and low demand for human translators, stopping most machine translation research at that time (Pierce and Carroll, 1966).

Machine translation became the domain of rule-based (or knowledge-based) systems. These systems were very labor-intensive, as they translation translated texts using hand created dictionaries, patterns, rules and exceptions.

The next breakthrough in machine translation came in the early 1990's when bilingual corpora became available. Machine translation team at IBM proposed a statistical approach to machine translation based on a probabilistic dictionary (Brown et al., 1990). Soon after, methods for automatic aligning of sentence pairs in a bilingual corpus were developed by Gale and Church (1991) and Brown et al. (1991). Brown et al. (1993) described five statistical models for machine translation, called the IBM Models. IBM models make the foundation for word-based statistical machine translation. These models show how a text could generate its word by word translation using Shannon's (1948) noisy channel model of communication, estimating parameters from a parallel corpus.

Machine Translation systems are built using different approaches today. Historically, first MT systems were rule-based and some of these systems are still in use commercially. But many

[1] Reprinted from a memorandum written by Weaver in 1949.

modern MT systems are built using corpus-based methods (statistical and example-based methods).

Machine translation has not reached a quality level where it can replace a human translator, and most probably it will not reach such level in a near future. However, machine translation has proven to be a very useful tool in such scenarios as gisting information in unknown languages and providing raw translation for post-editing. The need for fast and cheap translation has resulted in a number of commercial products (e.g. Systran, Promt, Reverso, LanguageWeaver) and several translation solutions are freely available as Web services (e.g. Google Translate, Bablefish, Bing Translator, Tilde Translator) demonstrating acceptable translation quality for widely used languages.

Cost-effectiveness is one of the key reasons why the statistical paradigm has come to be the dominant current framework for MT theory and practice, as it has proven to be the most effective solution both from the point of view of time and labor resources and translation output quality. As such, statistical approaches to MT have become the major focus for many research efforts (Hutchins, 2007b).

Until now SMT research has been mainly focused on widely used languages, such as English, German, French, Arabic, and Chinese. For “small” under-resourced languages MT solutions, as well as language technologies in general, are not as well developed due to the lack of linguistic resources and technological approaches that enable MT solutions for new language pairs to be developed cost effectively. This has resulted in a technological gap between these two groups of languages.

Nevertheless, online Google Translate which is freely available as an on-line service broadens the set of translation language pairs, incorporating the Baltic languages and many others. However, this service performs poorly on narrow domain texts. Typically such online translation solutions are exploited by occasional users to translate short texts.

The EuroMatrix project represents a major push in MT technology, applying the most advanced MT technologies systematically to all pairs of EU languages. The EuroMatrixPlus project is continuing the rapid advance of MT technology, creating sample systems for every official EU language and providing other MT developers with the infrastructure for building statistical machine translation models.

1.2 Motivation of the research

Despite the long history of research in machine translation field and many existing MT systems, the original goal of replacing human translators has not been met – current systems are far from being able to produce output of the same quality as a human translator (Hutchins, 2006).

Both rule-based and corpus-based machine translation methods have their advantages and disadvantages.

Rule-based MT systems provide a high quality translation if they have all necessary knowledge. Rule-based MT systems typically deal better with difficult language phenomena such as inflections, word agreements, long distance dependency, long distance reordering etc., they are better also in translating narrow domain texts or texts written in controlled languages. The output of rule-based MT systems is more consistent and predictable and it is easy to trace and fix the cause of translation mistakes. But the real human language is complex with many exceptions and ambiguities, usually dictionaries are deficient too. Therefore it is impossible to provide all knowledge necessary for high quality rule-based MT systems. It is possible to advance rule-based MT systems, but they can be advanced only to the certain level and further advancement becomes too complex and labor-intensive. A good linguistic and domain expertise is needed to create a rule-based MT system. Rules created for one rule-based MT system typically are not easy adaptable to other language pairs and domains. The rule-based MT is a big challenge for low resource languages due to the need of bilingual dictionaries and morphological and syntactic tools.

Corpus-based MT systems do not need handcrafted dictionaries and rules, they automatically learn linguistic phenomena of the languages from large corpora, they can be easily improved just by adding more training data, they can be easily adapted to new language pairs and domains, and they can easy learn phrasal and idiomatic expressions from the data. But, although latest corpus-based MT systems can be improved by adding models dealing with morphology and syntax of particular language pair, there still is a need for better MT systems dealing with highly inflectional and structurally complex languages. Corpus-based MT systems are still struggling with inflections, word agreements, long distance dependency, and long distance reordering. A large and good quality parallel corpus

is crucial for corpus-based MT, but often it is not available for smaller and less resourced languages.

As both MT methods have advantages and disadvantages and often one is weaker where other is stronger, it is reasonable to look for new MT methods combining existing approaches to overcome disadvantages of one method using advantages of the other. Latvian is a small language with a complex grammar and limited parallel corpus; therefore effective use of corpus based methods is difficult. Use of knowledge based methods is difficult due to complex grammar. It could be possible to get better results combining both approaches.

1.3 The aim of the research

This research focuses on the problem of combined use of rule-based and corpus-based methods in machine translation.

This work describes the issues related to machine translation, it describes limitations of current rule-based and corpus-based MT methods and gives suggestions how both methods can be combined to achieve a better MT quality.

The main challenge is to find method allowing to improve quality of MT for Latvian – a small, morphologically rich and under-resourced language, where none of current MT methods give good results. The use of rule-based methods is difficult and inefficient because we luck large and high quality dictionaries, high quality morphological and syntactic analyzers, and method by itself is too demanding with regard to human resources. The use of corpus-based methods is difficult because we luck a parallel corpus of reasonable size and current corpus-based methods still do not perform well working with morphologically rich languages. It is also important to look for new methods allowing adapting MT for an effective use in specialized domains and for a practical use in general.

Although the primary focus of this research is on MT issues related to Latvian, the aim of the research is to find methods which are applicable to other languages too.

For his research author has established the following hypothesis:

It is possible to achieve a better MT quality by combining knowledge and corpus based MT methods. And state-of-the-art corpus-based MT systems are flexible enough to be extended

with knowledge-based components and such extended MT systems provide a higher quality translation.

The goal of this research is to create a combined MT method that provides MT with a higher translation quality. That encompasses all the following major aspects:

- An analysis of both knowledge and corpus based MT methods looking for the ways how combine them;
- Experiments with different combined MT methods;
- MT quality evaluation;
- Applicability of methods for Latvian;
- Practical applications of MT, including MT tailoring to different domains;

1.4 Key results

The major contributions of this thesis are as follows:

- All major existing rule and corpus-based MT techniques have been researched focusing on ways how they can be combined to increase MT quality;
- Several combined methods have been implemented and evaluated:
  - Rule-based MT with a statistical disambiguation component (Skadiņš et al., 2008);
  - Rule-based MT with a dictionary extracted from a parallel corpus using statistical methods;
  - Factored phrase-based statistical MT with a rule-based morphological analyzer (Skadiņš et al., 2011a; 2011c; 2010; Vasiļjevs et al., 2011a);
  - Factored phrase-based statistical MT knowledge form the ontology (Skadiņš, 2010; 2011; Skadiņš et al. 2011b);
- Advantages and disadvantages of the proposed methods are analyzed;
- The method giving the best MT quality improvement is factored phrase-based statistical MT with rule-based morphology (Skadiņš et al., 2011a; 2011c). This method outperforms all other known English-Latvian MT systems in automatic evaluation achieving 35.0 BLEU points. The method is also tested on English-Lithuanian language pair and is applicable to other morphologically rich languages.

1.5 Practical implementation

The best English-Latvian machine translation system proposed in this research, combines statistical MT methods with a rule-based morphology. The system is available as a free online service http://translate.tilde.lv, it is included in a software package Tildes Birojs 2010 and it has been tested in a practical use for software localization where it helped to achieve the 32.9 % productivity increase (Skadiņš et al., 2011a; Vasiļjevs et al., 2011a).

Also the rule-based MT system with the statistical disambiguation component (Skadiņš et al., 2008) was released and included in a software package Tildes Birojs 2008; Latvian-Russian language pair of this system is also available as a free online translator on website http://translate.tilde.lv.

1.6 Author's publications and presentations related to the research

The author has presented the results of the research at 10 international conferences, workshops and seminars:

- Machine Translation Summit XII, Xiamen, China, 2011;
- The 15th International Conference of the European Association for Machine Translation - EAMT 2011, Leuven, Belgium, 2011;
- The 18th Nordic Conference of Computational Linguistics - NODALIDA 2011, Riga, Latvia, 2011;
- Research Workshop of the Israel Science Foundation - Machine Translation and Morphologically-rich Languages, Haifa, Israel, 2011;
- The Ninth International Baltic Conference DB&IS 2010, Riga, Latvia, 2010;
- The Fourth International Conference Baltic HLT 2010, Riga, Latvia, 2010;
- The Sixth Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 2008;
- The Third Baltic Conference on Human Language Technologies, Kaunas, Lithuania, 2007:
- The 16th Nordic Conference of Computational Linguistics – NODALIDA 2007, Tartu, Estonia, 2007;
- The Second Baltic Conference on Human Language Technologies, Tallinn, Estonia, 2005;

Research results are reported in the 17 papers published in the proceedings of the international conferences (see list of author's publications on page 100).

1.7 Outline of the thesis

The remainder of this document is structured as follows:

- Chapter 2 surveys existing MT methods and reviews advantages and disadvantages of different approaches.
- Chapter 3 describes other related work in the area of combining different MT approaches.
- Chapter 4 gives detailed description of methods for MT combining researched in this work. It starts with an overview of possible combination methods and their relevance to the aims of this research, then it gives detailed description of evaluation methods used in this work, then it gives detailed description of 4 major experiments made in this research, and finally it outlines possible future research directions.
- Chapter 5 summarizes results of the research and gives conclusions about this work.

2 BACKGROUND

MT has been a particularly difficult problem in the area of natural language processing since its beginnings in the early 1940-ies. From the very beginning of MT history, three main rule-based MT strategies have been prominent: direct, interlingua, and transfer. Rule-based MT strategy with a rich translation lexicon showed good translation results and found its application in many MT systems, e.g. Systran, Promt and others. However, this strategy requires immense time and human resources to incorporate new language pairs or to enhance translation quality. The more competitive SMT approach has occupied the leading position since the first research results performed in the late 1980s with the Candide project at IBM for English-to-French translation system (Brown et al., 1988a,b). The SMT strategy, first suggested in 1949 by Warren Weaver (1955) and then abandoned for various philosophical and theoretical reasons for several decades until the late 1980s (Brown et al., 1993), has proven to be a fruitful approach to foster development of MT. Cost-effectiveness and translation quality are the key reasons that the SMT paradigm has become the dominant current framework for MT theory and practice. As such, statistical approaches to MT have become the major focus for many research efforts (Hutchins, 2007b).

2.1 Rule-Based MT

Classical rule-based MT systems perform deep analysis of linguistic phenomenon of the given language pair. Rule-based systems usually consist of an MT engine and a set of transformation rules written by human expert or linguists. Linguistic knowledge is integrated into the MT system through those transformation rules. Rule-based MT engine is categorized into three different architectures: (i) direct, (ii) transfer and (iii) Interlingua. Direct MT is the primitive form of translation that replaces source language word with the target language word; Interlingua approach is based on in-depth semantic analysis whereas transfer approach is based on syntactic level and also deals with semantic on small scale. Transfer-based MT engine consist of three components: analysis, transfer, and generation. Source sentence is analyzed using parsers and morphological tools, gets transformed into intermediate representation using the transfer rules, and then target language sentence is generated from the intermediate representation.

Despite the impressive advances in statistical MT in the last 20 years, rule-based machine translation remains appropriate when the goal is publication quality translation of documents, especially within a restricted domain (Ranta et al., 2010). Furthermore, SMT is ruled out when bilingual corpora are unavailable for the language pairs of interest.

2.1.1 Interlingua

When human translators translate text from one language to the other, they read words in source language and form an understanding of the text guided by their knowledge of source language syntax and semantics. Then they rewrite it in target language using their knowledge of target language semantics and syntax. This view of translation is illustrated by the machine translation pyramid in Figure 2.1.1 (Vauquois, 1968).

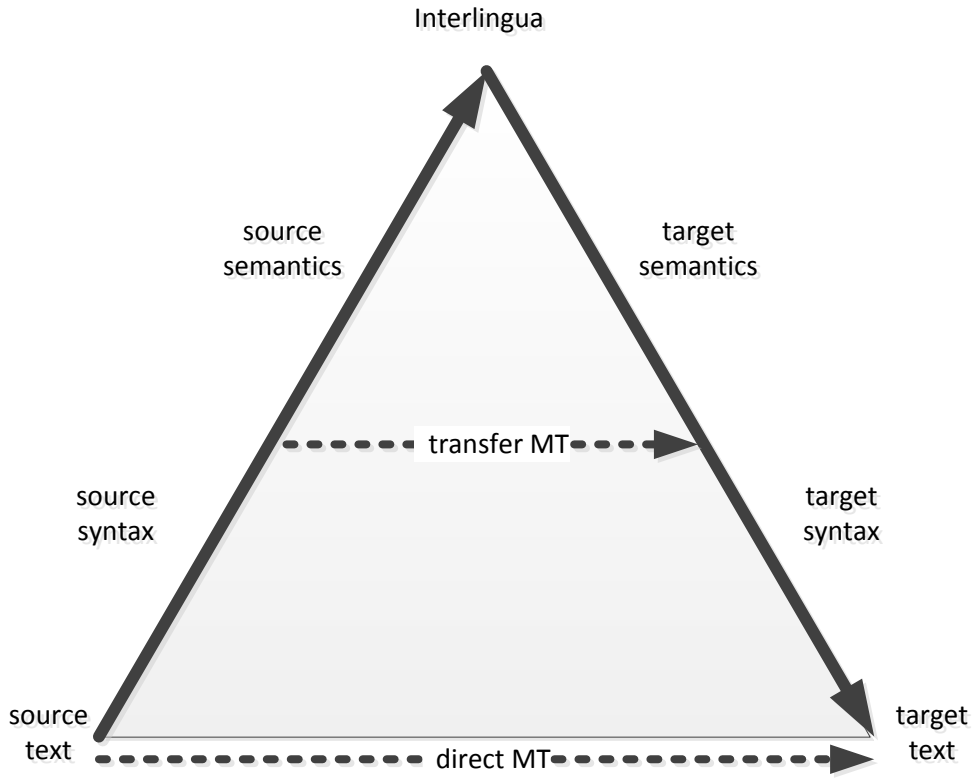


Figure 2.1.1 The machine translation pyramid

Efforts to build machine translation systems that follow this model of the translation process face challenges:

- the acquisition of syntactic and semantic knowledge to transform source text into language independent meaning representation;
- the generation of target text from the meaning representation.

But above all, the representation of meaning in the form of an Interlingua that is truly beyond language is really the hardest problem, since human understanding of meaning representation is very limited and often bound to our language.

Still, this did not stop attempts to build interlingua-based machine translation systems. These approaches are also called knowledge-based, since they require a vast amount of knowledge resources (lexicons, grammar rules, and especially world knowledge) to transform words into meaning representations.

A more detailed description of knowledge-based machine translation is given by Nirenburg et al. (1992) and Arnold et al. (1994), See also the description of the KANT system by Nyberg et al. (1992), which is an implementation of this approach.

Another attempt to build an Interlingua MT system is based on the Universal Networking Language (UNL) elaborated in the United Nations University by H. Uchida (Uchida et al. 1999, Uchida and Zhu 2001). UNL has an expressive power to represent relevant information conveyed by natural languages. The UNL consortium has developed modules for translating texts from different languages to UNL and vice versa. In UNL, the process of representing natural language sentences in UNL graphs is called enconverting, and the process of generating natural language sentences out of UNL graphs is called deconverting. The enconverting involves natural language analysis and understanding and it is supposed to be carried out semi-automatically. The deconverting is expected to be done fully automatically.

Interlingua MT approach is often used in combination with controlled natural languages. Mitamura et al. (1995) introduces controlled English in KANT MT system. Controlled languages such as Attempto Controlled English (ACE) by Fuchs and Schwitter (1996) are used also for OWL ontology authoring (Schwitter et al. 2008) and verbalization (Kaljurand, 2007; Grūzītis et al., 2010). As there are means for OWL ontology authoring and for ontology verbalization using controlled languages we can speak about the semantically precise translation (via OWL as Interlingua) among the controlled languages. (Vaivads, 2010)

Today, successful Interlingua or knowledge-based machine translation systems are limited to small domains where it is feasible to assemble the required knowledge. However, the challenge to scale such systems to larger domains (for example, news text) is one motivation behind various research efforts to build up such knowledge resources; and despite overall skepticism in MT community there are some Interlingua MT systems in use and even new

are emerging. For example, Synchronous Extensible Dependency Grammar based MT by Gasser (2011) and ABBYY Compreno MT system which is positioned as a knowledge-based MT system[2] (Pakhomov, 2011).

2.1.2 Transfer-based MT

Transfer-based machine translation methods are related to knowledge-based Interlingua methods in the sense that they also try to climb up the machine translation pyramid, but in contrast, not all the way to the top. The transfer from source language structure to target language structure takes place at some level below, ranging from limited syntax to some form of semantic representation.

The rules to create source language structure, the transfer rules (lexical and structural), and the generation rules are usually handcrafted. This requires some knowledge of comparative grammar of the language pair, i.e., what are the grammatical differences between the two languages and some knowledge of domain if MT is domain specific.

A number of transfer-based machine translation systems are reviewed by Hutchins and Somers (1992). Some of the concerns expressed about Interlingua approaches are also valid for transfer-based MT: the acquisition of grammar, transfer and generation rules is an endless process.

Over the last decade or two, SMT has gained significant momentum and success, both in academia and industry. SMT has many advantages, e.g. it is data-driven, language independent, does not need linguistic experts, and prototypes of new systems can be built quickly and at a low cost. On the other hand, the need for parallel corpora as training data in SMT is also its main disadvantage, because such corpora are not available for a myriad of languages, especially the so-called less-resourced languages, i.e. languages for which few, if any, natural language processing resources are available. When there is a lack of parallel corpora, transfer-based MT may be used and for many language pairs this approach gives the best results even if there is large training data available. The Workshops on Machine Translation (WMT) include also several transfer-based MT systems in MT system competition and evaluation results show that in English-Spanish, English-German, English-

[2] http://habrahabr.ru/company/abbyy/blog/115226/
(Archived by WebCite® at http://www.webcitation.org/61BAZtLk0)

Hungarian translation directions rule-based systems are outperforming statistical MT systems (Callison-Burch et al., 2009; 2010; 2011).

The Systran MT[3] system is probably the most famous and most used transfer-based MT system. It is a commercial MT system with a history of more than 40 years, and this system has gone through several generations of its technologies (Senellart et al. 2001). With the ability to facilitate communication in 52 language combinations and in 20 vertical domains, Systran's software is the choice of leading global corporations, portals including Apple, Yahoo! and AltaVista, and public agencies such as the US Intelligence Community and the European Commission. As any other transfer-based MT system Systran has faced the challenge of adaptation to new language pairs, to new domains and to specific customer needs. System adaptation is resource consuming and a lot of research and improvements has been done to lower the cost of system adaptation (Farghaly and Senellart 2003, Senellart et al. 2003a, 2003b, Senellart and Senellart 2005, Surcin 2007, Dugast 2009a). Systran's latest achievement, Systran Hybrid MT[4], combines the predictability and language consistency of rule-based machine translation with the fluency of statistical MT.

Another well-known platform for transfer-based MT today is the Apertium platform (Forcada et al. 2009). The Apertium shallow-transfer MT platform was originally aimed at the Romance languages, but has also been adapted for other languages, e.g. Welsh (Tyers and Donnelly 2009) and Scandinavian languages (Nordfalk 2009, Brandt 2011) and many more. There are MT systems for more than 40 language pairs systems build using the Apertium platform. The whole platform, both programs and data, is free and open source and all the software and data for the supported language pairs is available for download from the project website[5].

The Apertium platform consists of the following main modules: (i) a morphological analyser, (ii) an HMM-based (Cutting 1992) statistical part-of-speech tagger, (iii) a lexical selection, (iv) a lexical transfer based on a bilingual dictionary, (v) a structural transfer which performs local morphological and syntactic changes and (vi) a morphological generator.

[3] http://www.systran.co.uk/

[4] SYSTRAN Hybrid Technology. http://www.systran.co.uk/systran/corporate-profile/translation-technology/systran-hybrid-technology . Accessed: 2011-08-25. (Archived by WebCite® at http://www.webcitation.org/61CMZIw7K)

[5] http://www.apertium.org/

There are many more transfer-based MT systems each with its unique architecture, but all performing mainly rule-based (i) analysis of the source text, (ii) transfer from source structures to the target structures and (iii) generation of target text. Just to mention some more typical transfer-based MT systems: Lucy[6] (Gieselmann 2008), MetaMorpho[7] (Novák et al. 2008), ProMT[8,9], Personal Translator[10] (Aleksić and Thurmair, 2011), ETAP-3 (Boguslavsky, 1999), Tilde Translator (Skadiņš et al., 2008; Deksne et al., 2008).

In the area of rule-based MT systems, there has been some progress in recent years, ranging from better control of the translation flow to modifications in the knowledge acquisition phase and in the engine itself. New means of controlling the translation flow are described, for example, in Attnäs et al. (2005) using XML-based configuration files that can be edited by users to decide which processing steps can be taken. Approaches towards using corpus-based technology for bilingual term extraction, and importing such terms into the dictionary of a rule-based system have been researched in the EuroMatrix project (Eisele et al., 2008).

Changes in the MT engine's process of data-driven term selection in the transfer component show that disambiguation of transfer alternatives can be significantly improved using the corpus-based approach data-driven techniques (Thurmair, 2006). Research has also been conducted in intelligent post-processing of the output of rule-based systems using statistical models (Dugast et al. 2007). Each of these steps improves MT quality. The General tendency in a field of transfer-based MT is that almost all systems are integrating some corpus-based elements to improve translation quality and to adapt systems to new domains.

2.1.3 Advantages and Disadvantages

Rule-based MT systems are highly dependent on handcrafted rules which formally capture the language behavior.

Main advantages of the rule-based MT are:

---

[6] http://www.lucysoftware.com/

[7] http://www.morphologic.hu/en/Machine-Translation.html

[8] http://www.promt.com/

[9] PROMT. A Brief Guide to PROMT Machine Translation Technology. PROMT company website. 2011-08-25. URL: http://www.promt.com/company/technology/pdf/e_guide_promt_mt_technology.pdf. Accessed: 2011-08-25. (Archived by WebCite® at http://www.webcitation.org/61CTfW0O9)

[10] http://www.linguatec.net/products/tr/pt

- Provides good quality translation if given all necessary rules and dictionaries;
- Provides good quality translation in narrow domains or working with controlled languages;
- Output translation is consistent and predictable;
- Provide compliance with domain or corporate terminology included in dictionaries;
- Easy to trace and fix the cause of the translation mistakes;
- More precise on capturing difficult language phenomena such as inflections, word agreements, long distance dependency, long distance reordering etc.
- High performance, don't require expensive hardware;
- Allows translation also to functional languages or machine readable formats such as UNL, OWL etc.
- No large parallel corpus needed, since there is no training phase to build statistical model.

Main disadvantages or weaknesses of the rule-based MT are:

- Translation quality depends on the handcrafted rules; needs high linguistics and domain knowledge and is labor-intensive;
- Extensive human labor is required for analyzing syntactic structures of language pair and writing transformation rules;
- Insufficient amount of really good dictionaries. Building new dictionaries is expensive;
- The handcrafted rules are language and domain specific and mostly not adaptable to any other language (except for closely related languages) and are not easy adaptable even to other domains;
- Analysis and transfer modules are not reusable for other non-similar language pairs;
- Even if we have rule-based MT system for the language pair, it is difficult to adapt it to different domain with different terminology or language style;
- Hard to deal with rule interactions in big systems;
- Rule-based MT is a big challenge for low resource languages due to the need of bilingual dictionaries and morphological and syntactic tools;
- Since dictionaries and rules, which try to represent the language behavior, describe only part of real language, some language phenomena might not be covered;
- Rule-based MT systems typically struggle with word sense disambiguation;

- Hard to deal with idiomatic expressions.

2.2 Corpus-Based MT

In the corpus-based (or data-driven) MT large text corpus is used to develop the approximated generalized models of linguistic phenomena based on the actual examples of these phenomena given in text corpus. The corpus-based MT allows using the same MT system for translating any pair of languages if enough training data is available. The further classification of the corpus-based MT is made between (i) the **example-based MT**, where the basic idea is to do translation by analogy, and (ii) the **statistical MT**. In statistical MT, Bayes' rule and statistical decision theory are used to estimate the best translation from source language to target language. As statistical models are learned from corpus, adding more data into the system improves it.

Data-driven methods attempt to overcome the main problem of the traditional rule-based approach: the need for a large human effort of rule writing and linguistic analysis is replaced by the automatic acquisition of translation knowledge from a parallel corpus.

2.2.1 Example-Based MT

Researchers developing rule-based MT systems may consult a corpus of translated text as source of inspiration or validation and build their systems on the basis of corpus analysis and their intuition. But in example-based MT (EBMT) the machine learns to translate directly from a parallel corpus. In its simplest form of EBMT, a given input sentence is compared to a collection of sentences in parallel corpus and the closest match is used to construct the output translation.

There are different EBMT methods and they differ in their matching criteria for *closest match*, the length of input text that is being matched, the generalization of the stored translation examples, the degree of linguistic knowledge that is used for matching and for generalization, etc. Somers (1999) presents a good overview of EBMT methods. Example-based MT approach has its start in Nagao's work (Nagao, 1984) and, it is essentially translation by analogy. The basic principle is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. An EBMT system relies on past translations to derive the translation for a given input text and performs the translation in three steps: (i) matching, (ii) alignment and (iii) recombination (Somers, 2003).

