Charles University

Faculty of Mathematics and Physics

Institute of Formal and Applied Linguistics

HABILITATION THESIS

Pavel Pecina

Adaptation of Machine Translation to Specific Domains and Applications

Prague 2016

For my beloved ones

Contents

Preface

This work presents a compilation of our papers from the area of Machine Translation and its adaptation to specific domains and applications which were published during 2011–2016 in proceedings of major conferences and journals. The papers are preceded by an introductory part which puts all the works into context. The work was conducted within three projects funded by the European Union in the 7^{th} Framework Programme and H2020, namely **Panacea** (grant agreement no. 248 064), **Khresmoi** (grant agreement no. 257 528), and **KConnect** (grant agreement no. 644 753), at two research centers: Centre for Next Generation Localisation, Dublin City University, Ireland in 2010–2012 and Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic in 2012–2016.

The selection of publications includes the following conference papers and journal articles:

Pecina et al. (2011): Pavel Pecina, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou. **Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation**. In *Proceedings of the 15^{th} Annual Conference of the European Association for Machine Translation*, pages 297–304, Leuven, Belgium, 2011.

Pecina et al. (2012a): Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Josef van Genabith. **Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A Case Study**. In *Proceedings of the 16^{th} Annual Conference of the European Association for Machine Translation*, pages 145–152, Trento, Italy, 2012.

Pecina et al. (2012b): Pavel Pecina, Antonio Toral, and Josef van Genabith. **Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation**. In *Proceedings of the 24^{th} International Conference on Computational Linguistics*, pages 2209–2224, Mumbai, India, 2012.

Pecina et al. (2015): Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way, and Josef van Genabith. **Domain Adaptation of Statistical Machine Translation with Domain-focused Web Crawling**. In *Language Resources and Evaluation*, 49(1), pp. 147–193, Springer Netherlands, 2015.

Urešová et al. (2014): Zdeňka Urešová, Ondřej Dušek, Jan Hajič, and Pavel Pecina. **Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain**. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3244–3247, Reykjavik, Iceland, 2014.

Dušek et al. (2014): Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, and Daniel Zeman. **Machine Translation of Medical Texts in the Khresmoi Project**. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228, Baltimore, USA, 2014.

Pecina et al. (2014): Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. **Adaptation of Machine Translation for Multilingual Information Retrieval in the Medical Domain**. In *Artificial Intelligence in Medicine, 61 (3), Text Mining and Information Analysis of Health Documents*, pages 165–185, Elsevier, 2014.

Saleh and Pecina (2016c): Shadi Saleh and Pavel Pecina. **Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval**. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. The 7^{th} International Conference of the CLEF Association, Évora, Portugal, Volume 9822 of the series Lecture Notes in Computer Science*, pages 54–66, Springer International Publishing, 2016.

Part I

Introduction

1 Introduction and Motivation

Machine Translation (MT) aims to translate text (or speech) in one language to another language by means of computer software. It is a very active research field of Computational Linguistics which has been studied in many academic and industry labs over the world.

The variety of research problems, methods, and paradigms that have been proposed and investigated over the last few decades is enormous. Most of the major Computational Linguistics and Natural Language Processing conferences have large MT-focused sections (e.g., ACL[1], Coling[2], EACL[3], NAACL[4], EMNLP[5], and IJCNLP[6]) and there are also several conferences where MT is the central topic (e.g., EAMT[7], AMTA[8], and MT Summit). MT is also a very popular task in various public evaluation campaigns and shared tasks (e.g., WMT[9], IWSLP[10]) which actively contribute to the research in this field by providing common evaluation platforms for new methods, systems, and paradigms.

The research in MT has diversified and has been focusing on a wide range of sub-topics which cover various kinds of related problems. One of them, **Machine Translation Adaptation**, represents the main topic of this thesis. It investigates methods to adapt MT systems to specific domains and applications.

1.1 History of machine translation

The history of MT started soon after the development of the first electronic computers in 1940s (Goldstine and Goldstine, 1946) and the beginning of the Cold War in 1947 which created a demand for automated (mechanical) translation between Russian and English in the United States and Soviet Union. The idea of MT based on information theory and code breaking was first introduced by Warren Weaver in his *Memorandum on Translation* in 1949 (see, e.g., Weaver, 1955).