The two main approaches to EBMT are distinguished by the presence or absence of a training stage. EBMT systems that do not have a training stage are often called “pure” or “runtime” EBMT, e.g. (Lepage and Denoual, 2005). These EBMT systems have the advantage that they do not need time-consuming training stages, their runtime complexity can be significant. EBMT systems that incorporate a training stage are normally called “compiled”, as training consists of compiling units below the sentence level. Cicekli and Güvenir (2001) proposed an EBMT approach generalized over sequences of words. The underlying assumption is that translation templates can be learned based on the similarities in both the source and target sides. The same applies to the differing parts between two parallel sentences. Generalization in EBMT consists of replacing the similar or differing sequences with variables and producing a set of translation templates. Generalized templates are source and target language pairs in which certain parts have been replaced by variables. EBMT systems that make use of generalized templates are among the best-performing systems.

Other research was also carried out to learn translation templates based on syntactic generalization, e.g. by Kaji et al. (1992). A recent work has also done on morphological generalization (Phillips et al. 2007). EBMT is also linked with related techniques – translation memory (TM) and computer-aided translation (CAT). A TM stores source and target language translation pairs or translation units (TUs) for effective reuse of the previous translations. TM is often used to store examples so that they can later be used for EBMT systems and CAT tools. EBMT systems find the example (or a set of examples) from the TM which most closely matches the source-language and after retrieving a set of examples with their translations combine them to produce a grammatical translation (Somers 2003). CAT systems segment the input text and compare each segment against the TUs in the TM and produce one or more translations for the source segment and human translators select and recombine them (perhaps with modification) to produce a correct translation (Bowker, 2002). Both EBMT and CAT systems are developed based on a similar principle but in an EBMT, selection and recombination are done fully automatically without the help of a human.

Example-based MT seems to be a solution for under-resourced languages. Example-based MT is based on parallel corpus, which in contrast to statistical MT does not necessary have to include a large number of examples (Somers, 2003). Experiments by Dandapat et

al. (2011) also show that EBMT approaches work better compared to the SMT-based system for certain sentences when the amount of available resource is limited and integration of a sub-sentential TM with the EBMT framework improves translation quality.

However in general EBMT systems have often performed worse than statistical MT systems, e.g., Groves and Way (2005). The biggest problem with EBMT systems is that they do not combine translations of phrases well. This problem is known as boundary friction (Way 2001). Unfortunately it is particularly problematic when translating into a morphologically rich language such as Latvian which is in focus of this research.

2.2.2 Statistical MT

Statistical MT (SMT) may be viewed as example-based machine translation with probabilities. EBMT performs segment aligning and recombination using various heuristic methods (including statistical) while in SMT translation is done using statistical decision theory and everything is determined by statistical models built in the system.

However, historically it can be better understood as the continuation of methods that were highly successful in speech recognition: the decomposition of the problem into a generative statistical model.

Such a model is typically decomposed into a word to word (or phrase to phrase) translation model, a reordering model, and a language model. These models are trained to best explain the empirical data (training corpus). The statistical model is also used as scoring mechanism for possible translations.

SMT is a probabilistic framework for translating text from one language to another, based on models induced automatically from a parallel corpus. A statistical MT system is composed of three parts: a translation model, which captures the correspondence between words and phrases in different languages; a language model, which reflects the fluency of the target language; and a decoder, which incorporates the translation and language models to perform the actual translation.

Statistical MT came in the early 1990's when large bilingual corpora were readily available. Researchers at IBM proposed a statistical approach to machine translation based on a probabilistic dictionary (Brown et al. 1990). Shortly thereafter, methods for automatically aligning sentence pairs in a bilingual corpus were developed by Gale and Church (1991) and

Brown et al. (1991). Taking advantage of the available parallel corpora, Brown et al. (1993) described five statistical models for machine translation, called the IBM Models. These models form the basis for word-based statistical machine translation. These models show how a sentence could generate its translation, word by word, using Shannon's (1948) noisy channel model of communication and with parameters estimated from a parallel corpus.

Figure 2.2.1 shows how this probabilistic translation process can be used to model the machine translation problem.

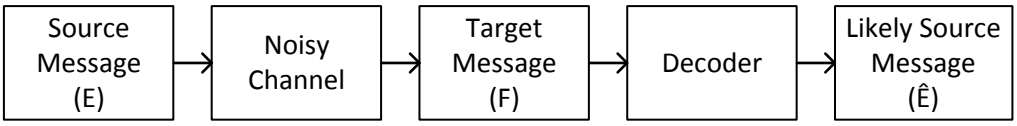


Figure 2.2.1 Translating with Shannon's noisy-channel model of communication

Brown et al. (1993) explained their perspective on the statistical translation process:

A string of English words, e, can be translated into a string of French words in many different ways. Often, knowing the broader context in which e occurs may serve to winnow the field of acceptable French translations, but even so, many acceptable translations will remain; the choice among them is largely a matter of taste. In statistical translation, we take the view that every French string, f, is a possible translation of e. We assign to every pair of strings (e, f) a number P(f|e), which we interpret as the probability that a translator, when presented with e will produce f as his translation. We further take the view that when a native speaker of French produces a string of French words, he has actually conceived of a string of English words, which he translated mentally. Given a French string f, the job of our translation system is to find the string e that the native speaker had in mind when he produced f. We minimize our chance of error by choosing that English string ê for which P(e|f) is greatest.

Brown et al. uses *English* and *French* to explain the stochastic translation process, but this may sometimes introduce misunderstanding. When they speak about translation, they always speak about translation into English. Equations get confusing if we start using them speaking about translation from English to French or other language. Therefore we will introduce more precise terms – *source language* and *target language*. So, to translate

source language sentence S to target language sentence we need to find target language sentence Ť which has the maximum probability P(T|S). See Eq. (1)

$$\check{T} = \underset{T}{\operatorname{argmax}} P(T|S) \tag{1}$$

Treating a source language sentence S as an encoding of a target language sentence means that the probability of T being the intended translation of S can be expressed using Bayes' rule:

$$P(T|S) = \frac{P(S|T) \cdot P(T)}{P(S)} \tag{2}$$

Because the denominator is independent of the translation hypothesis T, finding the most likely translation Ť can be reduced to maximizing the numerator. The resulting "Fundamental Equation of Machine Translation" is:

$$\check{T} = \underset{T}{\operatorname{argmax}}\ P(S|T) \cdot P(T) \tag{3}$$

The term P(T) represents the language model probability, and P(S|T) is the translation model probability. This transformation also allows for the use of a language model P(T), which can be trained independently. As in speech recognition, the a priori probability of T can be thought of as the likelihood that T is a fluent sentence in the target language. The language model probability is high for well-formed target language sentences, independent of their relationship to the source language sentence. The translation model probability is the probability that the sentence S can be generated from T. The translation model probability is large for target language sentences, regardless of their grammaticality, that have the necessary words in roughly the right places to explain the source language sentence. Equation (3) therefore assigns a high probability to well-formed target language sentences that account well for the source language sentence.

2.2.3 Word-based SMT

In word-based SMT the translation process is further decomposed into smaller steps that are modeled with probabilities that are conditioned on single words. (See Figure 2.2.2)

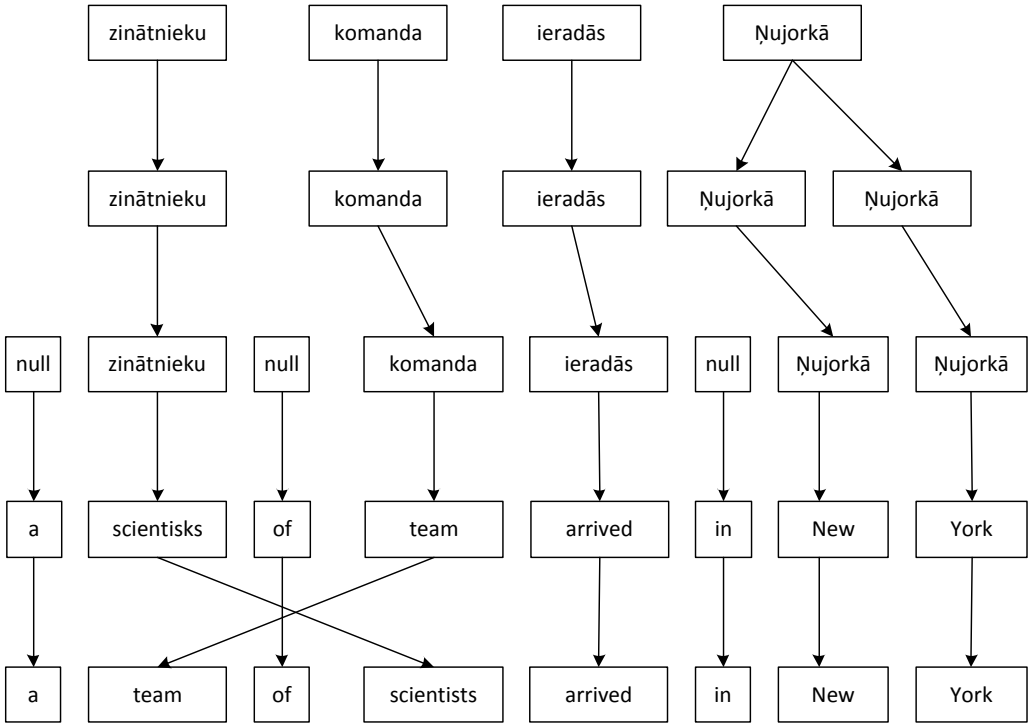


Figure 2.2.2 The translation process using IBM Model 4

Many concepts for SMT such as the expectation maximization (EM) approach for training from parallel corpora and the noisy channel approach for decoding were introduced in the Candide project at IBM by Brown et al. (1990).

Figure 2.1.1 gives an illustration of the translation process using IBM Model 4. The probability of the English sentence given the Latvian sentence is the product of a number of probabilities that model (i) word duplication (including multiplication and deletion), (ii) word insertion, (iii) word translation, and (iv) word reordering. Each of the arrows in the example denotes a probability that is used in the resulting product. The resulting product is the probability of sentence translation. This decomposition is mathematically motivated by marginalizing the joint probability distribution and a number of independence assumptions. Strong independence assumptions limit the conditioning to only the directly affected words, hence enabling sufficient statistical basis for the estimation of the probability distribution from the data.

The different models proposed by the IBM group differ only in the conditioning of the probability distributions. For instance IBM Model 4 uses relative movement (with respect to the previous word), while IBM Models 1-3 use absolute word reordering. Later work replaces

these probability distributions with maximum entropy classifiers that allow taking local context into account.

The decoding problem of finding a target language sentence for a given source language sentence is NP-complete for the IBM Models (Knight, 1999). Thus, to solve this problem search heuristics such as beam search (Och, 1998; Al-Onaizan et al., 1999; Och et al., 2001), or greedy hill-climbing (Germann et al., 2001) are needed.

Word-based MT sees translation as the task as the mapping of words from one language into another, with some reordering. However, often words have to be inserted and deleted without clear lexical evidence on the other side, and words do not always map one-to-one. As a consequence, the word-based models proposed by Brawn et al. (1993) were encumbered with additional complexities – word fertilities and NULL word generation. Word-based models have been all but abandoned over the last decade, and replaced by phrase-based models which view the task as mapping translation of small text chunks from one language into another, again with some reordering.

2.2.4 Phrase-based SMT

The term "phrase" denotes a multi-word segment, and not a syntactic unit such as a noun phrase. Using phrase alignments instead of word alignments to calculate the translation probability allows the inclusion of local context information in the translation model. Figure 2.2.3 shows an example of a phrase-aligned English to Latvian translation, including a one-to-many alignment ("*komanda"* to "*a team"*, and "*zinātnieku"* to "*of scientists"*) and a many-to-many alignment ("*ieradās Ņujorkā"* to "*arrived in New York"*), which are not allowed by word-aligned translation models. This leads to better word choice in translation and more accurate word reordering.

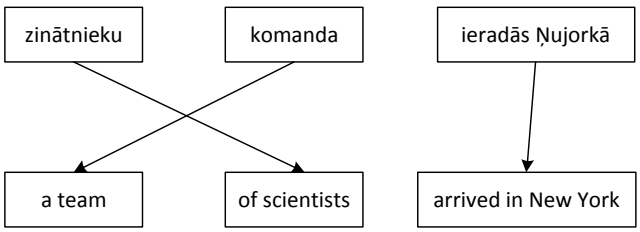


Figure 2.2.3 An example of phrase-aligned translation

Phrase-level alignments have been an active research topic in statistical machine translation. One approach has been to use phrases, corresponding to syntactic subtrees in a parsed sentence, as part of a syntax-based translation model (Yamada and Knight 2001). Phrase-based translation models by Och and Ney (2003) use the automatically-generated word-level alignments from the IBM models to extract phrase-pair alignments. The phrase-pair alignment probabilities are then used as the fundamental units of the translation model instead of the word alignment probabilities. Other alternative methods to learn phrase translation tables have also been proposed (Tillmann, 2003; Venugopal et al., 2003; He, 2007; Fraser and Marcu, 2007). Koehn et al. (2003) show that phrase-based translation systems outperform the syntax-based translation model of Yamada and Knight (2001).

The noisy-channel approach to model phrase-based translation (as in the word-based models) can be used, but phrase-based translation models tend to use log-linear models to model the phrase translation probabilities (Koehn et al., 2003).

2.2.5 Log-liner models in SMT

A log-linear model expresses the probability of a translation P(T|S) as a combination of several features that characterize some aspect of the translation of sentence S into T. The Fundamental Equation of Machine Translation (Eq. (3)) includes only two features – P(T|S) and P(T), but we can think about other features too. For example, the length of translated sentence could be a useful feature. Equation (3) also considers that both P(T|S) and P(T) has an equal weight or impact on translation result, but it might not be so if we have more features. If we have N features, then the log-linear translation model is expressed by equations (4) and (5):

$$P(T|S) = \frac{1}{Z}\prod_{i=1}^{N} \alpha_i^{\,x_i(T|S)} \tag{4}$$

Where Z is a normalizing constant, $x_i(T|S)$ is a feature that characterize some aspect of the translation of sentence S into T, and α_i is the weight assigned to feature $x_i(T|S)$

$$\check{\mathrm{T}} = \underset{T}{\operatorname{argmax}} \prod_{i=1}^{N} \alpha_i^{\,x_i(T|S)} \tag{5}$$

We see that the Fundamental Equation of Machine Translation (Eq. (3)) is a special case of equation (5), if N=2; $x_1(T|S) = \log_{\alpha_1} P(S|T)$ and $x_2(T|S) = \log_{\alpha_2} P(T)$.

An advantage of the log-linear model can be seen when taking log of equation (5).

$$\check{T} = \underset{T}{\operatorname{argmax}} \log \prod_{i=1}^{N} {\alpha_i}^{x_i(T|S)} = \underset{T}{\operatorname{argmax}} \sum_{i=1}^{n} \log {\alpha_i}^{x_i(T|S)} \tag{6}$$

$$= \underset{T}{\operatorname{argmax}} \sum_{i=1}^{n} x_i(T|S) \log \alpha_i = \underset{T}{\operatorname{argmax}} \sum_{i=1}^{n} \lambda_i \, x_i(T|S)$$

Where λ_i now is the weight assigned to feature $x_i(T|S)$.

The phrase-based log-linear model described by Koehn et al. (2003) includes several other features to select the most likely translation. For example, reordering probability (called distortion), and feature calibrating the length of output sentence, as the models otherwise have a tendency to produce shorter sentences.

The values for feature weight are estimated with Minimum Error-Rate Training (MERT) (Och, 2003). MERT optimizes the weights for the features to maximize the overall system's translation score. The most common metric used for MERT is BLEU (Papineni et al., 2002).

Use of log-liner models is not limited to phrase-based models, log-liner models are widely used in syntax-based MT too.

2.2.6 Factored SMT

The phrase-based models (Koehn et al., 2003; Och and Ney, 2004; Vogel et al., 2003; Tillmann, 2003), are limited to the mapping of small text chunks (phrases) without any explicit use of linguistic information, may it be morphological, syntactic, or semantic. Such additional information has been demonstrated to be valuable by integrating it in pre-processing or post-processing. However, a tighter integration of linguistic information into the translation model is desirable for two reasons:

1. Translation models that operate on more general representations, such as lemmas instead of surface forms of lexical units, can draw on richer statistics and overcome the data sparseness problems caused by limited training data.
2. Many aspects of translation can be best explained on a morphological, syntactic, or semantic level. Having such information available to the translation model allows the

direct modeling of these aspects. For instance: reordering at the sentence level is mostly driven by general syntactic principles, local agreement constraints show up in morphology, etc.

Factored SMT framework (Koehn and Hoang, 2007) is an extension of the phrase-based approach. It adds additional annotation at the lexical unit level. A lexical unit in our framework is not anymore only a token, but a vector of factors that represent different levels of annotation. The training data (a parallel corpus) has to be annotated with additional factors. For instance, if it is necessary to add part-of-speech information on the input and output side then part-of-speech tagged training data are required. Typically this involves running automatic tools on the corpus, since manually annotated corpora are rare and expensive to produce.

The Moses SMT system (Koehn et al, 2007) features factored translation models that allow integrating additional layers of data tightly into the process of translation.

2.2.7 Hierarchical and Syntax-based SMT

One of the fundamental properties of natural language is its recursive structure, which forms the basis of syntactic models that define language as process of recursive rule applications. It is hence intuitive that statistical machine translation models should also be based on such recursive rules, and in fact this notion was adopted fairly early on by the field (Wu, 1997; Alshawi et al., 1998; Yamada and Knight, 2001).

Such syntax-based translation models may be decorated with linguistically motivated syntactic annotation at the source (Huang et al., 2006; Liu and Gildea, 2009) or the target (Galley et al., 2006; Shen et al., 2008a), on both sides (Zhang et al., 2008), or neither (Chiang, 2005). One fundamental difference to the earlier, simpler word-based and phrase-based model is the use of a different decoding algorithm that builds a tree structure bottom-up, instead of left-to-right. This significantly increases computational complexity, a problem that has not yet been fully resolved.

Hierarchical SMT is statistical MT method which is based on classical Statistical MT techniques, but it uses hierarchical phrases. The method was introduced by Chiang (2005; 2007) who is the founder of this MT approach. Chiang's approach is purely statistical, he uses hierarchical phrases, but they are just hierarchical without any relation to hierarchical phrase structures used in computational linguistics.

Zollmann and Venugopal (2006) describes different kind of statistical MT approach (syntax-augmented SMT) which in general is based on the same idea as Chiang's system, but it uses linguistically motivated hierarchical phrases and other techniques in building of phrase table and in decoding.

Hierarchical SMT

Chiang (2007) describes popular approaches and describes limitations of current statistical MT systems and describes typical issues which cannot be solved with state-of-art MT systems. Main issue mentioned is long distance reordering. It is obvious that languages use phrases and that these phrases are hierarchical, Chiang proposes to introduce very general hierarchies in phrase-based translation model of statistical MT.

Chiang introduces Hiero MT system which is hierarchical phrase-based statistical MT system. System uses synchronous context free grammar (SCFG) to represent hierarchical phrases learned from parallel corpus. SCFG is context free grammar but it describes two languages simultaneously in each rule. Typical rule in SCFG used in Hiero system is:

X → <yu X1 you X2, have X2 with X1>

This rule says that Chinese phrase "yu X1 you X2" can be translated to English as phrase "have X2 with X1" where X1 and X2 can be any sequence of words. Only one non-terminal symbol X is used in grammar used by Hiero system. Grammar contains also rules with simple non-hierarchical phrases without non-terminals in right-hand side of the rule.

Grammar has also two so called glue rules:

S → <S1 X2, S1 X2>

S → <X1, X1>

These rules are used to glue together separate phrases to form sentences. Glue rules introduce one more non-terminal symbol – S.

The Hiero System learns rules from parallel corpus. It uses parallel corpus which is not annotated and is not parsed. System uses GIZA++ (Och and Ney, 2003) to align words in both directions then it takes union of both results. Then it extracts phrases just like any other phrase-based statistical MT system does, then Hiero system creates all possible rules from each phrase. In such way system can get enormous amount of SCFG rules which cannot be

stored and processed by the system, therefore only rules which comply with certain constraints are extracted. There are 6 constraints mentioned in the paper. The most significant restrictions are: only two non-terminals in a rule and no adjacent non-terminals in source language part. The second restriction is very important to avoid spurious ambiguity when equal source phrases are described with many different rules.

Rules extracted from parallel corpus are not linguistically motivated in any way, they have just one non-terminal X (plus S in glue rules) and they just represent hierarchy in phrases. In opposite to syntax based statistical MT systems (Zollmann and Venugopal 2006, Yamada and Knight 2001) in this approach rules are learned from unparsed parallel corpus. So there is no possibility to get any linguistically motivated phrases.

In general system uses model which is kind of classical log-linear phrase-based MT systems (Och and Ney, 2002). Language model is classical as in majority of statistical MT systems but translation model is different. Author gives deep and mathematically precise description of translation model. The decoder is source language parser which uses source language part of SCFG rules to parse source sentences. It would be reasonable to use probabilistic CYK (PCYK) parser, but source language rules are not in CNF form; right-hand side of rules can contain terminals along with one or two non-terminals. Parsing algorithm which is similar to PCYK, but can work with source language rules as they are, is used. Complexity of the algorithm is the same as for PCYK – $O(n^3)$.

Comparison of Hiero to Alignment Template System (Och and Ney, 2004), hereafter ATS, shows that Hiero is better. BLEU score for Hiero was 34.57 while BLEU score for ATS was 31.74. BLEU score for Hiero running with non-hierarchical phrases was 28.83. This shows that Hiero system benefits much from hierarchical phrases.

Author makes several conclusions: (i) the hierarchical phrase-based MT system performs better than state-of-art non-hierarchical MT system, (ii) the system might benefit from syntactically motivated phrases.

Syntax-based SMT

Zollmann and Venugopal (2006) are describing different hierarchical phrase-based MT system. System also uses SCFG in translation phrase model. But this system is trained on parallel corpus with parsed target side. As the result system uses SCFG with the same non-

terminal symbols as used in parsed corpus. In opposite to Hiero system (Chiang, 2007) this system uses hierarchical phrases which are linguistically motivated. Decoder is in many ways similar to decoder described by Koehn and others (Koehn et al. 2003) but it is based on chart parser which is parsing target side of SCFG rules.

Yamada and Knight (2001) present a syntax-based statistical translation model. The model transforms a source-language parse tree into a target-language string by applying stochastic operations at each node. These operations capture linguistic differences such as word order and case marking. To incorporate structural aspects of the language, the model accepts a parse tree as an input. The channel performs operations on each node of the parse tree. The operations are (i) reordering child nodes, (ii) inserting extra words at each node, (iii) and translating leaf words.

Comparison

Koehn et al. (2003) show that phrase-based translation systems outperform the early syntax-based translation model of Yamada and Knight (2001).

Hierachical and syntax-augmented approaches in SMT are relatively new and there is research going on to investigate do they really improve translation quality. Although evaluation of approaches and comparison to phrase-based systems was done also by Chiang (2007) and by Zollmann and Venugopal (2006), these evaluations were not faire in some aspects. Systems were compared with different parameters to overcome boundaries put by computational complexity. More recent and more precise large scale system comparison was done by Zollmann et al. (2008). Zollmann et al. discovered that PSCFG based approaches can yield substantial benefits for language pairs that are sufficiently non-monotonic, that is there is much reordering (Chines-English, Urdu-English). But PSCFG based approaches cannot yield substantial benefits for language pairs that are rather monotonic, that is there is no much reordering (Arabic-English). These results indicate that a phrase-based system with sufficiently powerful reordering features and LM might be able to narrow the gap to a hierarchical system or even outperform them for rather monotonic languages.

The gap (or non-gap) between phrase-based and PSCFG performance for a given language pair seems to be consistent across small and large data scenarios, and for weak and strong language models.

Hoang et al. (2009) show that despite many differences between phrase-based, hierarchical, and syntax-based translation models, their training and testing pipelines are strikingly similar. The Moses toolkit (Koehn et al., 2007) has been extended to implement also hierarchical and syntactic models, making it the first open source toolkit with end-to-end support for all three of these popular models in a single package. This extension substantially lowers the barrier to entry for machine translation research across multiple models. Hoang et al. (2009) also show that the hierarchical and syntactic models in Moses achieve similar quality to the phrase-based model, even though their implementation is less mature, for some languages (English- German) phrase-based models still outperform hierarchical and syntax-based models.

2.2.8 Language Modeling

As it was shown above all approaches in SMT use language model in one or other way. Language modeling is a fundamental base technology in SMT and it is usually trained and used independently form other models used in translation.

A statistical language model is a probabilistic way to capture regularities of a particular language, in the form of word-order constraints. A statistical language model expresses the likelihood that a sequence of words is a fluent sequence in a particular language. Reasonable sequences of words are given high probabilities whereas senseless ones are given low probabilities.

The dominant approach to language modeling is to use a simple n-gram language model, which probabilistically predicts the next word in a sequence based on the preceding few words. The most widespread statistical language model, the n-gram model, was proposed by Bahl et al. (1983) and has proved to be simple and robust. The n-gram language model has dominated the field since its introduction despite ignoring any essential linguistic properties of the language being modeled. Language is reduced to a sequence of symbols with no deep structure or meaning, but this simplification works.

There is also strong interest in using linguistic tools, such as parts-of-speech taggers, in SMT. Factored SMT framework (Koehn and Hoang, 2007) adds additional annotation at the lexical unit level. A lexical unit in our framework is not anymore only a token, but a vector of factors that represent different levels of annotation including part-of-speech or other information. Although factored SMT allows using of language models not only on the surface level but

also on other factors, language modeling over several factors is still done on each factor independently. An integrated factored language modeling is introduced by Axelrod (2006) in his M.Sc. Thesis.