In 1950s, an active research on MT started at several US universities (e.g., MIT in Boston, MA and Georgetown University in Washington, DC). The first systems were based on bilingual dictionaries and ad-hoc rules ensuring correct word order in the output. In 1954, the *Georgetown–IBM Experiment* publicly demonstrated the feasibility of MT on translation from Russian to English (Macdonald, 1954). The presented system was trivial – it exploited very limited vocabulary and simple rules – but stimulated a major increase of funding of MT research mainly in the US, Japan, and Russia. In 1960, IBM provided the US Air Force (USAF) with a technology to translate from Russian into English using a dictionary of 70 thousand words. Later, MT became inspired by promising developments of formal grammar models developed in Computational Linguistics.

In 1964, the so far bright atmosphere in MT changed and the research slowed down. The US National Academy of Sciences formed the *Automatic Language Processing Advisory Committee* to evaluate the progress of MT research (ALPAC, 1966) which concluded that MT quality could not compete with human translation (MT was found to be slower, less accurate,

[1] Association for Computational Linguistics

[2] Computational Linguistics Conference

[3] European Chapter of Association for Computational Linguistic

[4] North American Chapter of the Association for Computational Linguistics

[5] Conference on Empirical Methods in Natural Language Processing

[6] International Joint Conference on Natural Language Processing

[7] Annual Conferences of the European Association for Machine Translation

[8] Conferences of the Association for Machine Translation in the Americas

[9] Workshop/Conference on Statistical Machine Translation

[10] International Workshops on Speech and Language Processing

and more expensive than human translation) and as a consequence, the US funding was substantially reduced. The report, later criticized as narrow, biased, and short-sighted (Slocum, 1985), had a negative impact in other countries too but the research did continue, e.g., in Canada, France, and Germany, and several successful MT projects were realized during the 1970s (for a comprehensive overview see, e.g., Bruderer, 1977).

The first commercial MT company, *Systran*, was founded in 1968. From 1970, their Russian–English system was used by the USAF Foreign Technology Division and during the Apollo-Soyuz project in 1974–1975 by NASA. Since 1976, it had been also used by the Commission of the European Communities to translate their growing amounts of documentation, first, between English and French (Toma, 1977), later, systems for other languages of European Communities were developed too. In 1976, Systran was employed by General Motors in Canada to translate various manuals from English to French, and in 1978 by Xerox to translate their technical documents into six languages (Slocum, 1985). Another successful system, *METEO*, developed at the Université de Montréal, had been used to translate weather forecasts from English into French since 1977 (Thouin, 1982).

In 1980s, other commercial systems were brought on the market. *Logos*, originally employed by USAF to translate military equipment manuals from English to Vietnamese during the war in Vietnam (Sinaiko and Klare, 1973), was later used by several multi-national organizations for German–English and German–French translations (e.g., Nixdorf, Hewlet Packard). *Metal* originated at the University in Texas in 1960s focusing on German–English translation using advanced linguistic methods (such as German analysis based on context-free grammars) and features (Lehmann et al., 1981). In 1980, it was adopted by Siemens AG, further developed (e.g., Thurmair, 1990) and commercialized. *SPANAM* was based on the research conducted at the Georgetown University in 1960s and 1970s and internally developed by the Pan American Health Organization to translate between Spanish and English. Several systems for English–Japanese translation were also developed by Japanese computer companies (e.g., Sharp, NEC, OKI). During 1980s, the research in MT focused on more advanced methods usually based on indirect (interlingual) transfer using linguistic analysis and synthesis at various levels. The most notable research projects of the period include *SUSY* (Maas, 1987), *Mu* (Tsujii, 1987), *GETA/Ariane* (Guilbaud, 1987; Boitet, 1989), *Eurotra* (King, 1982), *Rosetta* (Appelo and Landsbergen, 1986), and *DLT* (Sadler, 1989).

In 1990s, the trends initiated in the previous era continued. MT technologies were used in large organizations for translation of in-house-created documents. Software localization industry working in the area of adaptation and translation of computer software and documentation for new markets became one of the major consumers of MT technologies. The growth of the market with personal computers increased the sales of MT software for personal use and soon after the development of Internet technologies, MT became provided as an on-line service for masses (e.g., *Altavista Babelfish* using the Systran technology, launched in 1997). During this decade, the MT methods started to shift from rule-based to data-driven which were based on exploitation of large parallel corpora (collections of texts and their translations). Those included example-based methods (Nagao, 1984) built on the idea of translation by analogy (reusing previously translated phrases extracted from a parallel corpus) and statistical methods exploiting statistical models with parameters also estimated on parallel corpora. The first fully statistical MT system *Candide* was developed in IBM (Berger et al., 1994). The MT research focused also on speech translation which integrates speech recognition, translation, and speech synthesis, e.g., the *Verbmobil* project funded by the German government (Kay et al., 1992).