From a language modeling point of view, rich morphology also poses a number of challenges. It increases the number of surface forms in the vocabulary, causing sparse data problems in model. N-gram monolingual language models have reduced context and consecutive-word-phrases have less coverage due to word order variation. Also, rich morphology typically increases the use of long-distance agreement constraints, which break the assumptions of simplistic but (for English) effective models such as n-gram language models, which only rely on neighboring words for context. The relatively free word order of languages with rich morphology creates a long-distance dependency problem for the traditional n-gram language modeling. This problem has been addressed by structured language modeling. Zhang (2009) provides an overview of various approaches to structured language modeling in the past and proposes a new framework. (see also Bilmes, 2003). Various ways how to use syntax and parsing in language modeling are also discussed by Charniak (2001), Charniak et al. (2003), Kirchhoff et al. (2006), Sarikaya and Deng (2007).

2.2.9 Advantages and Disadvantages

The main advantages of the corpus-based MT are:

- No handcrafted dictionaries and rules needed, since systems automatically learn from large corpora;
- No need for experts in language pair and domain;
- Development of corpus-based MT system is much cheaper compared to development of rule-based MT
- Translation quality can be easily improved by adding more data to the training cycle;
- Easy to apply the approach on different language pairs and domains if given adequate language resources for training ;
- Able to capture phrasal and idiomatic expressions occurring in the training data;
- Many publicly available and widely used statistical machine translation tools;
- SMT systems can be easily adapted for other language pairs or domains;
- Word senses can be easily disambiguated based on n-gram models;
- Produce more fluent translations because of huge language models;

- Highly dependent on statistical model which models linguistic aspect although with limited linguistics knowledge.

The main disadvantages or weaknesses of the corpus-based MT are:

- Large and good quality parallel corpus is required for building translation system for any language pair. For many language pairs such corpus is not available;
- The lack of reasonable size parallel corpus will produce a bad translation since the model it produces is inadequate to capture the language behavior;
- The domain of the training data also contributes to the translation quality. MT system trained on data from a different domain can produce bad translation result;
- Data sparseness is a common problem. Since training data, which tries to represent the language behavior, is just small part of language sample, some language phenomena might not be covered;
- There is still need of better systems for dealing with highly inflectional and structurally complex languages;
- Extensive hardware requirements for building and managing translation models on large data;
- Strict use of domain specific terminology is still a problem for corpus-based MT

2.3 MT for Latvian

The first MT system for Latvian was developed in the beginning of 60-ies at the Institute of Electronics (Гобземис et al., 1961). The system was a typical word-to-word (or direct) MT system to translate scientific texts from Russian into Latvian. Good morphological analysis tools were developed as part of the system, since the system provided translation between two highly inflected languages. Another direct MT system for aviation documentation was proposed at the Riga Aviation Institute in 1995 (Ореховский and Мишнев, 1995).

The rule-based approach to machine translation has been dominant in Latvia since mid-90-ies when the first version of the LATRA system has been developed at the Institute of Mathematics and Computer Science (IMCS) of the University of Latvia (Greitāne, 1997). Research on rule-based systems continued at IMCS until 2004 by elaborating LATRA with semantic properties and by adapting it to new domains.

In 1996 IMCS joined development of an Interlingua MT system is based on the Universal Networking Language (UNL) elaborated in the United Nations University by H. Uchida (Uchida et al., 1999; Uchida and Zhu, 2001). Through the project basic converting rules that allow translation from UNL into Latvian are developed (Skadiņa, 2004).

Tilde also has worked on a rule-based approach to develop a commercial system for users who have poor or no foreign language skills. The MT system Tildes Tulkotājs (Skadiņš et al., 2008) has been released in 2007 as part of Tildes Birojs 2008. The system translates texts from English into Latvian and from Latvian into Russian.

The recent advances in Latvian MT including the statistical MT and human language technologies in general are described by Skadiņa et al. (2010a). Research on Statistical Machine Translation (SMT) was started by IMCS with a LCS funded project "Evaluation of SMT Methods for English-Latvian Translation System" (2005-2008) through which the baseline English-Latvian system was created (Skadiņa and Brālītis, 2008). The system's performance in BLEU points was similar to other systems for inflected languages of that time. IMCS research on SMT continues with the project "Application of Factored Methods in English-Latvian SMT System" (Skadiņa and Brālītis, 2009), the latest version of the system is available on the Web at http://eksperimenti.ailab.lv/smt.

In 2009 Tilde started development of an English-Latvian SMT system. Besides publicly available resources, internal resources collected over time have been used for SMT training. Latvian-English SMT system has been developed as well. Both systems are publicly available at http://translate.tilde.com (Skadiņš et al., 2010).

In 2010 English-to-Latvian engines of Microsoft Translator was developed in close cooperation between Microsoft Research and Tilde (Microsoft Research, 2010). Tilde provided guidance in a number of technologies that touched machine translation, data, and Latvian specific tools and technologies, facilitating significant gains in quality for the Latvian language translations in Microsoft Translator.

Two SMT related EU projects: the ICT PSP program project LetsMT! (www.letsmt.eu) and the FP7 project ACCURAT (www.accurat-project.eu), both coordinated by Tilde, have been started in 2010. The LetsMT! project aims to build an innovative online collaborative platform for data sharing and MT generation. This platform will support the uploading of SMT training data and building of multiple customized MT systems (Vasiļjevs et al., 2011;

2010; Skadiņš et al., 2011a). The ACCURAT project researches novel methods that exploit comparable corpora to compensate for the shortage of linguistic resources to improve MT quality for under-resourced languages and narrow domains (Skadiņa et al., 2010b).

3 RELATED WORK: HYBRID MT

Recent MT evaluation campaigns (Callison-Burch et al. 2009; 2010, 2011) show that both rule-based and statistical MT systems reach comparable level of translation quality, but the level of output understandability even for the best language pairs is about 50%. This means, that the state-of-the-art methods in MT are far from being accepted by human readers, which limits the MT usage significantly. Error analysis (Chen et al., 2007; Thurmair, 2005) shows that the errors made by rule-based and statistical MT systems are complementary:

- Rule-based MT systems have weaknesses in lexical selection, and lack robustness when sentence analysis fails. However rule-based MT systems translate more accurately by trying to translate every piece of the input sentence.
- SMT systems are more robust and always produce output. The output of SMT systems is more fluent, due to the use of language models, and SMT systems are better in lexical selection. However, SMT systems have difficulties to deal with language phenomena which require linguistic knowledge, like word order, syntactic functions, and morphology. SMT systems more often lose adequacy because of missing or spurious translations (Vilar et al., 2006).

Systems which try to profit from the respective other approach (and avoid mistakes for which solutions already exist) are hybrid solutions, combining knowledge or rule-based and data-driven or statistical elements. The purpose of this chapter is to discuss recently proposed different architectures of hybrid MT systems.

Detailed overview of hybrid MT architectures is given by Thurmair (2009). According to him there are three main types of hybrid MT architectures: (i) coupling of systems (serial or parallel), (ii) architecture adaptations (integrating novel components into SMT or RBMT architectures), and (iii) genuine hybrid systems, combining components of different paradigms.

Coupling of MT systems means that two or more existing MT systems are combined to produce better MT output. Coupling can either be done in a serial or parallel way. Serial system coupling is called – Statistical Post-Editing (SPE), it uses statistical MT to post-edit output of a rule-based system. Coupling also can be done in parallel; in this case the best

translation is selected or produced from the output of two or more MT systems. Parallel MT system coupling are also known as MT System Combination or Multi-engine MT.

While system coupling means that the architecture of the involved MT systems is not changed, by architecture adaptation we mean that the system architecture basically follows the rule-based MT or statistical MT approach but is modified to include features (components, resources etc.) of the other approach. Modifications can be done as (i) pre-editing (the system data are pre-processed), or (ii) modification of the core functionality (e.g. extended phrase tables, enlarged dictionaries etc. using techniques of the other MT approach).

3.1 Statistical Post-Editing

Hybrid systems which are built using serial MT system coupling nearly exclusively modify a rule-based MT output by means of a SMT post-editing (See Figure 3.1.1). The SMT component is trained on a parallel training data consisting of output from the rule-based MT and "good" output (See Figure 3.1.2).

Figure 3.1.1 Serial MT system combination or Statistical Post-Editing.

Combinations of rule-based MT and SPE systems are highly competitive in MT quality (Schwenk et al., 2009). The output of SPE tends to be grammatical, and the main effect of the combination is an increase in lexical selection quality (Dugast et al., 2007) which is one of the weak points of pure rule-based MT systems. However, attention must be taken to avoid the introduction of new errors by the statistical post-processor. The statistical post-editor may introduce errors in the syntactic structure of the output (Ehara, 2007); accuracy may drop as some parts of the translation are omitted, and special attention needs to be taken to handle terminology and named entities well in the output (Dugast et al., 2009b). To avoid such quality degradation, Federmann et al. (2009) use a syntactic structure of rule-based MT system, and try only local alternatives. Statistical post-editing helps to improve the lexical

selection problem of the rule-based MT systems, but does not really deals with the parse-failure problem.

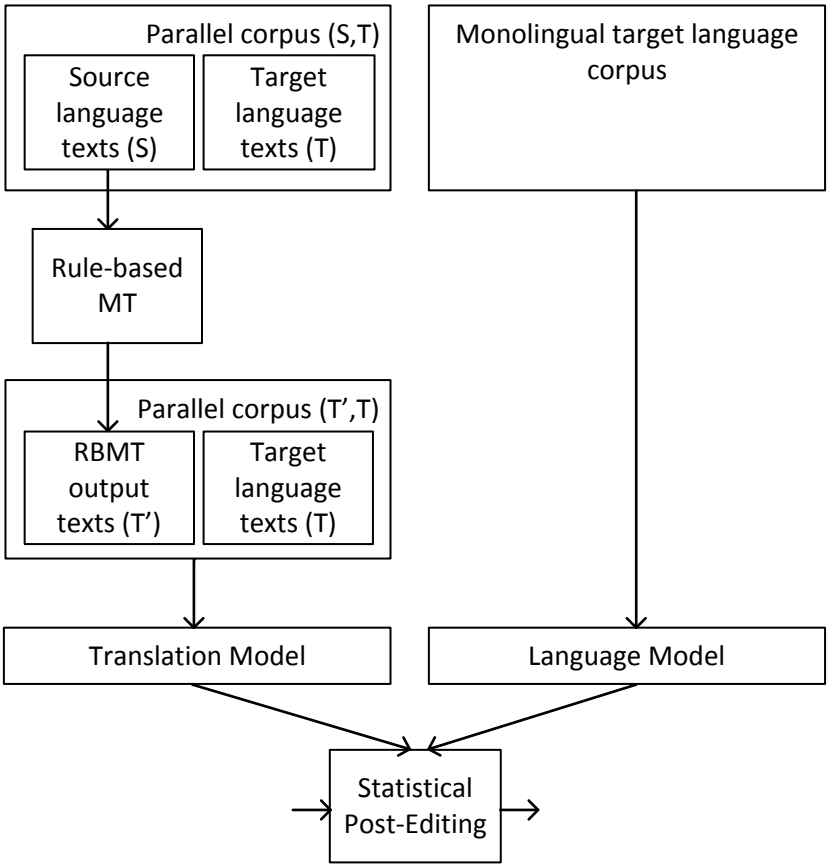


Figure 3.1.2 Training of Statistical Post-Editing system.

3.2 Multi-engine MT

This coupling employs several MT systems in parallel, and uses some mechanism to select or produce the best output from the result set.

Research in machine translation has led to many different translation systems, each with strengths and weaknesses. None of the different approaches to MT, whether statistical, example-based, rule-based or hybrid, does not provide the best results. System combination exploits these differences to obtain improved output. This is why some researchers have investigated the multi-engine MT (MEMT) systems (Eisele, 2005; Macherey and Och, 2007; Du et al., 2010) aimed to provide translations of higher quality than those produced by the isolated MT systems in which they are based on.

MT system combination has taken a great importance these past few years mainly due to the fact that single systems achieved good quality and the possibility of taking the most of their complementarity in a system combination framework is very attractive. The last Workshops on Machine Translation (WMT) included a system combination task; an overview is given in (Callison-Burch et al., 2009; 2010; 2011).

The effectiveness of system combination strongly depends on the relative performance of the systems being combined. In the 2009 WMT, Callison-Burch et al. (2009) conclude that "In general, system combinations performed as well as the best individual systems, but not statistically significantly better than them." A possible reason for this failure to improve on individual systems is given in 2010 WMT: "This year we excluded Google translations from the systems used in system combination. In last year's evaluation, the large margin between Google and many of the other systems meant that it was hard to improve on when combining systems. This year, the system combinations perform better than their component systems more often than last year." (Callison-Burch et al., 2010)

Many approaches to system combination exist. MEMT systems can be classified according to how they work.

- Systems that combine the translations provided by several MT systems into one consensus translation (Bangalore et al., 2002; Matusov et al., 2006; Heafield et al., 2009; Du et al., 2010); the output of MEMT system may differ from outputs provided by the individual MT systems it is based on.
- Systems that decide which translation is the most appropriate one among all the translations computed by the MT systems they are based on (Nomoto, 2004; Zwarts and Dras, 2008) and output this translation without changing it.
- In-between, there are MEMT systems that build a consensus translation from a reduced set of translations. Systems that (i) first chose the subset with the most promising translations, and then (ii) combine these translations to produce a single output (Macherey and Och, 2007).

Many different techniques are used for system combination. Some systems concern hypothesis selection using n-best list re-ranking based on various features (Hildebrand and Vogel, 2009). Some systems consider source text and systems outputs as parallel corpus and train a new SMT system on this corpus (Chen et al., 2009).

The system combination based on Confusion Network (CN) is probably the most popular method for system combination. There are numerous publications available on that subject, for example, by Rosti et al. (2007), Shen et al. (2008b) or Karakos et al. (2008). Such an approach is presented in Figure 3.2.1.

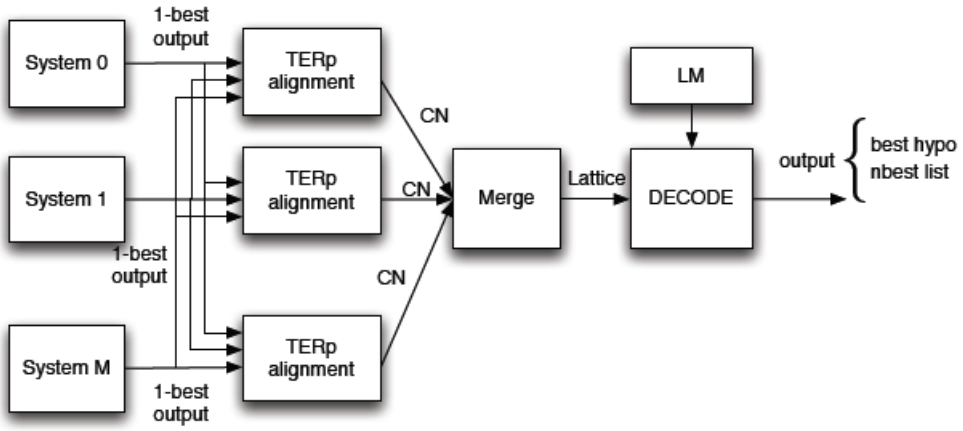


Figure 3.2.1 MT system combination using Confusion Networks (Barrault, 2010).

The system combination can be decomposed into three steps:

- 1-best hypotheses from all M systems are aligned in order to build confusion networks;
- All confusion networks are connected into a single lattice;
- A language model is used to decode the resulting lattice and the best hypothesis is generated.

Barrault (2010) presents machine translation system combination software, MANY, based on Confusion Networks described above.

There are Confusion Networks based system combination techniques which use confusion network for lexical choice and techniques which jointly resolves both word order and lexical choice. The Carnegie Mellon multi-engine machine translation system (Heafield and Lavie, 2010) is closely related to work by He and Toutanova (2009) who uses a reordering model like Moses (Koehn et al., 2007) to determine word order. This system shows significant improvement on some translation tasks, particularly those with systems close in performance. In NIST MT09 evaluation, the combined Arabic-English output scored 5.22 BLEU points higher than the best individual system.

Sennrich (2011) presents system combination architecture based on the phrase-based SMT framework by adding a second, dynamic phrase table similar to that described in (Chen et

al., 2007). While majority of approaches for system combination treat all systems as black boxes, needing only the 1-best output from each system and a language model, the combined system described by Chen et al.(2007) and Sennrich (2011) is an extension of an existing SMT system. The combination is achieved by taking a baseline SMT system and adding a second, dynamic phrase table to the existing primary one. Chen et al. (2007) propose that the dynamic phrase table is trained on the translation output of several rule-based translation systems. But later Chen et al. (2009) expand this concept by allowing for the inclusion of arbitrary translation systems. Sennrich (2011) shows that combined system can outperform system combination algorithms that only use information on the target side, i.e. the translation hypotheses and a language model.

One of the disadvantages of MEMT systems is a need to translate the input sentence as many times as different MT systems they use. This makes it difficult to use MEMT systems in applications where response time and required resources are constrained. To overcome or reduce this limitation Sánchez-Martínez (2011) presents an approach aimed to select the subset of MT systems, among a known set of systems, which will produce the most reliable translations for a given sentence by using only information extracted from that sentence.

3.3 Extending rule-based MT with statistical elements

The main approaches to improve rule-based systems with data-driven procedures are:

- Pre-editing. Both on the dictionary building, by running term-extraction tools and extracting dictionary candidates, and on the grammar building, by automatically extracting grammar rules from corpora. Word alignment and extracting grammar rules from corpora are typical techniques used in statistical MT.
- Modification of the system core functionality. Both by adding probability information to the analysis and parsing processes, and by improving the transfer and lexical selection processes.

Pre-Editing is done to prepare the language resources for rule-based MT. Main language resources, dictionaries and grammar rules can be set up and improved using data-driven technology.

MT quality improves moderately, depending on the amount of decrease of the out-of-vocabulary words, which depends on the size and coverage of the existing dictionary.

The approach helps to fill dictionary gaps, and to adapt to new domains. However, the problem of lexical selection exacerbate in rule-based MT systems with large dictionaries, as the amount of translations between which to select increases. This problem turns out to be much more difficult to solve than the problem of dictionary gaps. Current hybrid approaches focus more on translation selection in the transfer phase, which also is one of the weaknesses of rule-based MT systems, especially if dictionaries are big. The traditional approach to rule-based MT transfer selection relies on two techniques:

- Detection of subject domain;
- Detection of certain contextual or structural properties (like: colocations, certain prepositions, passive voice etc.), which leads to a specific translation.

An obvious solution is to use the more frequently used translation as default. But this technique is not sensitive to the specific context, and mostly returns the default. A second option is to use contextual disambiguation in the lexical selection process. As a result, core modifications in rule-based MT can significantly improve the transfer and lexical selection process; however they are less successful in case of robustness and parse failures.

3.4 Extending statistical MT with knowledge or rule-based elements

Like rule-based MT systems, SMT systems have also been extended to improve translation quality. Again,

- Pre-editing is used to prepare the data; the most commonly used steps are morphological analysis, POS information, syntactic information, and word reordering.
- System core modifications are used, by adding information coming from a rule-based MT to the phrase tables, and by using factored translation.

Morphology: Morphology has been researched rather extensively, especially in morphologically rich languages. Lemmatization and POS tagging was used both on the source side (de Gispert et al., 2006) and on the target side (Vandeghinste et al., 2006); the aim is to reduce data sparseness using lemma-based language models instead of surface form based models. It improves results for smaller corpora, but impact is not so significant when corpora grow. Also, it seems that both surface form and lemma based analysis should be done, as surface information has also shown to be beneficial (Koehn and Hoang, 2007).

Factored translation (see Chapter 2.2.6 “Factored SMT”) is able to work on both levels simultaneously.

Another pre-editing area in morphology is compounding (English) and de-compounding (German), to improve alignment (Stymne et al. 2008, Popović et al. 2006). In agglutinative languages, like Turkish, Estonian, Hungarian or Arabic, preprocessing is required to split complex words (including pronouns, case markers etc.) into parts to be able to align them.

Syntax: Syntactic pre-processing is also used e.g. by Hannemann et al. (2009); the idea is to parse source and target side of a corpus, and only syntactically well-formed phrases are allowed in the phrase table. Both corpora are parsed; matching sub-trees are identified and aligned in the phrase table or phrase table is filtered to remove un-matching phrases.

Importing rule-based MT resources into the phrase table: It was proposed e.g. by Eisele et al. (2008) to run rule-based MT systems before SMT systems, and enrich the SMT phrase tables by terms and phrases produced by rule-based MT systems. This approach makes use of the knowledge included in the dictionaries and rules of the rule-based MT systems. Sánchez-Cartagena et al. (2011) proposes different approach, they use lexicon and shallow-transfer rules of the existing rule-based MT system to enrich phrase tables of the SMT system. Results show that the coverage of the MT system can be increased, especially translating texts from different domains; however, as the SMT decoder runs last the output can be less grammatical than the one of the original rule-based MT system. Such hybrid architecture deals with the data sparseness issue of the SMT training but it does not help to solve word agreement and other the output issues related to the grammar.

Hierachical & Syntax-based SMT: Hierarchical and syntax-based statistical MT (see chapter 2.2.7 “Hierarchical and Syntax-based SMT”) is getting more and more popular. Although these SMT methods are corpus-based MT methods, they can be extended with additional knowledge. Syntax-based SMT systems can use rule-based parsers in training process. Hierarchical and syntax-based SMT systems work with probabilistic synchronous context-free grammars which they extract from syntactically annotated or un-annotated parallel corpus (Chiang, 2007; Zollmann and Venugopal 2006). Li et al. (2011) shows how human written rules can be integrated in hierarchical SMT system and such rules give significant quality improvement; human written rules can be integrated in syntax-based SMT as well.

Factored Translation: While using structural information for decoding attracts increasing interest, Factored Translation (Koehn and Hoang, 2007) aims at enriching phrase-based SMT systems 'bottom up', by providing additional information at the word level (see chapter 2.2.6 "Factored SMT"). It treats words not just as simple surface forms but as vectors of features, such features as the lemma, the POS, morphology, and others. Several papers (e.g. Stymne et al., 2008) show that phenomena like NP-agreement and compounding can be handled efficiently within a factored translation framework. Rule-based components such as morphological analyzers, POS taggers, syntactic analyzers and even semantic analysis can be used in training of factored phrase-based SMT systems. These rule-based components typically are used in pre-processing phase to annotate training data with additional knowledge.

3.5 Genuine hybrid architectures

Hybrid architectures described above are combinations or extensions of existing rule-based and corpus-based MT architectures, but there are also genuine hybrid architectures that do not just use extensions to their system architecture but combine whole system components of the rule-based and corpus-based approaches into novel system architectures. They use three basic components: (i) identification of source language "chunks" or fragments (words, phrases), (ii) transformation of such chunks into the target language using a bilingual resource, (iii) and generation of a target language sentence. At the moment active research is going on in the field of genuine hybrid MT (for example several EC co-financed research project: METIS[11], METIS II[12], PRESEMT[13]), but the current state-of-the-art technologies are still under-developed and they have not reached the same level of maturity as rule-based, statistical and example based MT.

One of the approaches used relies on rule-based analysis, bilingual dictionary, and target language model. Such an approach has been investigated in the METIS projects (Vandeghinste et al., 2006). Analysis of the input sentence is done using available natural language processing tools (such as lemmatisers, POS taggers, chunkers); transfer is done using existing dictionaries (consisting basically of lemma and POS in source and target

[11] http://www.ilsp.gr/metis/

[12] http://www.ilsp.gr/metis2/

[13] http://www.presemt.eu/

language; dictionaries also include multiword units); and generation uses a target language language model (tokenized and tagged English corpus). Evaluation shows that results are similar to basic SMT systems but worse than a well-established rule-based system like SYSTRAN (Vandeghinste et al., 2008).

Another hybrid architecture has been proposed by Carbonell et al. (2006) - an alternative data-driven approach instead of rule-based analysis. The only necessary resources are: (i) a bilingual dictionary (including all word forms), and (ii) an n-gram indexed target language corpus. In analysis phase, an n-gram window is moved over the sentence, and all words are translated using the dictionary; based on these translations, the target language corpus is searched for the closest n-gram (ideally containing all words of the source window, and no additional ones). The result is a lattice of n-gram translations. Segments with the strongest left and right overlaps, and the highest density of terms, are selected by the decoder from the lattice.

3.6 Domain adaptation

A special issue to be considered speaking about hybrid MT is domain adaptation. All kinds of MT systems must handle the fact that they can be used not only in the domain for which they had been developed but also for other domains. It is also important to have easy and relatively cheap ways how to adapt existing system to the new domain. Rule-based MT systems support adaptation by tailoring dictionary and rules; and this tailoring can be done using in-domain corpus, the situation is less obvious for statistical MT systems, and a significant drop of quality (up to 10 BLEU points) had been observed.

If we focus on building a domain specific SMT engine, pooling together all available data, especially a significant portion of data that is out of the desired domain, can lead to reductions in quality, since the out-of-domain training data will overwhelm the in-domain (Koehn and Schroeder, 2007). Unfortunately, the drawback of domain specific SMT, that is, where only in-domain data is used, is its failure to capture generalizations relevant to the target language, which can lead to poor translation quality again (Thurmair, 2004). A domain specific MT engine needs to capture the generalizations of an engine trained on a large and sufficient supply of parallel data, yet not lose the crucial domain orientation. It has been shown that to achieve this, SMT engine can be trained on all available parallel data including out-of-domain data, but language model training data must be split into in-domain and out-

of-domain sets, generating separate language models for each (Koehn and Schroeder, 2007; Lewis et al., 2010). Also the domain specific development corpus is necessary to tune model weights that favor the domain-specific language model over the out-of-domain one.

Experiments have also been performed for integrating domain specific terminology (a bilingual list of terms) into an SMT system (Itagaki and Aikawa, 2008). Artificial contexts are created to identify how the phrase tables will translate a source language term and then in a rule-based post-processing phase, the translations of the terms are replaced by the target expressions from the term list.

4 METHODS FOR COMBINING DIFFERENT MT APPROACHES AND EXPERIMENTS

This chapter:

1. Introduces methods which were selected and researched in this research to find the optimal way for combining best techniques used for rule-based and corpus-based MT;
2. Gives description of quality evaluation techniques used to evaluate these methods;
3. Describes experiments performed to test various combining methods and results.