Since 2000, the research has predominantly focused on the data driven methods, especially **Statistical Machine Translation** covering various approaches differing in the level of linguistic information used (ranging from word-based, phrase-based, to syntax-bases). The IBM word-based models were implemented in *GIZA++* (Och and Ney, 2003), the phrase-based translation in *Pharaoh* (Koehn, 2004a) and later in *Moses* (Koehn et al., 2007), currently the state-of-the-art system which implements several different approaches also. The most influential commercial MT projects include: *Google Translate* and *Microsoft Bing Translator*, both providing translation as a free service on the Internet.

1.2 Machine translation applications

The ultimate application of MT (and its long-term goal) is to take over the work of human translators and produce **high quality translations**. The quality of MT has improved a lot over the years, but due to the complexity of the task, MT still does not meet the quality requirements put on professional translation. In many areas, such as legal documents or drug information, where potential mistakes can seriously affect businesses or health, high translation quality is absolutely essential and MT can not (yet) replace human-produced translation.

On the other hand, there are many areas and situations where the translation quality is not that critical and MT can be successfully applied. One of them is on-line translation of web pages in a language the user does not understand. In that case, it is not necessary to obtain a perfect translation of the entire text, as long as the user finds what they searches for. This task is often called **gisting**.

Despite its imperfection, MT has been successfully applied in professional tools for **Computer-Assisted Translation** (CAT). Such tools support and facilitate the translation process performed by a human translator. Traditionally, the CAT tools were based on databases of previously translated texts (translation memories) which, for a given input text, provided a possible (rough) translation extracted from the database which must be edited by a human to correct errors and improve the overall quality. Modern CAT tools provide the option to use MT as an alternative to translation memories.

Another area of MT applications includes tasks where MT is not directly consumed by humans, rather it is fed into a subsequent algorithm solving a more complex task. One example is **Speech Translation** where both the input and output are in a spoken form but in different languages. This complex task includes speech recognition to convert the input speech in one language into a sequence of words, machine translation to translate the string to the other language, and speech synthesis to generate a spoken form of the translation. The overall quality of the output, of course, depends on the quality of the three steps and any imperfection in MT is immediately apparent in the spoken output.

Another example, which is highly relevant to the work presented in this thesis, is **Cross-Lingual Information Retrieval** (CLIR), a subfield of information retrieval where the retrieved information is in a language different from the language of a user's query. For instance, the query may be posed in English but retrieved documents written in French. Here, MT-based techniques is used to map the query and/or documents into a common representation space, typically one of the languages (either the query language or the document language). Often, the MT output is not visible to the users, which has two advantages. First, the users are not disappointed and distracted by (eventual) imperfect translation and, second, the translation can be represented in a non-human-readable form, better suitable for the algorithms solving the subsequent task (information retrieval).

1.3 Structure of the thesis

This thesis is devoted to adaptation of machine translation to specific domains and Applications. It consists of two parts. Part I presents the background and context for Part II which contains a compilation of our selected publications from the area of MT and the topics of this thesis. Section 1 in Part I introduces the research area, presents important milestones from its history, and provide an overview of its practical applications. Section 2.2 is devoted to Statistical Machine Translation (SMT), the prevailing paradigm of current MT approaches, and reviews the fundamental principles and state-of-the art methods in this field which is used in the work presented in this thesis. Section 2.4 is focused on the task of SMT adaptation. It presents on overview of methods to adapt an existing SMT system to specific domains and applications with a specific focus on the domain of medicine, which plays an important role in our experiments. Section 3 then reviews the papers presented in Part II, summarizes their content and main findings and puts them into context of each other. Section 4 concludes Part I and provides some final remarks summarizing our work in this area.

The publications presented in Part II of this thesis are each a joint work of several people. For all the papers, the author of this thesis is the main contributor to the presented work, either as the main/first author of the publication or the leader/supervisor of the team that co-authored the paper.

2 Background and Related Work

In this section, a theoretical background and an overview of work related to the papers presented later in this thesis is provided. We first describe the basic principles of Statistical Machine Translation (SMT) with the main focus on phrase-based Statistical Machine Translation. Then, we introduce the area of adaptation of SMT and acquisition of domain-specific data which plays an essential role in most of the domain-adaptation techniques. Finally, we introduce the area of cross-lingual information retrieval, a specific application which exploits SMT for translation of user search queries. In this section, a special attention is devoted to the domain of medicine which is central in many of our experiments presented later.