4.1 Methods

The goal of this research was to find effective means which would allow to build better MT systems overcoming limitations of the two main existing MT approaches by combining them using the strength of the one to overcome the weakness of the other. A special attention has been paid to the applicability of the methods to English-Latvian MT and to the possibility to use methods to adapt MT to particular domain.

A lot of research is going on in this field during the last years. Hybrid MT has recently emerged. As it is shown in Chapter 3 “Related Work: Hybrid MT” there are several ways how to build hybrid MT systems, and several of approaches has been researched by the author in this research.

Typically hybrid technologies develop starting from something what is well established by adding something new to bring the existing technology to the new level. If there is a good rule-based system, then its developers search for the ways how to improve it with elements from the corpus-based MT; and vice versa if there is a good corpus-based MT, then its developers look for possibilities to improve it using some rules or knowledge. This research also started with an existing English-Latvian rule-based MT system (Deksne et al., 2005). This was the first version of the first widely accessible English-Latvian MT system with the starting level of quality. The error analysis showed that the main weaknesses of the system are lexical selection and poor lexical coverage which both are typical issues for rule-based MT systems (Chen et al., 2007; Thurmair, 2005). Two attempts were made to improve the system quality using corpus-based MT technologies.

The use of corpus-based techniques to improve the lexical selection process of the rule-based MT system has been researched. The core functionality of Tilde's rule-based MT system (Deksne et al. 2005) was extended to incorporate a statistical disambiguation component to make the context aware lexical selection process (Skadiņš et al., 2008; Skadiņa et al., 2007). The statistical disambiguation component improved the quality of the MT system. A detailed description of the experiment and results are given in Chapter 4.3 "Rule-based MT with Statistical Lexical Disambiguat".

Another research was done to improve the lexical coverage using corpus-based techniques. Data extracted from the parallel corpus were used to extend a multi-word expression dictionary (Deksne et al., 2008) of the Tilde's rule-based MT system (Skadiņš et al, 2008). The extended multi-word dictionary also improved the quality of the rule-based MT. A detailed description of the experiment and results are given in Chapter 4.4 "Rule-based MT with a Dictionary Obtained from the Corpus".

Although both methods improved the quality of the MT system, they did not bring it to the new quality level. The rule-based MT system was still with the starting level of quality and without obvious, effective and relatively cheap way how to improve the quality using corpus-based techniques. The research done by Dugast et al. (2009a), Surcin et al. (2007) and Aleksić and Thurmair (2011) shows different results, which can be explained with the fact that they worked with well-established rule-based MT systems (Systran, PersonalTranslator) with rather high translation quality. From this we can see that the quality level of the rule-based MT system is very important, if we want to build a hybrid MT using the existing rule-based MT system. There is no easy way how to improve a rather low quality rule-based MT system using corpus-based techniques.

A rather high quality rule-based MT system is also a prerequisite to build a hybrid MT solution using the statistical post-editing or the multi-engine MT technique. If we have such a system, then we can improve it with a statistical post-editing component or combine with other MT systems. The development of the good rule-based MT system is a very time and resource consuming process; therefore such systems are not available for many language pairs, including English-Latvian which is in the focus of this research. Although there is English-Latvian rule-based MT system (Skadiņš et al., 2008) its quality is not high enough to be used in a hybrid MT.

Building of rather high quality statistical MT system is much more easy task than building of the same quality rule-based MT. To build the statistical MT system we need only a reasonable amount of training data (Koehn et al., 2009); and we can improve the quality of statistical MT just by adding more training data (Och, 2005). The amount of training data is always limited, therefore we can research methods how to use additional knowledge about the language pair and the domain to improve the quality of statistical MT system.

The use of knowledge about Latvian morphology to improve a factored phrase-based SMT system has been researched in this research. Several ways how to include morphology information in English-Latvian SMT system have been researched (Skadiņš et al., 2011a; 2011c; 2010; Vasiļjevs et al., 2011a; Šics, 2010). The created English-Latvian SMT system with Latvian morphology outperforms both a baseline SMT system without morphology and any other existing English–Latvian MT system (Skadiņš et al., 2011a; 2011c) in automatic evaluation achieving 35 BLEU points. The system produces more fluent output and helps to achieve the 32.9 % productivity increase in practical localization tasks (Skadiņš et al., 2011a; Vasiļjevs et al., 2011a). A detailed description of the experiment and results are given in Chapter 4.5 “Statistical MT with a rule-based morphological analyzer”.

Syntactic information also can be used as a factor in factored phrase-based SMT. The factored SMT with the syntactic information has been researched also for English-Latvian MT (Šics, 2010); unfortunately the syntactic information does not improve the MT quality as much as the morphological information.

The factored translation with different types of additional knowledge as factors has also been researched in this research. An experiment was performed using a spatial knowledge coming from a spatial ontology. The spatial ontology was used to disambiguate toponyms (Skadiņš, 2010; 2011; Skadiņš et al. 2011b). Results of this experiment show a slight improvement in the MT quality. The proposed method allows using not only the spatial knowledge, but also other type on the knowledge stored in domain ontologies. A detailed description of the experiment and results are given in Chapter 4.6 “Statistical MT with Knowledge”.

As it was mentioned before – more training data leads to a better quality of the statistical MT. Using of different methods to find parallel and comparable data in the web has been researched. Results show that it is possible to extract parallel sentences from comparable

data found in the web; and although the quality of such automatically extracted data is not very high, using of such data helps to improve the MT quality (Skadiņš et al., 2011a; 2011c; Skadiņa et al. 2010b).

4.2 Evaluation

There are several ways how the MT system quality is evaluated. Most popular are automatic evaluation metrics. These metrics are objective, they can be quickly calculated when needed, and it does not cost much. But automatic metrics just give a number which says how similar the MT system output is to the reference translation. Unfortunately the automatic MT quality evaluation does not always correlate with a human judgement (Callison-Burch et al., 2006). But the human evaluation of the MT quality is not as cheap as the automatic evaluation, because we need many human evaluators to compare many sentences produced by different MT systems to say which one is better. The human evaluation is expensive and it cannot be performed too often. There are also different ways how to perform the human evaluation. The simples and cheapest is a system comparison (Callison-Burch et al., 2009); it helps to distinguish which system is better, but it does not say anything about types of errors the MT system makes. Vilar et al. (2006) proposes more advanced (and more human labor demanding) method for the MT quality evaluation which gives also a detailed description of error types. Other way how to look on the MT quality is a usefulness of the system for the particular task. We can evaluate how well the MT system performs the task it is designed for. For example, we can evaluate whether it is helping to the professional translator to do translation faster. The following 3 sections describe evaluation methods (automatic, human and evaluation in localization) used to evaluate quality of methods researched in this research. Two human evaluation methods are used only to evaluate the best MT method proposed as these methods are too expensive to use in every experiment.

4.2.1 Automatic evaluation

For the automatic evaluation the two most popular and widely used metrics BLEU[14] (Papineni et al., 2002) and NIST (Doddington, 2002) were used. Automatic metrics are cost-effective and do not require much human intervention. They allow comparisons of two and

[14] BiLingual Evaluation Understudy

more systems, as well as different versions of one system in the process of its implementation and improvement as many times as necessary.

A balanced reference set of 500 English sentences was developed for the automatic evaluation purposes. The compiled corpus consists of original, natural, parallel sentences; and it is sentence-aligned, not annotated (morphologically, syntactically, and lexically unmarked), and is representational and balanced at the same time. English-Latvian parallel sentences were manually collected from the web and validated by a professional translator (a reference set to be compared with). English-Lithuanian reference corpus was manually translated by a professional translator. The breakdown of topics in the corpus is presented in Table 4.2.1.

Table 4.2.1 The breakdown of topics in the evaluation set.

Domain	Percentage
General information about the EU	12%
Specification and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
News and magazine articles	24%
Information technology	18%
Letters	5%
Fiction	5%

The procedure of the automatic evaluation consists of several sub-processes and the main idea, in general, is in the comparison of machine translation and reference sets. The higher the automatic scores are, the better the machine translation output quality is.

4.2.2 Human evaluation

A ranking of translated sentences relative to each other for manual evaluation of systems is used for human evaluation in this research. This was the official determinant of translation quality used in the 2009 Workshop on Statistical Machine Translation shared tasks (Callison-Burch et al., 2009). The same test corpus is typically used as in automatic evaluation.

In this research only two systems (ties were allowed) are compared. It was discovered that it is more convenient for evaluators to evaluate only two systems and results of such evaluations are easier to interpret as well. A web based evaluation environment was developed (Skadiņš et al., 2010) where we can upload sources sentences and outputs of two MT systems as simple txt files. Once evaluation of two systems is set up we can send a link of evaluation survey to evaluators. Evaluators are evaluating systems sentence by sentence.

Evaluators see source sentence and output of two MT systems. The order of MT system outputs in evaluation differs; sometimes evaluator gets the output of the first system in a first position, sometimes he gets the output of the second system in a first position. Evaluators are encouraged to evaluate at least 25 sentences, we allow evaluator to perform evaluation is small portions. Evaluator can open the evaluation survey and evaluate few sentences and go away and come back later to continue. Each evaluator never gets the same sentence to evaluate. We are calculating how often users prefer each system based on all answers and based on comparison of sentences.

When we calculate evaluation results based on all answers we just count how many times users chose one system to be better than other. In a result we get percentage showing how in many percent of answers users preferred one system over the other. To be sure about the statistical relevance of results we also calculate confidence interval of the results. If we have *A* users preferring the first system and *B* users preferring the second system, then we calculate percentage using Eq. (7) and confidence interval using Eq. (8).

$$p = \frac{A}{A+B}\ 100\% \tag{7}$$

$$ci = z\sqrt{\frac{p(1-p)}{A+B}}\ 100\% \tag{8}$$

where z for a 95% confidence interval is 1.96.

When we have calculated *p* and *ci*, then we can say that users prefer the first system over the second in *p±ci* percent of individual evaluations. We say that evaluation results are **weakly sufficient** to say that with a 95% confidence the first system is better than the second if Eq. (9) is true.

$$\mathrm{p} - \mathrm{ci} > 50\% \tag{9}$$

Such evaluation results are weakly sufficient because they are based on all evaluations but they do not represent system output variation from sentence to sentence. We can perform system evaluation using just one test sentence and get such weakly sufficient evaluation results. It is obvious that such evaluation is not reliable. To get more reliable results we have to base evaluation on sentences instead of all answers. We can calculate how evaluators have evaluated systems on a sentence level; if we have *A* evaluators preferring the particular

sentence from the first system and *B* evaluators preferring sentence from the second system, then we can calculate percentage using Eq. (7) and confidence interval using Eq. (8). We say that particular sentence is translated better by the first system than by other system if Eq. (3) is true. To get more reliable evaluation results we are not asking evaluators to evaluate sentences which have sufficient confidence that they are translated better by one system than by other. When we have *A* sentences evaluated to be better translated by the first system and *B* sentences evaluated to be better translated by the second system or systems are in tie, then we can calculate evaluation results on sentence level using Eqs. (7) and (8) again. And we can say that evaluation results are **strongly sufficient** to say that the first system is better than the second in the sentence level if Eq. (9) is true. We can say that evaluation is just **sufficient** if we ignore ties.

4.2.3 Evaluation in a localization scenario

Evaluation in a localization scenario (Skadiņš et al., 2011a; Vasiļjevs et al., 2011a) was based on the measurement of translation performance using SDL Trados Studio 2009 computer-aided translation tools with LetsMT! plug-in (Vasiļjevs et al., 2010; 2011b) Performance was calculated as the number of words translated per hour. The evaluation was made in the software localization domain.

For the evaluation two test scenarios were employed: (1) a baseline scenario with translation memory (TM) only and (2) an MT scenario with a combination of TM with MT. The baseline scenario established the productivity baseline of the current translation process using SDL Trados Studio 2009 when texts are translated unit-by-unit (sentence-by-sentence). The MT scenario measured the impact of using MT in the translation process when translators are provided not only matches from the translation memory (as in baseline scenario), but also MT suggestions for every translation unit that does not have 100% match in translation memory. Suggestions coming from the MT were clearly marked (see Figure 4.2.1).

In both scenarios translators were allowed to use whatever external resources needed (dictionaries, online reference tools etc.), just as during regular operations.

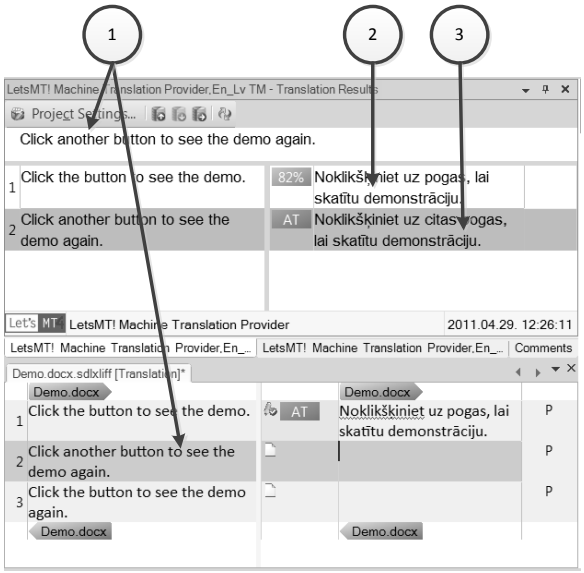


Figure 4.2.1 Translation suggestions in SDL Trados Studio 2009; 1 – a source text, 2 – a suggestion from the TM, 3 – a suggestion from the MT.

Five (5) translators with different levels of experience and average performance were involved in the evaluation.

The quality of each translation was evaluated by a professional editor based on the standard quality assurance process of the service provider. The editor was not made aware whether the text was translated using the baseline scenario or the MT scenario. An error score was calculated for every translation task. The error score is a metric calculated by counting errors identified by the editor and applying a weighted multiplier based on the severity of the error type. The error score is calculated per 1000 words and it is calculated using Eq (10).

$$ErrorScore = \frac{1000}{n} \sum_{i} w_i e_i \tag{10}$$

where

- n is a number of words in a translated text,
- e_i is a number of errors of type i,
- w_i is a coefficient (weight) indicating severity of type i errors

There are 15 different error types grouped in 4 error classes – accuracy, language quality, style and terminology. Different error types influence the error score differently because errors have a different weight depending on the severity of error type. For example, errors of type *comprehensibility* (an error that obstructs the user from understanding the information; very clumsy expressions) have weight 3, while errors of type *omissions/unnecessary additions* have weight 2. Depending on the error score the

translation is assigned a translation quality grade: Superior, Good, Mediocre, Poor and Very poor (Table 4.2.2).

The test set for the evaluation was created by selecting documents in the IT domain from the tasks that have not been translated by the translators in the organization before the SMT engine was built. This ensures that translation memories do not contain all the segments of texts used for testing.

Table 4.2.2 Quality evaluation based on the score of weighted errors

Error Score	Quality Grade
0…9	Superior
10…29	Good
30…49	Mediocre
50…69	Poor
>70	Very poor

To evaluate a usefulness of the MT system for the particular localization task the translator performance (as the number of words translated per hour) and error score was calculated. As this evaluation method requires much human work, it is used only to evaluate the best created MT system described in Chapter 4.5 “Statistical MT with a rule-based morphological analyzer”.

4.3 Rule-based MT with Statistical Lexical Disambiguator

This chapter gives an overview of research and experiments which has been done to improve English-Latvian-English and Latvian-Russian rule-based MT systems developed by Tilde (Deksne et al. 2005).

4.3.1 Motivation

As any other rule-based MT system Tilde’s MT system has to deal with an issue of lexical disambiguation. It is generally hard to resolve ambiguities in such systems, since there is no natural way to assign scores or probabilities to the dictionary entries and various rules.

4.3.2 MT system

Experiments with statistical disambiguation of lexical ambiguities where performed using existing Tilde’s multilingual transfer-based MT system which was extended with additional disambiguation module. The MT system is built from separate components, each of them having their own functionality (see Figure 4.3.1). Components are executed successively during the translation process. The system detects the language of source text, analyzes the

text, performs multi-word expression (MWE) processing, performs syntactic and lexical transfer, and establishes morphological agreement between words. Finally, the result is presented to the user.

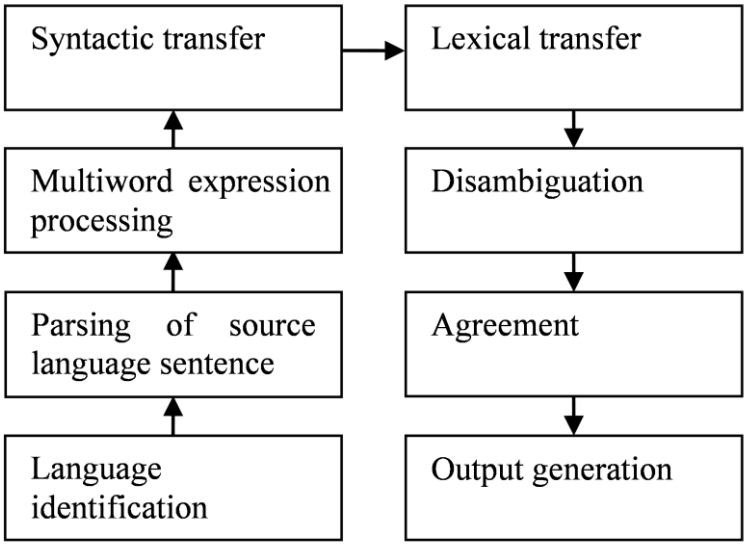


Figure 4.3.1 The chain of Tildes transfer-based MT system components

Language identification

Language identification module is developed to relieve the user from the need to select the source language every time the language of the source text changes. This module automatically identifies the language of the text and provides the appropriate information to the system. The MT system identifies English and Russian as source languages, but it is built on platform of comprehension assistant (Skadiņa et al. 2007) which can identify also Estonian, French, German, Latvian, and Lithuanian.

For language identification, the character n-gram approach is used (Grefenstette, 1995; Bashir Ahmed et al, 2004). The *language reference model* is based on the most frequent character n-grams of sizes 1, 2, 3 and 4.

Parser

The aim of the parser component is to obtain a fully or partially parsed sentence. As the parsers differ from language to language, a wrapper component is developed, which transforms the output of different parsers to a unified format necessary for further processing. English and Russian parsers are licensed from third party software vendors

Connexor[15] and Dictum[16]. The output of the parser component is a syntax tree, or a part of the syntax tree of the sentence in case when full sentence parsing fails.

Multiword expression processing

There are many cases in real texts when the meaning of collocation is not based on the meaning of its parts. Latvian is not an exception and is rich in idiomatic expressions. To improve the quality of rule-based MT system, MWEs should get a special treatment. There are different kinds of MWEs – phrasal verbs (e.g. “give up”, “have a lunch”), nominal compounds (e.g. “telephone box”), institutionalized phrases (e.g. “salt and pepper”) or phrases with truly idiomatic meaning (e.g. “early bird gets the worm”). Syntactic structure of translated phrase can be completely different from source phrase.

The MT system has a specially compiled dictionary of phrases and a set of MWE rules. Every entry in this dictionary is mapped to MWE rule ID (see Table 4.3.1).

Table 4.3.1 An excerpt from the dictionary of phrases

Source phrase	Target phrase	Rule ID
sound a false note	uzņemt nepareizu toni	V-DET-A-N-14
out of temper	Saniknots	ADV-PREP-N-1
have a swim	Izpeldēties	V-DET-N-1
get a cold	Saaukstēties	V-DET-N-1
have lunch	ēst pusdienas	V-N-3

MWE rule describes how the syntactical parse tree fragment will change in translation. In target tree description we specify not only the type of syntactic relation but also the position of the child node to the parent node (see Figure 4.3.2). The child node could be before the parent ('left'), before all parent's other children ('leftmost'), next to the parent ('right'), after all parent's other children ('rightmost').

```
IdiomRule(V-DET-A-N-14)
V1[comp:N4[?det:DET2,attr:A3]]=>V1[obj(right):N4[attr(left):A3]]

sound(V1) a(DET2) false(A3) note(N4) => uzņemt(V1) nepareizu(A3) toni(N4)
```

Figure 4.3.2 A sample MWE rule

[15] www.connexor.com

[16] DictaScope Syntax. 2011-08-29. http://www.dictum.ru/en/syntax-analysis/blog. (Archived by WebCite® at http://www.webcitation.org/61IQkJ08w)

The MWE processing algorithm is the following: we traverse the parse tree top-down trying to match the fragment of parse tree with a structure defined in MWE rule. If the match is found MWE rule looks up in the dictionary of phrases for a lexical match. If the matching entry is found in the dictionary, target tree fragment is created and lexical translations attached to the right nodes. The translated MWE is integrated into the target tree to use it later in transfer, agreement and other processes.

The simplest case of MWE rule – the structure of parse tree fragment does not change, the fragment with exactly same structure is created, and translations are attached to the corresponding nodes. But most MWE rules describe the changes in syntactical structure. More detailed description of MWEs and MWE processing is given in chapter 4.4.2 “MWE processing in rule-based MT system”.

Syntactic transfer

The next step is the syntactic tranfer which is responsible for the transformation of a source language syntactical tree into a corresponding target language syntactical tree by applying transfer rules. The developed rule formalism allows to change word order, delete or hide nodes, insert new nodes, transfer or assign syntactical, morphological or lexical properties, and change the type of syntactical relations between words.

The Figure 4.3.3 demonstrates a transfer rule that changes word order in a source language noun phrase into word order of target language noun phrase.

```
TransferRule (N<-mod-PREP<-pcomp-N) // team of scientists
{
        Child.SourceSpelling == ”of”;
move_to_left (GrandChild, Parent);
Grandchild.Case = genitive;
MakeLink (Child – hidden -> Parent);
MakeLink (GrandChild - mod -> Parent);
}
```

Figure 4.3.3 English-Latvian transfer rule for syntactic transfer of genitive phrase

When the transfer rule is applied to the English noun phrase ‘*team of scientists*’, the following transformations are carried out: a word *scientists* is placed before a head word *team*, the case of the word *scientists* is set to possessive (genitive), the preposition *of* is discarded.

Lexical transfer

After the syntactic transfer the system performs the lexical transfer. And here we deal with the translation *per se* based on a bilingual dictionary and the grammatical category of the part of speech identified by the parser component. For example, for the English word *rest* in a sentence '*we need a rest'*, noun translations (for Latvian: *atpūta, miers, pauze, pārtraukums*) are selected and verb translations (for Latvian: *palikt, atpūsties, balstīties, gulties*) are dismissed.

There can be no translation in a dictionary though. It can occur due to several reasons. Absence of derivatives in dictionaries should be mentioned, e.g., *assume* (stem word is included in a dictionary as a rule), *assumption*, *assumed* as an adjective, *assuming* as a noun (derivatives are unlikely to be included), *assumer*, *assumingly* (derivatives are likely not to be included). Then absence of some grammatical forms, e.g., non-finite forms of the verb: Russian participle *улыбающийся* is translated into Latvian in several steps *улыбаться – smaidīt – smaidošs*. Clash of opinions and diversity of approaches to the parts of speech in grammar science and language processing systems correspondingly also 'contributes', e.g., the Latvian word *mans* as a pronoun and the Russian word *мой* as an adjective in two morphological analyzers; as with the lack of convergence in grammatical categories of two languages, e.g., Russian qualitative, relative, and possessive adjectives comparing to Latvian qualitative and relative adjectives, therefore, a great number of relative and possessive adjectives correspond to Latvian nouns in genitive: *деревянный – koka, школьный - skolas, папин - tēva.* The system tries a different approach then, the dictionary lookup is attempted for alternate classes: an adjective instead of a participle, a noun instead of an adjective, and others.

Agreement

After disambiguation process, the syntactic tree of target language contains single target language word at each node. Each node has some morphological properties inherited from corresponding node of source language tree or set during parsing and transfer phases. These morphological properties are insufficient to generate fluent output sentence. Therefore agreement rules describe syntactic relations and grammar of the target language.

For example, to establish agreement in case, number and gender between noun and adjective in Latvian noun phrase, the morphological properties of noun is transferred to adjective (see Figure 4.3.4).

```
// Первый снег – pirmais sniegs, Первые шаги – pirmie soļi
// Small house – maza māja
Rule(A-attr->N)
{
        Child.Gender = Parent.Gender;
        Child.Number = Parent.Number;
        Child.Case = Parent.Case;
}
```

Figure 4.3.4 Agreement rule: parent (N) assigns gender, case, number to child (A)

4.3.3 Experiment – statistical lexical disambiguation

As the baseline for this experiment existing Tilde's transfer-based MT system without any disambiguation component was taken. The MT system will always take the first translation coming from the bilingual dictionary if there is no disambiguation component available. Typically there is more than one translation for a word in bilingual dictionary. The task of the disambiguation phase is to choose the most appropriate target language word from the several words selected in the lexical transfer phase. We use statistical methods for disambiguation. Traditionally bilingual corpus is used to get statistical data for disambiguation (Ide et al. 2002, Chan and Ng 2005). Three different translation directions with different resources available were used for the experiment. The experiment was done and Tilde's MT system was extended with statistical disambiguation component in 2007. At that time, availability of bilingual corpus was very limited. Part of the results of the experiments have been reported by Skadiņa et al. (2007), Skadiņš et al. (2008) and Gornostay et al. (2007) in papers introducing Tilde's rule-based MT system.

For **English-Latvian** translation direction we combined two approaches – (i) using a monolingual corpus and (ii) MWEs with their translation equivalents extracted from the bilingual dictionary.

We decided to take into account statistical data about the probability of syntactic pairs - two words being syntactically related in a phrase or sentence. This is a more advanced approach compared to bigram probability - probability of two words appearing next to each other in a sentence. We use several syntactic relations such as subject(noun, verb), object(verb, noun), attribute(adjective, noun) and attribute(noun, noun). We gathered a large corpus (about 3

mil. sentences) of Latvian texts from the web and news. We applied a shallow parser (Skadiņa et al. 2007) on this corpus to get pairs of syntactically related words. The frequency of each unique pair was calculated. Frequency data were normalized to get probability of syntactic pairs. We call the resulting data the syntactic language model (SLM) and use it for disambiguation. In the syntactic tree of the target language we have one or more Latvian language words mapped to every node (source language word). For every connected Latvian word pair in the tree we find probability from the Latvian SLM. Now we can disambiguate the syntactic tree by selecting those translations that give the highest probability for the whole tree representing the phrase or the sentence.

If we are translating sentence “cats chase mice”, then we have translation tree illustrated in Figure 4.3.5 before the disambiguation phase. A set of translations is attached to each node. For example, set of translations T_2 = {ķert, vajāt, padzīt, gravēt, …} is attached to node “chase”. Let’s label T_1 elements as $t_{1,1}$, $t_{1,2}$ etc.; and similarly T_2 elements as $t_{2,1}$, $t_{2,2}$, …, T_3 elements as $t_{3,1}$, $t_{3,2}$ ...

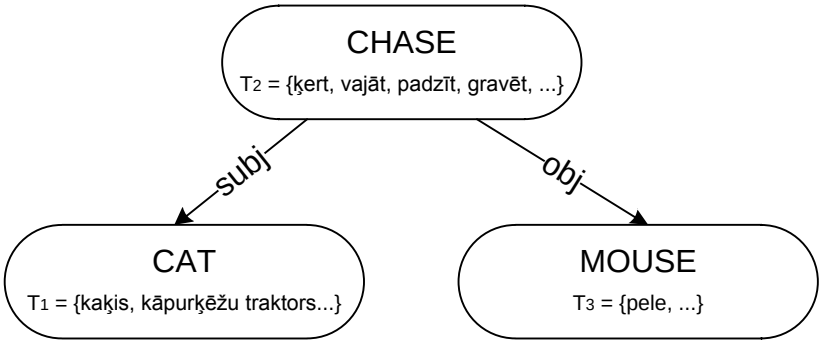


Figure 4.3.5 The translation tree before disambiguation, English-Latvian

To disambiguate this tree we have to find $t_{1,i} \in T_1$, $t_{2,j} \in T_2$ and $t_{3,k} \in T_3$ giving the highest probability for the whole tree. This is expressed by the equation (11).

$$\underset{t_{1,i} \in T_1 \,;\, t_{2,j} \in T_2 ;\; t_{3,k} \in T_3}{\operatorname{argmax}} P(\text{subj}\, t_{1,i}\; t_{2,j}) \times P(\text{obj}\, t_{2,j}\; t_{3,k}) \qquad (11)$$

where P(subj $t_{1,i}$ $t_{2,j}$) is a probability that $t_{1,i}$ and $t_{2,j}$ is syntactically related with relation *subj*, and so on.

The SLM based disambiguation improves the quality of the translation compared to the most primitive method of using just the first translation from the dictionary. But the drawback of

this method is usage of target language data only and ignoring the source language text in disambiguation.

Use of MWEs in lexical disambiguation is not a statistical process; it is completely rule-based based process. A list of phrases with the same syntactic structure was extracted from bilingual dictionary and added in MWE dictionary (see chapter "Multiword expression processing" on page 61). These MWEs do not change sentence structure in translation process, but they help to translate typical collocations much better compared to taking just the first translation for each word from the dictionary.

For the evaluation we used English-Latvian test corpus described in chapter 3. The aim of automatic evaluation task was to evaluate influence of disambiguation module on translation quality. Two popular evaluation metrics NIST (Doddington 2002) and BLEU (Papineni et al. 2002) were chosen for automatic evaluation. Evaluation results are summarized in Table 4.3.2. These results show that, although there is a quality improvement using the statistical disambiguator (8.13 vs 8.11 BLEU points), this improvement is not significant.

Table 4.3.2 Evaluation results for English-Latvian MT system with lexical disambiguation

System	BLEU	NIST
Baseline	7.74	3.84
Disambiguation with MWEs only	8.11	3.91
Disambiguation with both MWEs and SLM	8.13	3.92

For **English-Lithuanian** disambiguation, we tried a more advanced approach. We used an English-Lithuanian dictionary with a large number of phrase translations. We applied shallow parsing to it and aligned Lithuanian syntactic bigrams with the corresponding English syntactic bi-grams. Again the frequency and probability of such bilingual pairs was calculated. We call the resulting data the syntactic translation model (STM).

For English-Lithuanian translation, we find probability in the Lithuanian syntactic tree for every combination of English source and Lithuanian target words at one node connected with the same combination at other node. Probability for this bilingual pair (EN/LT –EN/LT) is found in the English-Lithuanian STM. Usage of the STM model should potentially provide improved disambiguation quality than the SLM model.

If we are translating sentence “cats chase mice”, then we have tree illustrated in Figure 4.3.6 before the disambiguation phase. A set of translations is attached to each node. For example, set of translations T_2 = {akinti, aptaisas, gainioti, …} is attached to node “chase”. Let’s label source words at tree nodes as s_1, s_2 and s_3; and T_1 elements as $t_{1,1}$, $t_{1,2}$ etc.; and similarly T_2 elements as $t_{2,1}$, $t_{2,2}$, …, T_3 elements as $t_{3,1}$, $t_{3,2}$ ...

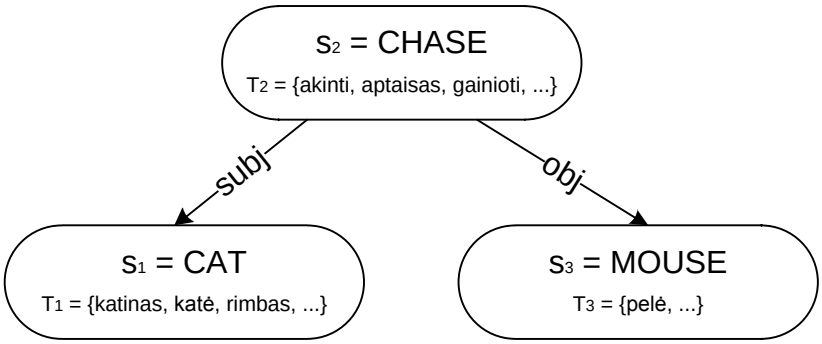


Figure 4.3.6 The translation tree before disambiguation, English-Lithuanian

To disambiguate this tree we have to find $t_{1,i} \in T_1$, $t_{2,j} \in T_2$ and $t_{3,k} \in T_3$ for the s_1, s_2 and s_3 giving the highest probability for the whole tree. This is expressed by the equation (12).

$$\underset{t_{1,i} \in T_1 \,;\, t_{2,j} \in T_2 ;\; t_{3,k} \in T_3}{\operatorname{argmax}} P\left(\text{subj}\, t_{1,i} s_1\, t_{2,j} s_2\right) \times P\left(\text{obj}\, t_{2,j} s_2\, t_{3,k} s_3\right) \qquad (12)$$

where P(subj $t_{1,i}$ s_1 $t_{2,j}$ S_2) is a probability that $t_{1,i}$ and $t_{2,j}$ is syntactically related with relation *subj*, given that T_1 is a set of translations of s_1 and T_2 is a set of translations of s_2. And so on.

It was discovered that for quality improvements we need much larger bilingual corpus of phrase translations than we have from the English-Lithuanian dictionary we used. The SLM model demonstrates slightly better results but the difference is not statistically significant, another comparison should be performed using a larger bilingual corpus.

For **Latvian-Russian** disambiguation, an approach similar to the one used for English-Lithuanian was used, but we ignored type of syntactic relation (*subj, obj* etc.) between words. This allowed us to calculate probabilities without parsing. This is important aspect because we did not have reliable parser able to parse Latvian text. We used two parallel corpora (i) Russian-Latvian dictionary with a large number of phrase translations and (ii) small parallel corpus consisting of news and legal documents (ca. 0.17 mil. sentences).

Parallel corpus was aligned using GIZA++ (Och and Ney, 2003) and phrase table was built using Moses SMT toolkit (Koehn 2003). We filtered the created phrase table to leave only phrases in length 2 and we converted all words in phrases to their base-forms. As the result Latvian bi-grams with the corresponding Russian bi-grams and probabilities where obtained.

If we are translating sentence “kaki ķer peles”, then we have a translation tree illustrated in Figure 4.3.7 before the disambiguation phase. A set of translations is attached to each node. For example, set of translations T_2 = {ловить, хватать, …} is attached to node “ķert”. Let’s label source words at tree nodes as s_1, s_2 and s_3; and T_1 elements as $t_{1,1}$, $t_{1,2}$ etc.; and similarly T_2 elements as $t_{2,1}$, $t_{2,2}$, …, T_3 elements as $t_{3,1}$, $t_{3,2}$ ...

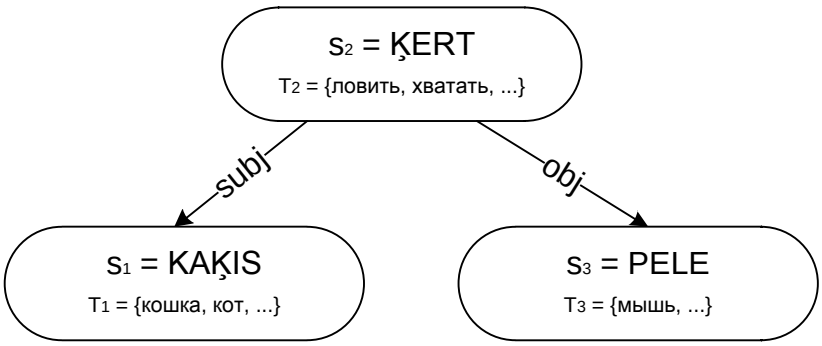


Figure 4.3.7 The translation tree before disambiguation, Latvian-Russian

To disambiguate this tree we have to find $t_{1,i} \in T_1$, $t_{2,j} \in T_2$ and $t_{3,k} \in T_3$ for the s_1, s_2 and s_3 giving the highest probability for the whole tree. But in contrast to previous two experiments, we do not have information about the syntactic relations between words in our model. This is expressed by the equation (13).

$$\underset{t_{1,i} \in T_1 \,;\, t_{2,j} \in T_2 ;\; t_{3,k} \in T_3}{\operatorname{argmax}} P\left(t_{1,i} s_1 \; t_{2,j} s_2 \right) \times P\left(t_{2,j} s_2 \; t_{3,k} s_3 \right) \qquad (13)$$

where $P(t_{1,i}\, s_1\, t_{2,j}\, S_2)$ is a probability that $t_{1,i}$ and $t_{2,j}$ is a bi-gram, and that T_1 is a set of translations of s_1 and T_2 is a set of translations of s_2. And so on.

Like in experiment with English-Lithuanian it was discovered that for quality improvements corpus of phrases extracted from bilingual dictionary is not sufficient. In fact disambiguating with the model trained on phrases from the dictionary quality even slightly decreased. But disambiguating with a model trained on parallel corpus we got slight quality improvement (14.84 vs 14.6 BLEU points). Evaluation results are summarized in Table 4.3.3.

Table 4.3.3 Evaluation results for Latvian-Russian MT system with lexical disambiguation

System	BLEU	NIST
Baseline	14.60	5.08
Disambiguation with a model trained on phrases from dictionary	14.42	5.03
Disambiguation with a model trained on parallel corpus	14.84	5.10

4.3.4 Interpretation of results

Three different experiments where performed to introduce a statistical lexical disambiguator into existing rule-based MT system. Several challenges where faced during these experiments:

- Luck of resources. Parallel corpus and syntactic parser is necessary to build disambiguation models. Both are an issue for under-resourced languages like Latvian. Small corpora extracted from bilingual dictionaries and ad hock shallow parsers where used to overcome this challenge.
- An increase of system complexity and hardware requirements. The MT system got much more complex after adding the disambiaguator; the translation speed decreased 3-5 times because probability maximization process needs much CPU time; the required disk space increased, statistical deisambiguation model built on quite small parallel corpus (0.17 mil. sentences) takes more than 500 MB disk space and this fact makes it difficult to deploy such MT system locally on end-user computers.

Experiments show that it is hard to resolve ambiguities in rule-based systems, since there is no natural way to assign scores or probabilities to the dictionary entries and various rules. Statistical disambiguator was integrated in a framework of existing MT system without significant redesign of the system. The evaluation of improved MT systems shows slight improvements in quality, but this improvement does not bring system to a new quality level.

To achieve bigger quality improvement using statistical techniques in disambiguation MT system needs to be redesigned to find more effective ways to add probabilities to the rules and dictionaries.

4.4 Rule-based MT with a Dictionary Obtained from the Corpus

4.4.1 Motivation

Treatment of Multiword Expressions (MWEs) is one of the most complicated issues in natural language processing, especially in MT. The paper by Deksne et al. (2008) presents

dictionary of MWEs for English-Latvian MT system, demonstrating a way how MWEs could be handled for inflected languages with rich morphology and rather free word order. The paper demonstrates this approach on different MWE types, starting from simple syntactic structures, followed by more complicated cases and including fully idiomatic expressions. Automatic evaluation shows that the described approach increases the quality of translation by 0.6 BLEU points. This increase is encouraging but development of MWE dictionary is very time and labor consuming task. Therefore an experiment was performed to research possibility to use corpus-based techniques used in statistical MT to automate development of MWE dictionary.

This research was also inspired by work of Dugast et al. (2009a) and Surcin et al. (2007) who also used corpus-based techniques to develop lexicons for rule-based MT systems.

4.4.2 MWE processing in rule-based MT system

Existing Tilde's multilingual transfer-based MT system (Skadina et al. 2007b, Deksne et al. 2008) was used to test methods developed for extracting MWE dictionary from parallel corpus. The overall description of the system architecture and main modules is given in chapter 4.3.2. This chapter gives more detailed description of MWEs and MWE processing in rule-based MT system.

Multiword Expressions

There are many cases in real texts when the meaning of collocation is not based on the meaning of its parts. Usually such phrases are called Multiword Expressions (MWEs). MWEs include a large range of linguistic phenomena, such as nominal compounds, phrasal verbs, idiomatic expressions, terminology and institutionalized phrases.

MWEs cannot be treated by general, compositional methods of linguistic analysis due to unclear semantics. Such approach causes over-generation in cases when the meaning could be inferred from the words, e.g., 'telephone box' (Sag et al, 2002). Sag points to the idiomaticity problem for MWEs with opaque semantics: how to predict cases when MWE has a meaning which is unrelated to the meanings of its constituents (words), e.g., the meaning of idiom 'raining cats and dogs' is not related to 'cats' and 'dogs'.

Although meaning of MWEs cannot be derived from its component words, MWEs behave like any other phrase in a sentence, e.g., they take inflections, undergo syntactic operations

etc.; at the same time, when MWE is translated, its syntactic structure in the translated phrase can be completely different from the source phrase. Different strategies have been used for encoding of MWEs in different lexical resources. For languages with minimal inflection a lot of MWEs can be fixed in the lexicon as words with spaces. This approach is inappropriate for highly inflected languages with rather free word order where each MWE can have a lot of different morphological variants and can be used in the sentence in different syntactic roles.

Alvey Tools Lexicon (Carroll and Grover, 1989) provides good coverage of phrasal verbs with detailed information about syntactic aspects, but without distinguishing compositional from non-compositional entries and not specifying entries that can be productively formed. WordNet (Fellbaum, 1998) covers a large number of MWEs, but does not provide information about their variability. Neither of these resources covers idioms (Villavicencio et al., 2004). According to Villavicencio, “the challenge in designing adequate lexical resources for MWEs, is to ensure that the variability and the extra dimensions required by the different types of MWE can be captured”. Calzolari et al. (2002) focus on MWEs that are productive and present regularities which can be generalised and applied to other classes of words with similar properties.

Following this approach, Deksne et al. (2008) proposes flexible architecture for a lexical encoding of MWEs, which allows the unified treatment of different kinds of MWE in the translation process, taking into account syntactic similarities. In rule-based MT system processing of MWEs is one of the modules in the system which allows identifying, translating and generating MWEs as part of the sentence.

Transfer of source language syntactic structures into the corresponding target language syntactic structures during the translation process could be implemented in many different ways. Mel'čuks lexical functions (LFs) (Mel’čuk, 1974) establish a semantic relation between one word or word combination, which is called function argument, and another word or word combination, which is called function value corresponding to this argument. LFs are universal regarding the language and therefore the translation could be acquired by identifying the arguments and the value of the LF during parsing and by substituting with the correct value from the target language dictionary during generation (Apresjan et al 2002).

A different approach is the usage of Lexicalized Tree Adjoining Grammar (LTAG) (Abeillé et al, 1990). The transfer between two languages can be realized by directly putting large elementary units into correspondence without going through interlingual representation and without major changes to the source and target grammars. Transfer rules are stated as correspondences between nodes of trees which are associated with words.

In Tilde's MT system dictionary of MWEs consists of (i) a lexicon of phrases and (ii) a set of MWE rules. The lexical entry consists of a normalized source language MWE, its translation equivalent and an identifier of MWE rule describing syntactic structures of the source and the target MWE (see Table 4.4.1). Usually one rule describes tens, hundreds or even thousands of MWEs. Depending on the syntactic structure of the MWE, normalized MWE could be a list of the words in a base form or/and inflected or conjugated forms of the words.

Table 4.4.1 Lexicon of phrases

Source phrase	**Target phrase**	**Rule ID**
talk around	runāt apkārt	V-ADV-7
clever boots	slīpēts zellis	A-N-9
have a swim	izpeldēties	V-DET-N-1
get a cold	saaukstēties	V-DET-N-1
sound a false note	uzņemt nepareizu toni	V-DET-A-N-14
out of temper	saniknots	ADV-PREP-N-1
have lunch	ēst pusdienas	V-N-3

The MWE rule describes the syntactic structure of MWE in the source language and its transformation into the corresponding structure of the target language.

In simplest cases the source and target MWEs have the same syntactic structure and translations of words are attached to the corresponding nodes of syntactic tree. Figure 4.4.1 shows the rule for such type of MWEs consisting of a main verb (V) and an adverb (ADV). It starts with the rule identifier V-ADV-7 followed by the syntactic structure of MWE in the source and target language (V1[advl:ADV2]=>V1[advl:ADV2]) and providing characteristics of the normalized phrase, e.g., V1.SourceBaseform stands for the verb in base form, ADV2.SourceSpelling stands for the adverb in its written form.

```
IdiomRule(V-ADV-7)
V1[advl:ADV2]=> V1[advl:ADV2]
{
V1.SourceBaseform;
ADV2.SourceSpelling;

V1.TargetBaseform;
ADV2.TargetSpelling;
}
//talk(V1) around(ADV2) => runāt(V1) aplinkus(ADV2)
```

Figure 4.4.1 Example of a simple MWE rule

The MWE rule can also include morphological restrictions for a certain source language parse tree node and assign morphological features for a certain target language parse tree node. Figure 4.4.2 shows the rule for the English noun phrase ’clever boots’ in plural and the corresponding Latvian noun phrase ‘slīpēts zellis’ in singular.

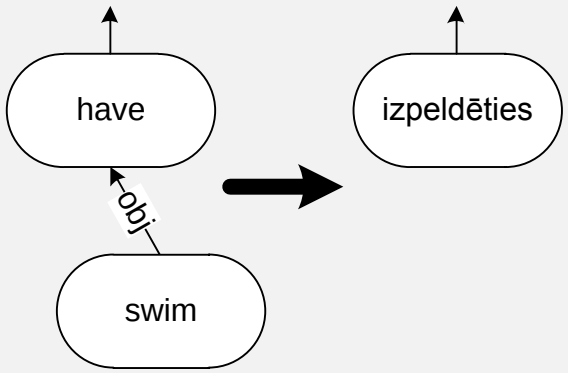
```
IdiomRule(A-N-9)
N2[attr:A1]=> N3[attr:A1]
{
N2.Number == plural;
A1.SourceSpelling;
N2.SourceSpelling;

       A1.TargetBaseform;
N3.TargetBaseform;
N3.Number = singular;
}
//clever(A1) boots(N2) => slīpēts(A1) zellis(N3)
```

Figure 4.4.2 Example of a simple MWE rule

Although the simplest MWEs form a considerable part of the MWE dictionary, most of the MWE rules are more complicated and describe transformation of parse tree between the source and target languages. Some nodes can be dropped from the source tree, some new ones can be added in the target tree during a transfer. In the most complicated cases the head node of the fragment tree can be changed into a different one. Figure 4.4.3 shows how English MWE ‘have a swim’ is transformed into a single Latvian word ‘izpeldēties’.

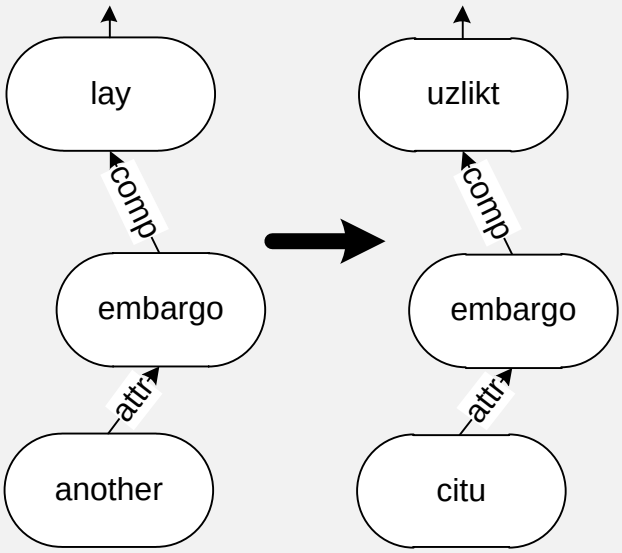
Similar syntactic structures can be translated differently depending on the context they are used. Figure 4.4.4 and Figure 4.4.5 show the translation process for the MWE ‘lay an embargo’ in two cases: as a single verb ‘apķīlāt’ or a verb phrase ‘uzlikt embargo‘. The rule from the Figure 4.4.5 will be applied only if the noun node N2 has no other children as the only optional determiner DET3; in this case the translation is a single verb and we can drop

the N2 node in the target tree. In other cases the rule from the Figure 4.4.4 is performed, i.e., the same tree structure is kept.




```
IdiomRule(V-DET-N-1)
V1[obj:N2[?det:DET3,null]]=>V1
{
V1.SourceBaseform;
DET3.SourceSpelling;
N2.SourceBaseform;

V1.TargetBaseform;
}

//have(v1) swim(N2) => izpeldēties(V1)
```

Figure 4.4.3 MWE rule where the source and target tree have different syntactic structures




```
V1[comp:N2[?det:DET3]]=>V1[comp(right):N2]
```

Figure 4.4.4 Translation of ‘to lay an embargo’: similar syntactic structures in source language have different target language tree

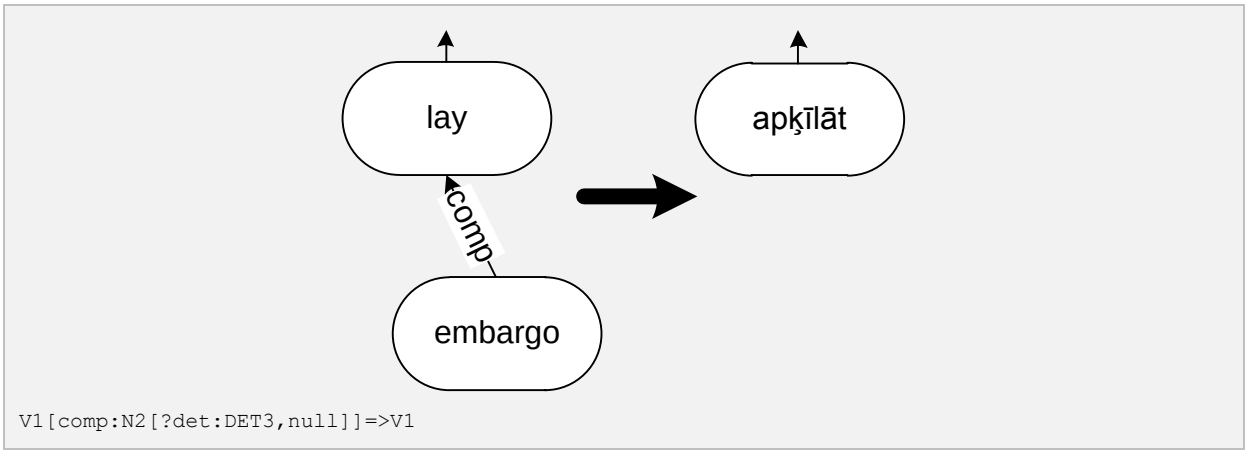


Figure 4.4.5 Translation of 'to lay an embargo': similar syntactic structures in source language have different target language tree

Not only the structure of the syntactic tree, but also the word order can be changed during the translation process. Therefore, in the description of the target language tree, we specify not only the parse tree and the syntactic relations but also the word order, i.e., the position of the child node in respect to the parent node. The child node can be inserted directly before the parent ('left'), at the beginning of phrase ('leftmost'), directly after the parent ('right') or at the end of phrase ('rightmost').

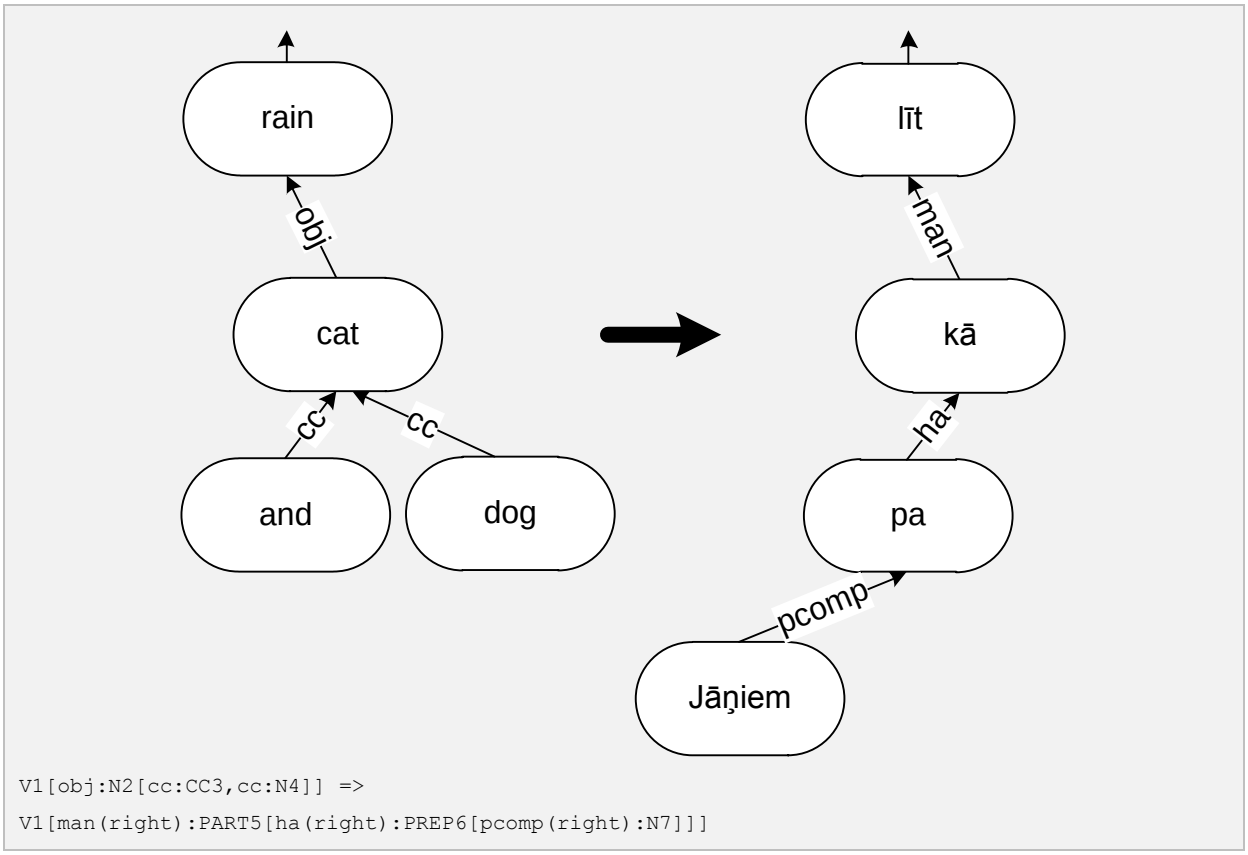


Figure 4.4.6 Source and target tree for idiomatic expressions

Truly idiomatic expressions have completely different phrase structure in the source and the target languages. Figure 4.4.6 illustrates the translation of the idiom ‘raining cats and dogs’ into ‘līst kā pa Jāņiem’ ('it's raining like on Midsummer's Day'). Only the main verb node V1 is kept in target tree during the transfer, all other target nodes have been replaced with different ones.

Processing of Multiword Expressions

The English-Latvian MT system is built from separate components, each of them having their own functionality. Components are executed successively during the translation process: the system detects the language of the source text, builds the syntactic parse tree, performs MWE processing, performs syntactic and lexical transfer, disambiguates word translations, and establishes morphological agreement between words.

Input of the MWE module is the parse tree of the source language sentence. The MWE processing module traverses the parse tree top-down trying to identify the potential MWEs, i.e., patterns (fragments of parse tree) defined in MWE rules. If a match is found, the MWE rule looks for a lexical match in the lexicon of phrases. If the matching entry is found in the lexicon of phrases, the target tree fragment is created and lexical translations are attached to the right nodes.

The translated MWE is integrated into the target tree to be used later during transfer, agreement and other processes. In these modules MWE is treated in the same way as other words in sentence (conjugated, declined, etc.) to create a fluent target language sentence.

Impact on quality

MWE dictionary (Deksme et al. 2008) has a lexicon of 19,790 English MWEs with their translations, and 914 rules. The most frequent phrases are adjective-noun phrases (6995 entries), noun-noun phrases (3912 entries), verb-noun phrases (2597 entries), noun-preposition-noun phrases (1674), and verb-preposition-verb phrases (1010 entries).

For the evaluation we used English-Latvian test corpus described in chapter 3. The aim of automatic evaluation task was to evaluate impact of MWE processing module on translation quality. Two popular evaluation metrics NIST (Doddington 2002) and BLEU (Papineni et al. 2002) were chosen for automatic evaluation. The evaluation results for MWE processing module in English-Latvian MT are summarized in Table 4.4.2.

Table 4.4.2 Evaluation results for English-Latvian MT with MWE processing

System characteristics	BLEU	NIST
Without MWE processing	7.74	3.84
With MWE processing	8.13	3.92

BLEU score rose by 0.4 points while NIST score rose by 0.08 points when MWE processing module was included. MWE processing module detected and provided translations for 83 MWEs in the test corpus.

4.4.3 Experiment – extracting MWE dictionary from parallel corpus

Extracting MWE lexicon

This experiment describes work which has been done to find automated way how to increase MWE lexicon.

MWE lexicon consists of aligned phrases which are attached to MWE rules. Statistical MT systems also use aligned phrases in translation model. Tools used for SMT training were used use to get a list of aligned phrases from the English-Latvian part of JRC-Acquis parallel corpus (Steinberger et al., 2006). Parallel corpus was aligned using GIZA++ (Och and Ney, 2003) and phrase table was built using Moses SMT toolkit (Koehn et al., 2007). Typically SMT systems do not have any linguistic constraints on phrases, which mean that phrases are not linguistically motivated in any way. SMT phrase table was filtered to remove all phrases which do not match structure of MWE rule. Baseline MT system operates with a set of MWE rules contains 914 rules. A subset of 3 main MWE rules was selected to validate hypothesis that automated MWE extraction is possible and it gives MT quality improvement; working with all 914 rules would be too time and labor consuming to perform the experiment. The tree selected MWE rules are:

- to take a N → V
  (to take a swim → izpeldēties)
- N 1 of N2 → Ngen2 N1
  (number of errors → kļūdu skaits)
- N1 N2 → Ngen1 N2
  (training data → treniņa dati)

These are rules with simple and easy structure and they are commonly used in texts. The majority of lexicon of current MWE dictionary is attached to rules with such or very similar structure.

Typically GIZA++ is used to create phrase tables for SMT. GIZA++ is run in both directions and phrases are extracted from the union or intersection of both results. For this experiment the union was used, because we are interested in big amount of phrases not in high precision of alignment. Wrongly aligned phrases are expected to be filtered out during the filtering phase of the experiment. GIZA++ usually works with texts as they are or with lemmatized texts. It is not enough for this experiment because we have to find phrases with a certain phrase structure during filtering phase. The base form in combination with the part of speech information is used for alignment in this experiment; it means that texts were preprocessed and all words were replaced with their base forms and part of speech. The Connexor English parser[17] and Latvian POS-tagger (Pinnis and Goba, 2011) was used in preprocessing. The base form and part of speech information does not give complete information necessary for filtering, therefor the full morpho-syntactic tag returned by Latvian POS-tagger was used for Latvian .

The created phrase table was filtered to leave only phrases matching 3 selected MWE rules. Filtering was implemented as simple Perl script.

Evaluation of lexicon extraction

The MWE dictionary of the baseline MT system has a lexicon of 19,790 English MWEs with their translations, and 914 rules. As expected – a large number of phrases relevant to the selected MWE rules were found during the experiment. (See Table 4.4.3.)

Table 4.4.3 Comparison of the size of baseline MWE lexicon with size of MWE lexicon obtained during the experiment

MWE type	Lexicon of baseline MT system	New phrases found
to take a N → V	2,527	3,673
N 1 of N2 → Ngen2 N1	1,674	47,325
N1 N2 → Ngen1 N2	3,912	73,446

[17] http://www.connexor.com/

The phrase table filtering to find phrases with a specific structure is a typical Information Extraction (IE) task. Therefore popular IE evaluation metrics (Precision and Recall[18]) are used to evaluate this task. 100 random phrases attached to each of 3 MWE rules where manually checked to calculate precision and 25 random sentences containing phrases with required structure were manually checked to calculate recall. (See Table 4.4.4)

Table 4.4.4 Precision and Recall of phrase filtering

MWE type	Precision	Recall
to take a N → V	0.92	1.00
N 1 of N2 → Ngen2 N1	0.73	0.65
N1 N2 → Ngen1 N2	0.85	0.70

Main reasons of low precision (wrong phrases found) are as follows: (i) discontinuous phrases are not fond properly, (ii) parsing and POS-tagging errors. Main reasons of low recall (phrases where not found) are as follows: (i) parsing and POS-tagging errors, (ii) insufficient information about phrase structure for filtering, (iii) discontinuous phrases.

Experiment looks promising. Although it has been performed only on very small set of MWE rules and many simplifications were made during this experiment, many new and relevant MWE lexicon items were found. Precision of phrase filtering is not very high, but it looks good enough. This experiment only shows that it is possible to use methods described to automatically obtain MWE lexicon from corpus.

Evaluation in MT system

This chapter deals with a question: does automatically obtained MWE lexicon improve quality of MT system.

Several things had to be done to add the newly obtained MWE lexicon to the system. The tree new MWE rules were created because rules used in the experiment are not a part of the baseline system. The baseline system has many other MWE rules which match MWE rules used in this experiment. The original MWE rules are more restrictive, for example, system has 18 MWE rules with the same structure like the second rule (N 1 of N2 → Ngen2 N1); some of these 18 rules add restrictions to case and number of source words, some requires to use specific number and gender in a target phrase. MWE processing was also changed so

[18] http://en.wikipedia.org/wiki/Precision_and_recall

that the new generic MWE rules will always be preferred in cases were both original and new rules can be applied. The evaluation results are summarized in Table 4.4.5.

For the evaluation English-Latvian test corpus described in Chapter 4.2.1 was used. Two popular automatic evaluation metrics NIST (Doddington, 2002) and BLEU (Papineni et al., 2002) was used for automatic evaluation. To evaluate an impact of automatically learned MWE lexicon on quality of MT system, the newly obtained lexicon was added to the baseline system and BLEU and NIST scores were calculated.

Table 4.4.5 Evaluation results for English-Latvian MT with automatically obtained MWE lexicon

System characteristics	BLEU	NIST
Baseline system	8.13	3.92
Modified system – with new generic MWE rules and automatically obtained lexicon	8.34	4.07

4.4.4 Interpretation of results

The experiment showed that it is possible to use methods described above to automatically obtain MWE lexicon from parallel corpus and a large number of phrases relevant to the selected MWE rules were found during the experiment. It also showed that automatic extraction of the lexicon from corpus is difficult. Very simplistic approach was used in the experiment.

Results of the experiment look promising but they do not bring quality of the whole MT system to the new level.

4.5 Statistical MT with a rule-based morphological analyzer

Part of the results of the experiment has been reported in papers (Skadiņš et al., 2010, Skadiņš et al., 2011a; 2011c) introducing Tilde's statistical MT system and the LetsMT! platform.

4.5.1 Motivation

Besides Google machine translation engines and research experiments with statistical MT for Latvian (Skadiņa and Brālītis, 2009) and Lithuanian, there are both English-Latvian (Skadiņš et al., 2008) and English-Lithuanian (Rimkute and Kovalevskaite, 2008) rule-based MT systems available. Both Latvian and Lithuanian are morphologically rich languages with quite free phrase order in a sentence and with very limited parallel corpora available. All mentioned

aspects are challenging for SMT systems. The aim of the experiment was not to build yet another SMT using publicly available parallel corpora and tools, but also to add language specific knowledge to assess the possible improvement of translation quality. Another important aim of this experiment was the evaluation of available MT systems; we wanted to understand whether we can build SMT systems outperforming other existing statistical and rule-based MT systems.

4.5.2 SMT system

We used Moses SMT toolkit (Koehn et al., 2007) for SMT system training and decoding. The baseline SMT models were trained on lowercased surface forms only for source and target languages. The SMT baseline models were trained for reference point to assess the relative improvement of additional data manipulation, factors, corpus size and language models.

The phrase-based approach in SMT allows translating source words differently depending on their context by translating whole phrases, whereas target language model allows matching target phrases at their boundaries. However, most phrases in inflectionally rich languages can be inflected in gender, case, number, tense, mood and other morphosyntactic properties, producing considerable amount of variations.

Both Latvian and Lithuanian belong to the class of inflected languages which are the most complex from the point of view of morphology. Latvian nouns are divided into 6 declensions. Nouns and pronouns have 6 cases in both singular and plural. Adjectives, numerals and participles have 6 cases in singular and plural, 2 genders, and the definite and indefinite form. The rules of case generation differ for each group. There are two numbers, three persons and three tenses (present, future and past tenses), both simple and compound, and 5 moods in the Latvian conjugation system. Latvian is quite regular in the sense of forming inflected forms however the form endings in Latvian are highly ambiguous. Nouns in Latvian have 29 graphically different endings and only 13 of them are unambiguous, adjectives have 24 graphically different endings and half of them are ambiguous, verbs have 28 graphically different endings and only 17 of them are unambiguous. Lithuanian has even more morphological variation and ambiguity. Another significant feature of both languages is the relatively free word order in the sentence which makes parsing and translation complicated.

The inflectional variation increases data sparseness at the boundaries of translated phrases, where a language model over surface forms might be inadequate to estimate the probability

of target sentence reliably. The baseline SMT system was particularly weak at adjective-noun and subject object cases.

Following the approach of English-Czech factored SMT (Bojar et al., 2009) we introduced an additional language model over disambiguated morphologic tags in the English-Latvian system. The tags contain morphologic properties generated by a rule-based morphological analyzer and statistical morphology tagger. The tags contain relevant morphologic properties (case, number, gender, etc.) that are generated by a morphologic tagger (Pinnis and Goba, 2011). The order of the tag LM was increased to 7, as the tag data has significantly smaller vocabulary.

When translating from morphologically rich language, the SMT baseline system will not give translation for all forms of word that is not fully represented in the training data. The solution addressing this problem would be to separate richness of morphology from the words and translate lemmas instead. Morphology tags could be used as additional factor to improve quality of translation. However, as we do not have a morphologic tagger for Lithuanian we used a simplified approach, splitting each token into two separate tokens containing the stem and an optional suffix. The stems and suffixes were treated in the same way in the training process. Suffixes were marked (prefixed by a special symbol) to avoid overlapping with stems.

The suffixes we used correspond to inflectional endings of nouns, adjectives and verbs, however, they are not supposed to be linguistically accurate, but rather as a way to reduce data sparsity. Moreover, the processing always splits the longest matching suffix, which produces errors with certain words.

We trained another English-Latvian system with a similar approach, using the suffixes instead of morphologic tags for the additional LM. Although the suffixes are often ambiguous (e.g. the ending *-a* is used in several noun, adjective and verb forms), our goal was to check whether we can get improvement in quality by using knowledge about morphology in case we do not have morphological tagger, and to assess how big is this improvement compared with using the tagger.

Table 4.5.1 gives an overview of SMT systems trained and the structure of factored models.

Table 4.5.1 Structure of Translation and Language Models

System	Translation Models	Language Models
EN-LV SMT baseline	1: Surface → Surface	1: Surface form
EN-LV SMT suffix	1: Surface → Surface, suffix	1: Surface form 2: Suffix
EN-LV SMT tag	1: Surface → Surface, morphology tag	1: Surface form 2: Morphology tag
LT-EN SMT baseline	1: Surface → Surface	1: Surface form
LT-EN SMT Stem/suffix	1: Stem/suffix → Surface	1: Surface form
LT-EN SMT Stem	1: Stem → Surface	1: Surface form

4.5.3 Training resources

For training the SMT systems, both monolingual and bilingual sentence-aligned parallel corpora of substantial size are required. The corpus size largely determines the quality of translation, as has been shown both in case of multilingual SMT (Koehn et al. 2003) and English-Latvian SMT (Skadiņa and Brālītis 2009).

For all of our trained SMT systems the parallel training corpus includes DGT-TM, OPUS and localization corpora. The DGT-TM corpus is a publicly available collection of legislative texts available in 22 languages of European Union. The OPUS translated text collection (Tiedemann and Nygaard 2004, Tiedemann 2009) contains publicly available texts from web in different domains. For Latvian we chose the EMEA (European Medicines Agency) sentence-aligned corpus. For Lithuanian we chose the EMEA and the KDE4 sentence-aligned corpus. Localization parallel corpus was obtained from translation memories that were created during localization of software content, appliance user manuals and software help content. We additionally included word and phrase translations from bilingual dictionaries to increase word coverage.

Both parallel and monolingual corpora were filtered according to different criteria. Suspicious sentences containing too much non-alphanumeric symbols and repeated sentences were removed.

Monolingual corpora were prepared from the corresponding monolingual part of parallel corpora, as well as news articles from Web for Latvian and LCC (Leipzig Corpora Collection) corpus for English.

Table 4.5.2 Bilingual corpora for English-Latvian system

Bilingual corpus	Parallel units
Localization TM	~1.29 mil.
DGT-TM	~1.06 mil.
OPUS EMEA	~0.97 mil.
Fiction	~0.66 mil.
Dictionary data	~0.51 mil.
Total	**4.49 mil.**
	(3.23 mil. filtered)

Table 4.5.3 Bilingual corpora for Lithuanian-English system

Bilingual corpus	Parallel units
Localization TM	~1.56 mil.
DGT-TM	~0.99 mil.
OPUS EMEA	~0.84 mil.
Dictionary data	~0.38 mil.
OPUS KDE4	~0.05 mil.
Total	**3.82 mil.**
	(2.71 mil. filtered)

Table 4.5.4 Monolingual corpora

Monolingual corpus	Words
Latvian side of parallel corpus	60M
News (web)	250M
Fiction	9M
Total, Latvian	**319M**
English side of parallel corpus	60M
News (WMT09)	440M
LCC	21M
Total, English	**521M**

The evaluation and development corpora were prepared separately. For both corpora we used the same mixture of different domains and topics (Table 4.5.5) representing the expected translation needs of a typical user. The development corpus contains 1000 sentences, while the evaluation set is 500 sentences long.

Table 4.5.5 Topic breakdown of evaluation and development sets

Topic	Percentage
General information about European Union	12%
Specifications, instructions and manuals	12%
Popular scientific and educational	12%
Official and legal documents	12%
News and magazine articles	24%
Information technology	18%
Letters	5%
Fiction	5%

4.5.4 Results and Evaluation

Automated evaluation

We used BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) metric for automatic evaluation. The summary of automatic evaluation results is presented in Table 4.5.6.

Table 4.5.6 Automatic evaluation BLEU scores

System	Language pair	BLEU
Tilde rule-based MT	English-Latvian	8.1
Google[19]	English-Latvian	32.9
Pragma[20]	English-Latvian	5.3
SMT baseline	English-Latvian	24.8
SMT suffix	English-Latvian	25.3
SMT tag	English-Latvian	25.6
Google	Lithuanian-English	29.5
SMT baseline	Lithuanian-English	28.3
SMT stem/suffix	Lithuanian-English	28.0

For Lithuanian-English system we also measured the out of vocabulary (OOV) rate on both per-word and per-sentence basis (Table 4.5.7). The per-word OOV rate is the percentage of untranslated words in the output text, and the per-sentence OOV rate is the percentage of sentences that contain at least one untranslated word. It was not possible to automatically determine the OOV rates for other translation systems (e.g. Google), as the OOV rates were calculated by analyzing the output of Moses decoder.

Table 4.5.7. OOV rates for Lithuanian-English

System	Language pair	OOV, Words	OOV, Sentences
SMT baseline	Lithuanian-English	3.31%	39.8%
SMT stem/suffix	Lithuanian-English	2.17%	27.3%

Human evaluation

We used a ranking of translated sentences relative to each other for manual evaluation of systems. This was the official determinant of translation quality used in the 2009 Workshop on Statistical Machine Translation shared tasks (Callison-Burch et al. 2009). The same test corpus was used as in automatic evaluation. The summary of manual evaluation results is presented in Table 4.5.8.

[19] Google Translate (http://translate.google.com/) as of July 2010

[20] Pragma translation system (http://www.trident.com.ua/eng/produkt.html)

Table 4.5.8. Manual evaluation results for 3 systems, balanced test corpus

System	Language pair	BLEU	NIST	Average rank in manual evaluation
Tilde Rule-Based MT	English-Latvian	8.1	3.82	1.98 ± 0.08
SMT Baseline	English-Latvian	21.7	5.32	2.06 ± 0.07
SMT tag	English-Latvian	23.0	5.40	1.59 ± 0.07

We did evaluation both ranking several systems simultaneously and ranking only two systems (ties were allowed). We discovered that it is more convenient for evaluators to evaluate only two systems and results of such evaluations are easier to interpret as well. A detailed description of human evaluation method is given in chapter 4.2.2 "Human evaluation"

Table 4.5.9. Manual evaluation results. Comparison of two systems

System1	System2	Language pair	p	ci
SMT tag	SMT baseline	English-Latvian	58.67 %	±4.98 %
Google	SMT tag	English-Latvian	55.73 %	±6.01 %
SMT stem/suffix	SMT baseline	Lithuanian-English	52.32 %	±4.14 %

Best factored systems where compared to baseline systems and best English-Latvian factored system to Google SMT system using system for manual comparison of two systems described above. Results of manual evaluation are given in Table 4.5.9. Manual comparison of English-Latvian factored and baseline SMT systems shows that evaluation results are sufficient to say that factored system is better than baseline system, because in 58.67% (± 4.98%) of cases users judged its output to be better than the output of baseline system. Manual comparison of English-Latvian factored and Google systems shows that Google system is slightly better, but evaluation results are not sufficient to say that it is really better, because the difference between systems is not statistically significant (55.73 – 6.01% < 50%). Manual comparison of our best Lithuanian-English and the baseline systems shows that system with stems and suffixes is slightly better, but evaluation results are not sufficient to say that with a strong confidence, because difference between systems also is not statistically significant (52.32 – 4.14% < 50%).

Evaluation in a localization scenario

Evaluation in localization was done using latest version of Tilde's SMT system; it is trained using more training data. A larger selection of parallel data was used which was automatically extracted from comparable web corpus (0.9 M sentences) and from 104 works

of fiction (0.66 M sentences). The total size of the English-Latvian parallel data used to train the translation model is 5.37 M sentence pairs and the total size of the Latvian monolingual corpus was 391 M words. The BLEU (Papineni et al., 2002) metric for automatic evaluation was used. The BLEU score of the SMT system is 35.0.

For application in the localization scenario, LetsMT! platform (Vasiļjevs et al. 2011) which provides a plug-in for the SDL Trados 2009[21] CAT environment was used. The MT system is running on the LetsMT! platform and are accessible using a web service interface. A detailed description of human evaluation in localization is given in Chapter 4.2.3 "Evaluation in a localization scenario"

The results were analyzed for 46 translation tasks by analyzing average values for translation performance (translated words per hour) and an error score for translated texts. Usage of MT suggestions in addition to the use of the translation memories increased productivity of the translators in average from 550 to 731 words per hour (32.9% improvement). There were significant performance differences in the various translation tasks; the standard deviation of productivity in the baseline and MT scenarios were 213.8 and 315.5 respectively. At the same time the error score increased for all translators. Although the total increase in the error score was from 20.2 to 28.6 points, it still remained at the quality evaluation grade "Good".

Grouping of errors identified by error classes reveal the increase of number of errors shown in Table 4.5.10.

Table 4.5.10 Comparison by error classes (error score)

Error Class	Baseline scenario	MT scenario
Accuracy	6	9
Language quality	6	10
Style	3	4
Terminology	5	7

More detailed description of the evaluation in the localization is given in papers by Skadiņš et al. (2011a) and Vasiļjevs et al. (2011a).

[21] http://www.trados.com/en/sdl-trados/default.asp

4.5.5 Interpretation of results

The MT system evaluation shows that used automatic metrics are unreliable for comparing rule-based and statistical systems, strongly favoring the latter. Both Pragma and Tilde rule-based systems have received very low BLEU score. This behavior of automated metrics has been shown before (Callison-Burch et al. 2006).

By development of factored EN-LV SMT models we expected to improve human assessment of quality by targeting local word agreement and inter-phrase consistency. Human evaluation shows a clear preference for factored SMT over the baseline SMT, which operates only with the surface forms. However, automated metric scores show only slight improvement on balanced test corpus (BLEU 21.7% vs 23.8%).

By developing of the LT-EN SMT Stem/suffix model we expected to increase overall translation quality by reduction of untranslated words. The BLEU score slightly decreased (BLEU 28.0% vs 28.3%), however the OOV rate differs significantly. Human evaluation results suggest that users prefer lower OOV rate despite slight reduction in overall translation quality in terms of BLEU score.

Current development of SMT tools and techniques has reached the level where they can be implemented in practical applications addressing the needs of large user groups in a variety of application scenarios. Evaluation results in localization promise important advances in the application of SMT in localization. The results of our experiment clearly demonstrate that it is feasible to integrate the current state of the art SMT systems for highly inflected languages into the localization process.

Error rate analysis shows that overall usage of MT suggestions decrease the quality of the translation in all error categories, particularly in language quality. At the same time this degradation is not critical and the result is acceptable for production purposes.

The created English-Latvian SMT system with Latvian morphology outperforms both a baseline SMT system without morphology and any other existing English –Latvian MT system (Skadiņš et al., 2011a; 2011c) in automatic evaluation achieving 35.0 BLEU points.

4.6 Statistical MT with Knowledge from the Ontology

Part of the results of the experiment has been reported in papers (Skadiņš et al. 2011b; Skadiņš, 2010; Skadiņš, 2011) introducing SOLIM project[22] results.

4.6.1 Motivation

Rule based MT systems use rules and knowledge in different levels of analysis, some systems deal only with morphology and shallow syntax, others use also deep syntactic analysis and some systems use even semantic analysis. Modern SMT methods use different kinds of additional knowledge (e.g. morphological or syntactical) to build more sophisticated statistical models and improve the output quality of machine translation (see, for example, factored SMT (Koehn et al., 2007), tree-based SMT (Chiang, 2007; Marcu et al., 2006; Li et al., 2009); treelet SMT (Quirk et al., 2005). However SMT systems currently are using only morphologic and syntactic information.

MT systems could benefit from various kinds of semantic knowledge in various stages of translation or training processes. Semantic information might be used in word breaking, part-of-speech tagging, syntactic disambiguation, word sense disambiguation etc. All mentioned areas are clearly distinguished in rule-based MT systems, but they are somewhat vague in SMT. There is no words sense disambiguation component in SMT, but SMT models are built so that they can deal with word ambiguities. Although various kinds of semantic knowledge could be used to improve various translation aspects, such as word sense disambiguation, translation selection or phrase reordering, the research behind the experiment described in this chapter is focusing on using of spatial information for word sense disambiguation in factored phrase-based SMT.

Toponyms are geographical names, or names of places. A natural language is ambiguous and toponyms are not exceptions. This fact makes toponyms difficult for processing, and due to their linguistic and extra-linguistic nature toponyms require special treatment (Gornostay and Skadiņa, 2009).

There are cases when real-world geographical knowledge is required for the resolution of ambiguous toponyms. The implemented SMT system deals with two types of ambiguity (see Leidner (2007) for the description of possible types of toponym ambiguity). The first type is a

[22] http://www.solim.eu

referential ambiguity, where a toponym may refer to more than one location of the same type, for example:

- Georgia as the US state and the country in Caucasus (English);
- Riga as the populated place and the capital of Latvia and as the populated place in the USA, state Michigan (Latvian);
- Šveicarija as the village in Lithuania and as the country in Europe (Lithuanian).

The second type of ambiguity is a feature type ambiguity, where a toponym may refer to more than one place of a different type, for example:

- Tanfield refers to the populated place as well as the castle in the United Kingdom (English);
- Gauja refers to the populated place as well as the river in Latvia (Latvian);
- Šventoji as the town near the Baltic Sea as well as the name of 3 different rivers in Lithuania (Lithuanian).

4.6.2 Spatial Ontology

The spatial ontology to be integrated into the machine translation process was developed using the ontology language, designed and implemented in the web ontology language (OWL) using RCC-8 properties (Region Connection Calculus) (Randell et al. 1992), and tools developed in the SOLIM project. RCC-8 properties are as follows: externally connected (EC), disconnected (DC), covered by/tangential proper part (TPP), inside/non-tangential proper part (NTPP), equal (EQ), partial overlap (PO), covers/tangential proper part inverse (TPPi), and contains/non-tangential proper part inverse (NTPPi), a visual illustration of the relations is given in Figure 4.6.1.

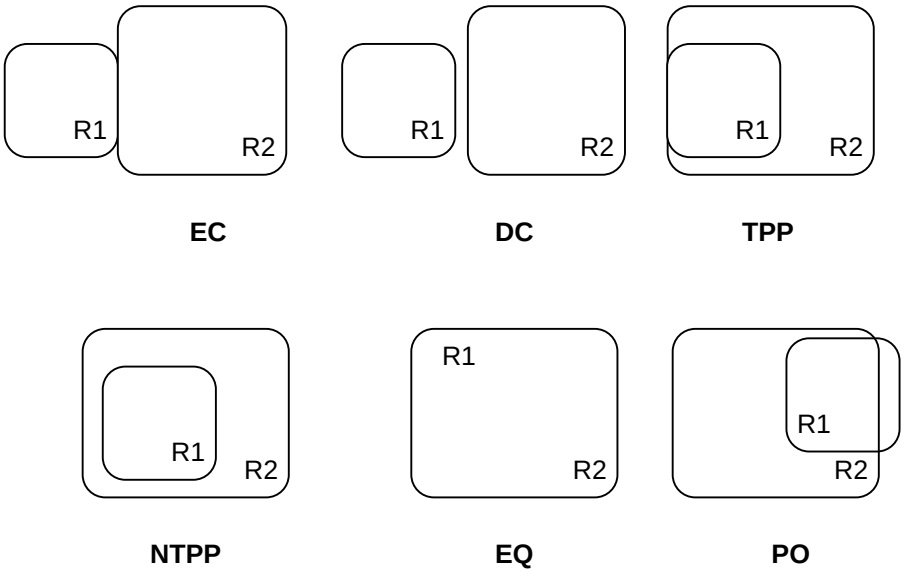


Figure 4.6.1 The standard 'base relations' of RCC and similar calculi.

The spatial ontology consisted of three sub-ontologies: basic and two language ontologies. The basic ontology contained concepts and spatial properties. The two language ontologies contained English and Lithuanian toponyms. Words in language ontologies were matched with concepts in the basic ontology (e.g. United States, US and USA represent the same concept USA). All locations in language ontologies were represented by a geo-info.owl code and lexically represented by a hasLexrep relation.

A list of instances was created on the basis of the GeoNames[23] database (7 continents, 193 countries, 51 USA states, 6359 USA cities, 6955 Lithuanian place names, 1869 cities from top 10 cities of other countries). The GeoNames database contains information about continents, countries and cities and it contains information about spatial relations between these objects. RCC-8 relations were extracted from the GeoNames database.

To query the spatial ontology we used the function GetSpatialRelations(A,B) to get spatial knowledge about relations between A and B. This information can be inferred from the spatial ontology, whereas we cannot get false or unknown information, for example:

- GetSpatialRelations(Georgia,Armenia)= "EC" only if there is enough information in the ontology to infer this relation;
- GetSpatialRelations(Georgia,Latvia)= "DC" if this relation can be inferred;

[23] www.geonames.org

- GetSpatialRelations(Georgia, Latvia)= ””, if there is not enough information in the ontology to infer the DC or any other spatial relation.

More detailed information about SOLIM spatial otology and about inferring with it is given in a paper by Skadiņš (2010).

4.6.3 MT System

The core functionality of the presented system is a disambiguation of toponyms during the machine translation process. The implemented SMT system uses semantic knowledge to improve the quality of translation, in particular with regard to the disambiguation of geographical names, or toponyms. Spatial knowledge is added to toponyms in the source text as additional semantic tags, or factors.

As a baseline system a statistical phrase-based machine translation system based on the Moses toolkit was trained on the following publicly available and proprietary corpora:

- DGT-TM parallel corpus[24] – a publicly available collection of legislative texts in 22 languages of the European Union;
- OPUS parallel corpus – a publicly available collection of texts from the web in different domains[25] (Tiedemann, 2004; Tiedemann, 2009).
- Localization parallel corpus obtained from translation memories that have been created during the localization of software, user manuals and helps.

We also included word and phrase translations from bilingual dictionaries and term translations from EuroTermBank[26] to increase word coverage.

Monolingual corpora for the training of language models were prepared from corresponding monolingual parts of parallel corpora, as well as Lithuanian news articles collected from the web. Bilingual and monolingual resources prepared and used for the baseline SMT system development are represented in Table 4.6.1 and Table 4.6.2.

[24] http://langtech.jrc.it/DGT-TM.html

[25] We chose the EMEA (medical domain) and KDE4 (IT domain) sentence-aligned corpora.

[26] www.eurotermbank.com

Table 4.6.1 Monolingual training data

Monolingual corpus	**Units**
Lithuanian side of parallel corpora	~4.04 mil. (filtered)
Web news	~5.22 mil.
Total	**~9.26 mil. (filtered)**

Table 4.6.2 Bilingual training data

Bilingual corpus	**Parallel units**
Localization TM	~5.21 mil.
DGT-TM	~1.08 mil.
OPUS EMEA	~1.04 mil.
Dictionary data	~0.27 mil.
EuroTermBank data	~0.10 mil.
KDE4	~0.05 mil.
Fiction	~0.01 mil.
Total (used for the baseline system)	**~7.76 mil.**

For the implemented system with spatial knowledge we used the same training corpora as for the baseline system, as well as prepared two more corpora from the ontology – a translation dictionary (~0,02 mil. units) and spatial relation dictionary (~0,42 mil. units).

The developed baseline SMT system was a pure phrase-based SMT system which dealt only with surface forms of words. Its translation model contained simple probabilities like:

- P(Georgia|Gruzija) – a probability that *Georgia* is the English translation of the Lithuanian word *Gruzija*;
- P(Georgia|Džordžija) – a probability that *Georgia* is the English translation of the Lithuanian word *Džordžija*.

It also contained probabilities for all morphological variants of Lithuanian words and phrases. However, it was difficult to choose the correct Lithuanian translation of a given ambiguous English toponym since both probabilities were similar:

$$P(\text{Georgia}|\text{Gruzija}) \cong P(\text{Georgia}|\text{Džordžija}). \quad (14)$$

The factored phrase-based SMT (Koehn and Hoang, 2007) is an extension of the phrase-based approach. It contains an additional annotation at a lexical unit level. The lexical unit is no longer just a token, but a vector of factors that represent different levels of annotation. The training data (a parallel corpus) has to be annotated with additional factors. For instance, it is possible to add lemma or part-of-speech information on source and target sides.

The implemented SMT system with spatial knowledge is based on the Moses toolkit (Koehn et al., 2007) that features factored translation models allowing the integration of additional layers of data directly into the process of translation. Spatial knowledge was used during training and translation processes as additional semantic factors integrated with the source language data. All toponyms in the source text were analyzed and tagged (annotated) with semantic factors (spatial knowledge) inferred from the spatial ontology with a reasoner. For example, a toponym *Georgia* is ambiguous: it can refer to the *USA state* or the *Caucasian country*. See the example sentences:

- There are Lithuanians living in Georgia, Florida and other states.
- Experts have failed to travel to Georgia at the Tbilisi airport.

In the first sentence *Georgia* refers to the *USA state*, while in the second one it refers to the *Caucasian country*. To resolve this type of ambiguity, spatial knowledge was used to determine spatial relations between corresponding toponyms within one sentence. For example, in the first sentence *Georgia* was annotated with *EC.Florida* since that information had been inferred from the spatial ontology (*Georgia* is externally connected to *Florida*). In the second sentence *Georgia* was annotated with *NTPPi.Tbilisi* (*Tbilisi* is a city in *Georgia*). We searched a sentence for toponyms and queried the spatial ontology for their relations. If there were more than two toponyms in a sentence we used just one (the first found, but not DC) annotation to each toponym. Compared with a simple unfactored translation model, that kind of factored translation model contained more useful information for toponym disambiguation since it might contain probabilities like:

- P(Georgia/EC.Florida|Džordžija) – a probability that *Georgia* is the English translation of a Lithuanian word *Džordžija* given that *Georgia* is externally connected to *Florida*;
- P(Georgia/NTPPi.Tbilisi|Gruzija) – a probability that *Georgia* is the English translation of Lithuanian word *Gruzija* given that *Georgia* encloses *Tbilisi*.

The translation model with probabilities about words and phrases with spatial knowledge helps to perform more accurate toponym disambiguation, because spatial context is included in the translation model. For example, if we have almost equal probabilities for *Georgia*, being a translation of both *Gruzija* and *Džordžija* in the translation model of the baseline system, probabilities with spatial knowledge are significantly different:

P(Georgia/EC.Armenia|Gruzija) ≫ P(Georgia/EC.Armenia|Džordžija) (15)

P(Georgia/EC.Florida|Džordžija) ≫ P(Georgia/EC.Florida|Gruzija) (16)

Thus, during the machine translation process semantic factors inferred from the spatial ontology provide additional information for the Moses decoder. As a result, it helps in choosing the appropriate translation equivalent. Therefore, SMT training data annotated with the proposed kind of spatial knowledge leads to a better machine translation quality.

It should also be mentioned that two SMT systems with spatial knowledge were trained. The first system (later referred as Spatial-8) was trained using corpora annotated with all eight RCC-8 spatial relations. The second system (later referred as Spatial-7) was trained using only seven RCC-8 relations since initial experiments, proved with the linguistic analysis, showed that using the DC:disconnected relation did not help in toponym disambiguation.

4.6.4 Results and Evaluation

A multifaceted evaluation with three procedures was applied to the evaluation of the output quality of machine translation: (i) automatic (black-box) evaluation, (ii) human evaluation and (iii) linguistic analysis.

Automatic Evaluation

For the evaluation, the test corpus described in chapter 3 was used. The aim of automatic evaluation task was to evaluate influence of spatial knowledge on translation quality. Two popular evaluation metrics NIST (Doddington, 2002) and BLEU (Papineni et al., 2002) were chosen for automatic evaluation.

BLEU and NIST scores for the baseline system were 27.35 and 5.90 correspondingly. BLEU and NIST scores for the implemented system with spatial knowledge were 27.97 (BLEU) and 5.97 (NIST) for the system "Spatial-8" and 27.47 (BLEU) and 5.91 (NIST) for the system "Spatial-7" (see Table 4.6.3).

Table 4.6.3 Results of the automatic evaluation

System	BLEU	NIST
Baseline	27.35	5.90
Spatial-8	27.97	5.97
Spatial-7	27.47	5.91

As a result, a slight improvement in the output quality of machine translation with spatial knowledge can be observed. In general, this improvement is not high and is not sufficient for the objective and an integrated evaluation procedure. Results of the automatic evaluation can be explained so that general-purpose development and evaluation corpora used for the evaluation did not contain many ambiguous geographical names. Therefore, the evaluation with the task-specific evaluation corpus was performed during the human evaluation. Nevertheless, automatic scores were set as a threshold for further experiments.

Human Evaluation

A test set of 464 English sentences containing ambiguous toponyms was developed for human evaluation purposes. A ranking of translated sentences relative to each other was used for the manual evaluation of systems. This was the official determinant of translation quality used in the 2009 Workshop on Statistical Machine Translation shared tasks (Callison-Burch et al., 2009).

A web-based human evaluation environment (Skadiņš et al., 2010) was used where source sentences and translation outputs of the two SMT systems could be uploaded as simple txt files. Once the evaluation of the two systems was set up, a link to the evaluation survey was sent to evaluators. Evaluators were evaluating the systems sentence by sentence. Evaluators saw the source sentence and the translation output of the two SMT systems – baseline and the one implemented with spatial knowledge. The frequency of preferring each system based on evaluators' answers and a comparison of the sentences was calculated. About 20 evaluators participated, each comparing translations of 50 sentences.

The manual comparison of the two systems (Baseline vs. Spatial-8) has shown that the implemented SMT system with spatial knowledge is slightly better than the baseline system: in 50.66% of cases evaluators judged its output to be better than the output of the baseline system. Results of the human evaluation do not allow us to say with certainty either the spatial SMT system is significantly better or it is disambiguating toponyms better, since the difference is not convincing and evaluators have been comparing sentences using subjective criteria and not paying a special attention to the translation of toponyms.

Linguistic Evaluation of Toponym Disambiguation

A detailed linguistic analysis of toponym disambiguation during the machine translation process was performed. The same corpus as for the human evaluation was used and the

accuracy of the toponym translation was evaluated. The accuracy of the baseline system was 84.09%. The accuracy of the Spatial-8 system was 83.87%. Since results for the baseline system were better, it was decided to analyse the impact of each spatial relation to toponym disambiguation. It was discovered that the accuracy could be increased to 88.00% if the DC:*disconnected* relation was ignored (system Spatial-7).

4.6.5 Interpretation of results

We can see that the quality of machine translation can be improved by using the semantic information from the spatial ontology. Nevertheless improvement is not big. But improvement is noticeable in specific translations. Improvement is not obvious if we are translating general texts, but it is noticeable when we translate texts with ambiguous toponyms.

The proposed approach to toponym disambiguation is not limited to:

- machine translation *per se* and can be regarded as generic, i.e. it can be also applied to other fields of natural language processing, e.g. information retrieval;
- use of spatial knowledge only: other types of implicit or inferred knowledge can be used in a similar way.

Spatial information is just one type of semantic knowledge which can be added to SMT system. Enriching SMT with other types of semantic knowledge coming from other types of ontologies is also a perspective research direction.

5 CONCLUSIONS

This research is the first large-scale work dedicated to the combined use of rule-based and corpus-based methods in machine translation for Latvian. Both theoretical and practical guidelines are provided covering aspects of building machine translation system for a small, morphologically rich and under-resourced language. Although majority on the research in this work is focused on issues related to machine translation for Latvian, several combined MT methods have been also applied and verified for other morphologically rich languages – Lithuanian and Russian.

The analysis and conclusions are based on extensive studies of the existing MT methods, best practice in the field and an evaluation of different proposed combined MT methods for their applicability and adaptation in real life scenarios.

For major combined methods are proposed – rule-based MT with a statistical lexical disambiguator, rule-based MT with a dictionary obtained from the corpus, statistical MT with a rule-based morphological analyzer, and statistical MT with knowledge form the ontology. Although there could be other combined methods, these four were selected because they comply with the scope and goals of this research – practical MT for morphologically rich and under-resourced language.

All four mentioned combined methods are analyzed and evaluated to find the optimal. Each of these methods has its strength and an area of application where it is most appropriate. The factored phrase-based statistical MT with a rule-based morphology gives the biggest quality improvement for general domain MT and for MT in software localization domain. This method outperforms all other known English-Latvian MT systems in automatic evaluation achieving 35.0 BLEU points. The method is also tested on English-Lithuanian language pair and is applicable to other morphologically rich languages too.

Author proposes the method how to integrate knowledge from the domain ontology in the statistical MT. This method helps to solve a problem of the lexical ambiguity in a domain specific MT.

A practical application of the research results in the public on-line MT system http://translate.tilde.lv and software package Tildes Birojs 2011 serves as a proof-of-concept for the proposed approach.

We can conclude that results of the thesis work prove the research hypothesis that it is possible to achieve a better MT quality by combining knowledge and corpus based MT methods; and state-of-the-art statistical MT systems can be extended with knowledge-based components and such extended MT systems provide a higher quality translation.

BIBLIOGRAPHY

Authors's publications

A list of author's publications related to this research in a chronological order. Publications where the author is not the first author are also listed below in a section "Other publications" to make a reference lookup easier.

Vasiļjevs, A., Skadiņš, R., & Skadiņa, I. (2011a). Towards Application of User-Tailored Machine Translation in Localization. In Zhechev, V. (Ed.) *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators" JEC 2011* (pp. 23-31). Luxembourg

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2011b). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In *Proceedings of the 13th Machine Translation Summit* (pp. 507-511). Xiamen, China

Skadiņš, R., Puriņš, M., Skadiņa, I., & Vasiļjevs, A. (2011a). Evaluation of SMT in localization to under-resourced inflected language. *Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011* (pp. 35-40). Leuven, Belgium

Skadiņš, R., Gornostay, T., & Šics, V. (2011b). Toponym Disambiguation in English-Lithuanian SMT System with Spatial Knowledge. *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011* (pp. 191-197). Riga, Latvia

Skadiņš, R., Goba, K., & Šics, V. (2011c). Improving SMT with Morphology Knowledge for Baltic Languages. Research Workshop of the Israel Science Foundation - Machine Translation and Morphologically-rich Languages, Haifa, Israel

Deksne, D., & Skadiņš, R. (2011). CFG Based Grammar Checker for Latvian. *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011* (pp. 275-278). Riga, Latvia

Skadiņš, R. (2011). Spatial Ontology in Factored Statistical Machine Translation. In Barzdins, J., & Kirikova, M. (Eds.) *Databases and Information Systems VI, Selected papers from*

the Ninth International Baltic Conference DB&IS 2010, Frontiers in Artificial Intelligence and Applications, Vol. 224 (pp. 153-166). Riga, Latvia: IOS Press

Vasiļjevs, A., Gornostay, T., & Skadiņš, R. (2010). LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219* (pp. 133-140). Riga, Latvia: IOS Press

Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219* (pp. 125-132). Riga, Latvia: IOS Press

Skadiņa, I., Auziņa, I., Grūzītis, N., Levāne-Petrova, K., Nešpore, G., Skadiņš, R., & Vasiļjevs, A. (2010a). Language Resources and Technology for the Humanities in Latvia (2004–2010). In *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219* (pp. 15-22). Riga, Latvia: IOS Press

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., & Gornostay, T. (2010b). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In *Proceedings of 3rd Workshop on Building and Using Comparable Corpora*. BUCC 2010, Valletta, Malta

Skadiņš, R. (2010). Spatial Ontology in Statistical Machine Translation. *Proceedings of the Ninth International Baltic Conference Baltic DB&IS 2010* (pp. 409-421). Riga, Latvia

Deksne, D., Skadiņš, R., Skadiņa, I. (2008). Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (pp. 1401-1405). Marrakech, Morocco

Skadiņš, R., Skadiņa, I., Deksne, D., & Gornostaja, T. (2008). English/Russian-Latvian Machine Translation System. In Čermák, F., Marcinkevičienė, R., Rimkutė, E., & Zabarskaitė, J. (Eds.) *Proceedings of the Third Baltic Conference on Human Language Technologies* (pp. 287-296). Vilnius, Lithuania

Skadiņa, I., Vasiļjevs, V., Deksne, D., Skadiņš, R., & Goldberga, L. (2007). Comprehension Assistant for Languages of Baltic States. *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007* (pp. 167-174). Tartu, Estonia

Горностай, Т., Васильев, А., Скадиньш, Р., Скадиня, И. (2007). Опыт латышско↔русского машинного перевода. *Труды международной конференции «Диалог 2007»* (pp. 137-146). Бекасово

Deksne, D., Skadiņa, I., Skadiņš, R., & Vasiļjevs, A. (2005). Foreign language reading tool – first step towards English-Latvian commercial Machine Translation. In *Proceedings of the Second Baltic Conference on Human Language Technologies* (pp. 113-118). Tallinn, Estonia

Other publications

Abeille, A., Schabes, Y., & Joshi, A. (1990). Using Lexicalized Tags forMachine Translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)* (pp. 1-6). Helsinki, Finland

Alegria, I., Ezeiza, N., & Fernandez, I. (2008). Translating Named Entities using Comparable Corpora. *Proceedings of the Workshop on Comparable Corpora*, *LREC'08* (pp. 11-17)

Aleksić, V., & Thurmair, G., (2011). Personal Translator at WMT 2011 – a rule-based MT system with hybrid components. *Proceedings of the 6th Workshop on Statistical Machine Translation* (pp.303-308). Edinburgh, Scotland, UK

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I. D., Och, F. J., Purdy, D., Smith, N. A., & Yarowsky, D. (1999). Statistical Machine Translation: Final Report. Johns Hopkins University 1999 Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA

Alshawi, H., Bangalore, S., & Douglas, S. (1998). Automatic acquisition of hierarchical transduction models for machine translation. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL)*

Apresjan, J.D., Boguslavsky, I.M., Iomdin, L.L., Tsinman, L.L. (2002). Lexical Functions in NLP: Possible Uses, Computational Linguistics for the New Millennium: Divergence or Synergy? In Klenner, M., & Visser, H. (Eds.), *Proceedings of the International Symposium* (pp. 55-72). Ruprecht-Karls-Universität Heidelberg, Frankfurt am Main

Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. L., & Sadler, L. (1994). Machine Translation: an Introductory Guide. Blackwells-NCC.

Attnäs, M., Senellart, P., & Senellart, J. (2005), Integration of SYSTRAN MT systems in an open workflow. *Proceedings of MT Summit*, Phuket, Thailand

Axelrod, A. E. (2006). *Factored Language Models for Statistical Machine Translation.* M.Sc. Thesis. University of Edinburgh

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5.179–190

Bangalore, S., Murdock, V., & Riccardi, G. (2002). Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *Proceedings of 19th International Conference on Computational Linguistics* (pp. 1–7). Taipei, Taiwan

Barrault, L. (2010). MANY: Open Source Machine Translation System Combination. *The Prague Bulletin of Mathematical Linguistics 93* (pp. 147–155). ISBN 978-80-904175-4-0

Bilmes, J., & Kirchhoff, K. (2003). Factored Language Models and Generalized Parallel Backoff. In *Proceedings of HLT/NACCL*

Boguslavsky, I. (1999). Translation to and from Russian: the ETAP system. EAMT Workshop, Prague

Bojar, O., Mareček, D., Novák, V., Popel, M., Ptáček, J., Rouš, J., & Žabokrtský, Z. (2009). English-Czech MT in 2008. *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics

Bowker, L. (2002). Translation Memory Systems. *Computer-aided translation technology: a practical introduction* (pp. 92-127). Ottawa: University of Ottawa Press

Brandt, M. D., Loftsson, H., Sigurþórsson, H., Tyers, F. M. (2011). Apertium-IceNLP: A rule-based Icelandic to English machine translation system, *Proceedings of the 15th International Conference of the European Association for Machine Translation EAMT 2011*. Leuven, Belgium

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, F., & Roossin, P. (1988) A statistical approach to French/English translation. In *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages.*

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Lafferty, J., Mercer, R., & Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics* 16.79–85.

Brown, P., Lai, J., & Mercer, R. (1991). Aligning sentences in parallel corpora. In *ACL 1991: Proceedings of the 29th Meeting of the Association for Computational Linguistics* (pp. 169–176). University of California, Berkeley

Brown, P., Pietra, V. D., Pietra, S. D., & Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19.263–311.

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research, In *Proceedings of EACL 2006*

Callison-Burch, C., Koehn, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 1-28). Athens, Greece: Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., & Zaidan, O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR* (pp. 17-53). Uppsala, Sweden: Association for Computational Linguistics.

Callison-Burch, C., Koehn, P., Monz, C., & Zaidan, O. (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 22–64). Edinburgh, Scotland: Association for Computational Linguistics

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., & Frey, J. (2006). Context-based Machine Translation. In *Proceedings of the 7th Conference of the Association for*

Machine Translation in the Americas, "Visions for the Future of Machine Translation" (pp. 19-28). Cambridge, Massachusetts, USA

Carroll, J., & Grover, C. (1989). The derivation of a large computational lexicon of English from LDOCE. In B. Boguraev and E. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*. Longman

Chan, Y. S., Ng, H. T. (2005). Scaling up word sense disambiguation via parallel texts. In *Proceedings of the 20th National Conference on Artificial Intelligence* (pp. 1037–1042). AAAI, Pittsburgh, PA

Charniak, E. (2001). Immediate-head parsing for language models. In *Proceedings of the Assocation for Computational Linguistics 2001* (pp. 116–123). New Brunswick, NJ: ACL

Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based language models for machine translation. In *Proceedings of MT Summit IX*

Chen, Y., Eisele, A., Federmann, C., Hasler, E., Jellinghaus, M., & Theison, S. (2007). Multi-engine machine translation with an open-source decoder for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, *StatMT-07* (pp. 193-196). Morristown, NJ, USA: Association for Computational Linguistics

Chen, Y., Jellinghaus, M., Eisele, A., Zhang, Y., Hunsicker, S., Theison, S., Federmann, C., & Uszkoreit, H. (2009). Combining multi-engine translations with Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT-09* (pp. 42-46). Morristown, NJ, USA: Association for Computational Linguistics.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. Proceedings of the 43rd Annual Meeting of the ACL (pp. 263–270), Ann Arbor, MI

Chiang, D. (2007), Hierarchical Phrase-Based Translation. *Computational Linguistics 33*(2): 201-228.

Cicekli, I., & Güvenir, H. A. (2001). Learning translation templates from bilingual translation examples. *Applied Intelligence* 15(1): 57-76.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992). A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing. Association for Computational Linguistics. Proceedings of the Conference* (pp. 133–140). Trento, Italy.

Dandapat, S., Morrissey, S., Way, A., & Forcada, M.L. (2011). Using example-based MT to support statistical MT when translating homogeneous data in a resource-poor setting. In Forcada, M.L., Depraetere, H., Vandeghinste, V. (Eds.), *Proceedings of the 15th conference of the European Association for Machine Translation* (pp.201-208). Leuven, Belgium

Deksne, D., Skadiņa, I., Skadiņš, R., & Vasiļjevs, A. (2005). Foreign language reading tool – first step towards English-Latvian commercial Machine Translation. In *Proceedings of the Second Baltic Conference on Human Language Technologies* (pp. 113-118). Tallinn, Estonia

Deksne, D., Skadiņš, R., Skadiņa, I. (2008). Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (pp. 1401-1405). Marrakech, Morocco

Deksne, D., & Skadiņš, R. (2011). CFG Based Grammar Checker for Latvian. *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011* (pp. 275-278). Riga, Latvia

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT-02*

Du, J., Pecina, P., & Way, A. (2010). An augmented three-pass system combination framework: DCU combination system for WMT 2010. In *Proceedings of the Fifth ACL Workshop on Statistical Machine Translation* (pp. 271–276). Uppsala, Sweden

Dugast, L., Senellart, J., Simard, M., & Koehn, P. (2007). Statistical Post-Edition on SYSTRAN Rule-Based Translation System. *Proceedings of the Workshop on Statistical MT*. Prague: ACL

Dugast, L., Senellart, J., & Koehn, P. (2009a). Selective addition of corpus-extracted phrasal lexical rules to a rule-based machine translation system. *MT Summit XII: proceedings of the twelfth Machine Translation Summit* (pp.222-229). Ottawa, Ontario, Canada

Dugast, L., Senellart, J., & Koehn, P. (2009b). Statistical Post Editing and Dictionary Extraction: SYSTRAN/Edinburgh submissions for ACL-WMT2009. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 110-114). Athens, Greece: Association for Computational Linguistics.

EAGLES (1996). Preliminary recommendations on corpus typology. Electronic resource: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Ehara, T. (2007). Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *MT Summit XI Workshop on patent translation* (pp.13-18). Copenhagen, Denmark

Eisele, A. (2005). First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 155–158). Ann Arbor, MI, USA

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., & Chen, Y. (2008). Hybrid Machine Translation Architectures within and beyond the EuroMatrix project. *Proceedings of EAMT.* Hamburg.

Farghaly, A., & Senellart, J. (2003). Intuitive Coding of the Arabic Lexicon. In *Proceedings of MT Summit IX*

Federmann, C., Theison, S., Eisele, A., Uszkoreit, H., Chen, Y., Jellinghaus, M., & Hunsicker, S. (2009). Translation Combination using Factored Word Substitution. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 70-74). Athens, Greece

Fellbaum, C. (1998). Towards a representation of idioms in WordNet. *Proceedings of the workshop on the use of WordNet in Natural Language Processing Systems (Coling-ACL 1998).* Montreal

Forcada, M. L., Tyers, F. M., & Ramírez Sánchez, G. (2009). The Apertium machine translation platform: Five years on. In *Proceedings of the First International Workshop on Free/Open- Source Rule-Based Machine Translation*. Alacant, Spain

Fraser, A., & Marcu, D. (2007) Getting the structure right for word alignment: LEAF. In *Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning* (pp. 51–60). Prague, Czech Republic

Fuchs, N. E., & Schwitter, R. (1996). Attempto Controlled English (ACE). CLAW 96, The First International Workshop on Controlled Language Applications, Katholieke Universiteit Leuven

Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *ACL 1991: Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL)* (pp. 177–184). University of California, Berkeley

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., & Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 961–968). Sydney, Australia: Association for Computational Linguistics

Gasser, M. (2011). Toward Synchronous Extensible Dependency Grammar. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Germann, U., Jahr, M., Knight, K., Marcu, D., & Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of the 39th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Gieselmann, P. (2008). Architecture of the Lucy translation system. Second Machine Translation Marathon, Wandlitz, Berlin, Germany

de Gispert, A., Gupta, D., Popović, M., Lambert, P., Mariño, J., Federico, M., Ney,H., & Banchs, R. (2006). Improving Statistical Word Alignments with Morpho-syntactic Transformations. In *Proceedings of 5th International Conference on Natural Language Processing, FinTAL'06* (pp. 368-379). Turku

Gornostay, T., & Skadiņa, I. (2009). English-Latvian Toponym Processing: Translation Strategies and Linguistic Patterns. In *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation* (pp. 81-87). Barcelona, Spain: Universitat Politècnica de Catalunya

Goutte, C., Cancedda, N., Dymetman, M., & Foster, G. (eds.) (2009). *Learning Machine Translation*. Cambridge, Massachusetts, London, England: The MIT Press.

Greitāne, I. (1997). Mašīntulkošanas sistēma LATRA, *LZA Vēstis* Nr.3./4 (1997), 1-6.

Groves, D., & Way, A. (2005). Hybrid Example-Based SMT: the Best of Both Worlds? In *ACL-05: Building and Using Parallel Texts: Data- Driven Machine Translation and Beyond, Proceedings of the Workshop* (pp. 183–190). University of Michigan, Ann Arbor, Michigan, USA

Grūzītis, N., Nešpore, G., & Saulīte, B. (2010). Verbalizing Ontologies in Controlled Baltic Languages. In *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219*. Riga, Latvia: IOS Press

Hanneman, G., Ambati, V., Clark, J.H., Parlikar, A., & Lavie, A. (2009). An Improved Statistical Transfer System for French-English Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 140-144). Athens, Greece

He, X. (2007). Using Word Dependent Transition Models in HMM based Word Alignment for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 80–87). Prague: Association for Computational Linguistics

He, X., & Toutanova, K. (2009). Joint optimization for machine translation system combination. In *Proceedings of EMNLP*: Association for Computational Linguistics

Heafield, K., Hanneman, G., & Lavie, A. (2009). Machine translation system combination with flexible word ordering. In *Proceedings of the Fourth ACL Workshop on Statistical Machine Translation* (pp. 56–60). Suntec, Singapore

Heafield, K., Lavie, A. (2010). Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. In *The Prague Bulletin of Mathematical Linguistics 93* (pp. 27-36). ISBN 978-80-904175-4-0. doi:10.2478/v10108-010-0008-4

Hewavitharana, S., & Vogel, S. (2008) Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources. *Proceedings of the Workshop on Comparable Corpora, LREC'08* (pp. 7-10)

Hoang, H., Koehn, P., & Lopez, A. (2009). A uniform framework for phrase-based, hierarchical and syntax-based machine translation. In *Proceedings of the International Workshop on Spoken Language Translation* (pp. 152-159). Tokyo, Japan.

Hildebrand, A. S., & Vogel, S. (2009). CMU system combination for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 47–50). Athens, Greece

Huang, L., Knight, K., & Joshi, A. (2006) Statistical syntax-directed translation with extended domain of locality. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*. Boston, Massachusetts

Hutchins, W. J., & Somers, H. L. (1992). *An Introduction to Machine Translation*. London: Academic Press

Hutchins, J. (2006). Machine translation: history of research and use. In *Encyclopedia of Languages and Linguistics*. 2nd edition, edited by Keith Brown (Oxford: Elsevier 2006), vol.7, pp.375-383.

Hutchins, J. (2007a). Machine translation in Europe and North America: current state and future prospects. *JAPIO 2007 Yearbook.* Tokyo: Japan Patent Information Organization

Hutchins, J. (2007b). Machine translation: a concise history. In Chan Sin Wai *(Ed.), Computer aided translation: Theory and practice*. : Chinese University of Hong Kong.

Ide, N., Erjavec, T., & Tufis, D. (2002). Sense discrimination with parallel corpora. In *Proceedings of ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* (pp. 54–60). Philadelphia, PA

Itagaki, M., & Aikawa, T. (2008). Post-MT Term Swapper: Supplementing a statistical machine translation system with a user dictionary. In *LREC 2008: 6th Language Resources and Evaluation Conference*. Marrakech, Morocco

Kaji, H., Kida, Y., & Morimoto, Y. (1992). Learning translation templates from bilingual text. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)* (pp. 672-678). Nantes, France

Kaljurand, K., & Fuchs, N.E. (2007). Verbalizing OWL in Attempto Controlled English. 3rd International OWLED Workshop, Innsbruck, Austria

Karakos, D., Eisner, J., Khudanpur, S., & Dreyer, M. (2008). Machine translation system combination using ITG-based alignments. In *46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 81–84). Columbus, Ohio, USA

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., & Stolcke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech & Language* 20(4). 589-608

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607-615.

Koehn, P., Och, F. J., Marcu, D. (2003). Statistical phrase based translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL)*.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X*.

Koehn, P., & Hoang, H. (2007). Factored translation models. In *Proceedings of EMNLP-CoNLL* (pp. 868–876)

Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the ACL 2007 Demo and Poster Sessions* (pp. 177-180), Prague

Koehn, P., & Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague.

Koehn, P., Birch, A., & Steinberger, R. (2009). 462 Machine Translation Systems for Europe, *Proceedings of MT Summit XII*.

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis. Institute for Communicating and Collaborative Systems School of Informatics, University of Edinburgh

Lepage, Y., & Denoual, E. (2005). Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251-282.

Lewis, W., Wendt, C., & Bullock, D. (2010). Achieving Domain Specificity in SMT without Overt Siloing. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W., Weese, J., & Zaidan, O. (2009). Joshua: An Open Source Toolkit for Parsing-based Machine Translation. *Proceedings of the Workshop on Statistical Machine Translation (WMT09)*

Li, X., Lü, Y., Meng, Y., Liu, Q., & Yu, H. (2011). Feedback Selecting of Manually Acquired Rules Using Automatic Evaluation. In *Proceedings of the 13th Machine Translation Summit* (pp. 52-59). Xiamen, China

Liu, D., & Gildea, D. (2009). Bayesian learning of phrasal tree-to-string templates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 1308–1317). Singapore: Association for Computational Linguistics.

Macherey, W., & Och, F. J. (2007). An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 986–995). Prague, Czech Republic

Maia, B., & Matos, S. (2008). Corpógrafo V.4 – Tools for Researchers and Teachers Using Comparable Corpora. *Proceedings of the Workshop on Comparable Corpora, LREC'08* (pp. 79-82)

Marcu, D., Wang, W., Echihabi, A., & Knight, K. (2006). SPMT: statistical machine translation with syntactified target language phrases. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 44-52*)*. ACL Workshops. Association for Computational Linguistics, Sydney, Australia

Matusov, E., Ueffing, N., & Ney, H. (2006). Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33–40). Trento, Italy

May, J., & Knight, K. (2007). Syntactic Re-Alignment Models for Machine Translation. *Proceedings of EMNLP-CoNLL'07*

Mel'čuk, I. (1974). Opyt teorii lingvisticheskix modelej "Smysl ↔ Tekst". *Nauka.* Moscow

Microsoft Research. (2010). *Microsoft ® Translator Partners*. Retrieved on September 6, 2011 from http://www.microsofttranslator.com/partner/

Mitamura, T., Nyberg, E. H., 3rd. (1995). *Controlled English for Knowledge-Based MT: Experience with the KANT System*. Pittsburgh: Carnegie Mellon University, Center for Machine Translation

Munteanu, D., Fraser, A., Marcu, D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL'04.*

Munteanu, D., & Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4): 477-504.

Munteanu, D. (2006). Exploiting Comparable Corpora (for automatic creation of parallel corpora). *Online presentation. Electronic resource*: http://content.digitalwell.washington.edu/msr/external_release_talks_12_05_2005/14008/lecture.htm

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. *Proceedings of the international NATO symposium on Artificial and human intelligence* (pp. 173–180). New York, NY, USA: Elsevier North-Holland, Inc.

Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. (1992). Machine Translation: A knowledge-based approach. Morgan Kaufman, San Mateo.

Nomoto, T. (2004). Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 494–501). Barcelona, Spain

Nordfalk, J. (2009). Shallow-transfer rule-based machine translation for Swedish to Danish. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*. Alacant, Spain

Novák, A., Tihanyi, L., & Prószéky, G. (2008) The MetaMorpho translation system. In *Third Workshop on Statistical Machine Translation, Proceedings* (pp. 111-114). The Ohio State University, Columbus, Ohio, USA

Nyberg, E. H., III , & Mitamura, E. H. (1992). The Kant System: Fast, Accurate, High-Quality Translation in Practical Domains. *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Och, F. J. (1998). *Ein beispielsbasierter und statistischer Ansatz zum maschinellen Lernen von naturlichsprachlicher iibersetzung*. Master's thesis, Universitat Erlangen-Niirnberg.

Och, F. J., Ueffing, N., & Ney, H. (2001). An efficient A* search algorithm for statistical machine translation. In *Proceedings of the workshop on Data-driven methods in machine translation - Volume 14*

Och, F. J., & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. *Proceedings of the 40th Annual Meeting of the ACL* (pp. 295–302). Philadelphia, PA.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *ACL 2003: Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 160–167)

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19-51, March.

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics,* 30:417–449.

Och, F. J. (2005). Statistical Machine Translation: Foundations and Recent Advances. Tutorial at the Tenth Machine Translation Summit. Phuket, Thailand.

Pakhomov, E. (2011). Introducing ABBYY Compreno – new approach to machine translation. TAUS User Conference 2011, Santa Clara (CA), USA

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics. :* ACL

Pierce, J. R., & Carroll, J. B. (1966). *Language and machines: Computers in translation and linguistics*. Washington, DC, USA: National Academy of Sciences/National Research Council

Phillips, A. B., Cavalli-Sforza, V., & Brown, R. D. (2007). Improving example based machine translation through morphological generalization and adaptation. In *Proceedings of the 9th Machine Translation Summit (MT Summit IX)* (pp. 369-375). Copenhagen, Denmark

Pinnis, M., & Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. *Proceedings of the Second Workshop on Systems and Frameworks for Computational Morphology, Communications in Computer and Information Science, Volume 100*

Popović, M., Stein, D., & Ney, H. (2006). Statistical Machine Translation of German Compound Words. In *FinTAL - 5th International Conference on Natural Language Processing* (pp. 616-624): Springer Verlag, LNCS

Quirk, C., Menezes, A., & Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. *Proceedings of ACL 2005*

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning* (pp. 165–176). San Mateo: Morgan Kaufmann

Ranta, A., Angelov, K., and Hallgren, T. (2010). Tools for multilingual grammar-based translation on the web. In *Proceedings of the Association for Computational Linguistics System Demonstrations*, Beijing

Resnik, P., & Smith, N. (2003). The Web as a Parallel Corpus. *Computational Linguistics* 29(3):349-380.

Rimkute, E., Kovalevskaite, J. (2008). Linguistic Evaluation of the First English-Lithuanian Machine Translation System. In Čermák, F., Marcinkevičienė, R., Rimkutė, E., & Zabarskaitė, J. (Eds.) *Proceedings of the Third Baltic Conference on Human Language Technologies* (pp. 257-264). Vilnius, Lithuania

Rosti, A.-V.I., Matsoukas, S., & Schwartz, R. (2007). Improved word-level system combination for machine translation. In *Association for Computational Linguistics* (pp. 312–319)

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword Expressions: a pain in the neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–15). Mexico City, Mexico

Sánchez-Cartagena, V. M., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2011). Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of the 13th Machine Translation Summit* (pp. 562-569). Xiamen, China

Sánchez-Martínez, F. (2011). Choosing the best machine translation system to translate a sentence by using only source-language information. In Forcada, M. L., Depraetere, H., & Vandeghinste, V. (Eds.), *Proceedings of the 15th International Conference of the European Association for Machine Translation* (pp. 97-104). Centre for Computational Linguistics, Katholieke Universiteit Leuven, Leuven, Belgium

Sarikaya, R., & Deng, Y. (2007). Joint morphological-lexical language modeling for machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers on XX*, (pp.145-148). New York

Schwenk, H., Abdul-Rauf, S., Barrault, L., & Senellart, J. (2009). SMT and SPE machine translation system for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation* (pp. 130-134). Athens, Greece

Schwitter, R., Kaljurand, K., Cregan, A., Dolbear, C., & Hart, G. (2008). A Comparison of Three Controlled Natural Languages for OWL 1.1. 4th International OWLED Workshop, Washington DC

Senellart, J., & Dienes, P. (2001). New Generation SYSTRAN Translation System, In *Proceedings of MT Summit 8*

Senellart, J., Boitet, C., & Romary, L. (2003a). SYSTRAN New Generation: The XML Translation Workflow. In *Proceedings of MT Summit IX*

Senellart, J., Yang, J., Rebollo, A. (2003b) SYSTRAN Intuitive Coding Technology. In *Proceedings of MT Summit IX*

Senellart, P., & Senellart, J. (2005). SYSTRAN Translation Stylesheets: Machine Translation driven by XSLT. In *Proc. XML Conference & Exposition*. Atlanta, USA

Sennrich, R. (2011). Combining Multi-Engine Machine Translation and Online Learning through Dynamic Phrase Tables. In Forcada, M. L., Depraetere, H., & Vandeghinste, V. (Eds.), *Proceedings of the 15th International Conference of the European Association for Machine Translation* (pp. 89-96). Centre for Computational Linguistics, Katholieke Universiteit Leuven, Leuven, Belgium

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27.379–423 and 623–656.

Shen, L., Xu, J., & Weischedel, R. (2008a). A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08: HLT* (pp. 577–585). Columbus, Ohio: Association for Computational Linguistics

Shen, W., Delaney, B., Anderson, T., & Slyh, R. (2008b. The MIT-LL/AFRL IWSLT-2008 MT System. In *International Workshop on Spoken Language Translation* (pp. 69–76). Hawaii, USA

Skadiņa, I. (2004). Machine Translation for Latvian. In *Proceedings of First Baltic Conference „Human Language Technologies – the Baltic Perspective"* (pp. 102-106). Riga

Skadiņa, I., Vasiļjevs, V., Deksne, D., Skadiņš, R., & Goldberga, L. (2007). Comprehension Assistant for Languages of Baltic States. *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007* (pp. 167-174). Tartu, Estonia

Skadiņa, I., Brālītis, E. (2008). Experimental Statistical Machine Translation System for Latvian, In *Proceedings of the 3rd Baltic Conference on HLT* (pp. 281-286)

Skadiņa, I., Brālītis, E. (2009). English-Latvian SMT: knowledge or data. In *Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA, NEALT Proceedings Series, Vol. 4 (2009)* (pp. 242–245). Odense, Denmark

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., & Gornostay, T. (2010b). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In *Proceedings of 3rd Workshop on Building and Using Comparable Corpora*. BUCC 2010, Valletta, Malta

Skadiņa, I., Auziņa, I., Grūzītis, N., Levāne-Petrova, K., Nešpore, G., Skadiņš, R., & Vasiļjevs, A. (2010a). Language Resources and Technology for the Humanities in Latvia (2004–2010). In *Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219* (pp. 15-22). Riga, Latvia: IOS Press

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation* 14:113-157.

Somers, H. (2003). An overview of EBMT. In Carl, M. , & Way, A. (Eds.), *Recent Advances in Example- Based Machine Translation* (pp. 3-57). Dordrecht, The Netherlands: Kluwer Academic Publishers

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation: LREC'06*

Stymne, S., Holmqvist, M., & Ahrenberg, L. (2008). Effects of Morphological Analysis in Translation between German and English. In *StatMT '08 Proceedings of the Third Workshop on Statistical Machine Translation*

Surcin, S., Lange, E., & Senellart, J. (2007). Rapid development of new language pairs at SYSTRAN. In Maegaard, B. (Ed.) *Proceedings Machine Translation Summit XI* (pp. 443-449). Copenhagen, Denmark

Šics, V. (2010). *Uz Moses bāzes izstrādātas statistiskās mašīntulkošanas sistēmas pielāgošana latviešu valodai.* M.Sc. Thesis. University of Latvia

Thurmair, G. (2004). Comparing rule-based and statistical MT output. In *LREC-2004. Workshop, 25th May 2004: The amazing utility of parallel and comparable corpora* (pp. 5-9). Lisbon

Thurmair, G. (2005). Hybrid Architectures for Machine Translation Systems. *Language Resources and Evaluation*, 39(1)

Thurmair, G. (2006). Using Corpus Information to Improve MT Quality. *Proceedings of the Workshop LR4Trans-III.* Genova: *LREC*

Thurmair, G. (2009). Comparing different architectures of hybrid Machine Translation systems. In *Proceedings of the twelfth Machine Translation Summit* (pp. 340-347). Ottawa, Ontario, Canada

Tiedemann, J., Nygaard, L. (2004). The OPUS corpus - parallel & free. in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (Eds.), *Recent Advances in Natural Language Processing* (vol V) (pp. 237-248). Amsterdam/Philadelphia: John Benjamins

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural language processing* (pp. 1-8).

Tyers, F. M., & Donnelly, K. (2009). Apertium-cy - a collaboratively-developed free RBMT system for Welsh to English. *Prague Bulletin of Mathematical Linguistics,* 91:57–66.

Uchida, H., Zhu, M., & Della Senta, T. (1999). *The UNL, a gift for a millennium*. Tokyo: IAS/UNU

Uchida, H., & Zhu, M. (2001). *The Universal Networking Language beyond Machine Translation*. Seoul, Korea: UNDL Foundation

Vaivads, J. (2010). Semantiski korekta tulkošana kontrolētās dabīgās valodās, izmantojot OWL. M.Sc. Thesis. University of Latvia

Vauquois, B., (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation, *IFIP Congress-68* (pp. 254-260). Edinburgh

Vandeghinste, V., Schuurman, I., Carl, M., Markantonatou, S., & Badia, T. (2006). METIS-II: Machine Translation for Low Resource Languages. In *Fifth International Conference on Language Resources and Evaluation. Proceedings* (pp. 1284-1289). Genoa, Italy

Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M., & Schmidt, P. (2008). Evaluation of a Machine Translation System for Low Resource languages: METIS-II. In *LREC 2008: 6th Language Resources and Evaluation Conference*. Marrakech, Morocco

Vasiļjevs, A., Skadiņš, R., & Skadiņa, I. (2011a). Towards Application of User-Tailored Machine Translation in Localization. In Zhechev, V. (Ed.) *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators" JEC 2011* (pp. 23-31). Luxembourg

Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2011b). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In *Proceedings of the 13th Machine Translation Summit* (pp. 507-511). Xiamen, China

Vasiļjevs, A., Gornostay, T., & Skadiņš, R. (2010). LetsMT! – Online Platform for Sharing Training Data and Building User Tailored Machine Translation. *Proceedings of the*

Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 219 (pp. 133-140). Riga, Latvia: IOS Press

Venugopal, A., Vogel, S., & Waibel, A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL)*.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings* (pp. 697-702). Genoa, Italy

Villavicencio, A., Copestake, A., Waldron, B., & Lambeau, F. (2004). Lexical Encoding of MWEs. In T.Tanaka, A. Villavicencio, F. Bond, A. Korhonen (Eds.), *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*. Barcelona.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., Waibel, A. (2003). The CMU Statistical Machine Translation System. *Proceedings of MT-Summit IX*.

Way, A. (2001). Translating with Examples. In *MT Summit VII: Workshop on Example-Based Machine Translation, Proceedings of the Workshop* (pp. 66–80). Santiago de Compostela, Spain

Weaver, W. (1955). Translation. In Locke, W. N., & Boothe, A.D. (Eds.) *Machine translation of languages* (pp. 15–23). Cambridge, MA: MIT Press. Reprinted from a memorandum written by Weaver in 1949.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3)

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 523–530). Toulouse, France

Yamada, K., & Knight, K. (2002). A Decoder for Syntax-Based Statistical MT. In *Proceedings of the Conference of the Association for Computational Linguistics, ACL'02*.

Zhang, H., & Gildea, D. (2008). Efficient multi-pass decoding for synchronous context free grammars. In *Proceedings of ACL-08: HLT* (pp. 209–217). Columbus, Ohio: Association for Computational Linguistics.

Zhang, Y. (2009). *Structured Language Models for Statistical Machine Translation*. PhD thesis. Carnegie Mellon University

Zollmann, A., & Venugopal, A. (2006). Syntax Augmented Machine Translation via Chart Parsing. In *NAACL 2006 - Workshop on statistical machine translation*. New York.

Zollmann, A., Venugopal, A., Och, F. J., & Ponte, J. (2008). A Systematic Comparison of Phrase-Based, Hierarchical and Syntax-Augmented Statistical MT. In *Proc. of 22nd International Conference on Computational Linguistics (Coling).* Manchester, U.K.

Zwarts, S., & Dras, M. (2008). Choosing the right translation: a syntactically informed classification approach. In *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 1153–1160). Manchester, UK

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)

Горностай, Т., Васильев, А., Скадиньш, Р., Скадиня, И. (2007). Опыт латышско↔русского машинного перевода. *Труды международной конференции «Диалог 2007»* (pp. 137-146). Бекасово

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)