2.1 Machine translation

Methods of Machine Translation can be structured into several basic groups:

Rule-based (RBMT) methods exploit a set of rules and dictionaries manually created by language experts to map grammatical structures and lexical items from one language to another. Most of the early MT systems were based on this approach (e.g. Toma, 1977). The manual labor involved in development of an RBMT system is quite substantial with a significant impact on translation quality of the output. Rule-based systems are found to be especially effective in very limited domains (e.g. weather forecast, Thouin, 1982).

Example-based (EBMT) methods introduced by Nagao (1984) rely on a different type of resources. Instead of applying the translation principles encoded in the rules and dictionaries, they extract this knowledge from existing translations stored in a form of parallel texts. The input is decomposed to phrases which are then *translated by analogy* to previous translations found in the parallel texts. An overview of this area can be found, e.g., in Carl et al. (2004).

Statistical (SMT) methods also employ knowledge extracted from parallel texts, here in a form of proper statistical models, which are directly used to generate the translations. This paradigm is based on the idea of code breaking by Weaver (1955) motivated by Shannon's information theory (Shannon, 1948). Nowadays, it is the most widely studied approach to machine translation (Koehn, 2010).

Hybrid methods are based on a combination of multiple approaches, typically rule-based and statistical. Statistics is either used to smooth the output of a rule-based system (e.g., by applying a language model), or rules are used to preprocess SMT input to better fit the translation process, or rules are used to postprocess the SMT output (e.g., by improving grammatical agreement, Rosa, 2014).

Neural network (NMT) methods have been introduced to the MT area only recently (Cho et al., 2014) but brought a radical change and diversification in MT research. These methods employ (large) neural networks trained directly to produce translated texts. Several configurations have been proposed and evaluated so far, mostly based on two recurrent neural networks (one for encoding a source sentence into a fixed-length vector representation and another network for decoding this representation into the target sentence), and in some cases already achieve results comparable to the state of the art (Sutskever et al., 2014).

The research in the area of MT focuses also on another task, such as system combination (Du and Way, 2009), exploiting comparable corpora (Munteanu and Marcu, 2002), incorporating linguistic information (Axelrod, 2006), and many others.

2.2 Statistical machine translation

Formally, SMT translates text according to the probability distribution $p(\mathbf{e}|\mathbf{f})$, where $\mathbf{f}$ is a input sentence of l_f words (tokens) in the source language ($\mathbf{f} = f_1^{l_f}$) and $\mathbf{e}$ is its output translation into the target language ($\mathbf{e} = e_1^{l_e}$).

The two principal problems in this approach are: **modeling** – how to model and estimate the probability distribution $p(\mathbf{e}|\mathbf{f})$, and **decoding** – how to find the optimal translation with the highest probability. Formally, the decoding problem is defined in the following way:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{e}^*} p(\mathbf{e}|\mathbf{f}), \tag{1}$$

where $\hat{\mathbf{e}}$ is the optimal translation and $\mathbf{e}^*$ is the set of all sentences in the target language. Naturally, a complete search through this set is intractable and must be solved by a heuristics limiting the search space. Modelling $p(\mathbf{e}|\mathbf{f})$ can be approached in various ways. The standard way is the (generative) noisy channel model which applies the Bayes Theorem and decompose the conditional probability distribution into two independent components: **translation model** (TM) and **language model** (LM):

$$p(\mathbf{e}|\mathbf{f}) = \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})} \propto p(\mathbf{f}|\mathbf{e})p(\mathbf{e}), \tag{2}$$

where $p(\mathbf{f}|\mathbf{e})$ is the probability of the source string given the target string (translation model) and $p(\mathbf{e})$ is the probability of the target sentence irrespectively of the source sentence (language model). $p(\mathbf{f})$ is constant and does not affect the maximisation. The decoding is then converted into the following form:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{e}^*} p(\mathbf{f}|\mathbf{e})p(\mathbf{e}), \tag{3}$$

The above transformation, however, does not alleviate the problems of modeling and decoding: the translation search space does not change at all and instead of one conditional distribution it is needed to estimate virtually the same (but inverse) translation model and a language model. However, this decomposition allows to control two fundamental aspects of translation: the translation model controls **adequacy** (correct lexical choice) and the language model controls **fluency** (correct grammar) of the output. Moreover, language models have been well studied and explored in other related areas (e.g., speech recognition, Jelinek, 1997) and plenty of methods can be used "out-of-the-box" (e.g., n-gram models and smoothing, Chen and Goodman, 1996). Using the chain rule, the language model is defined as:

$$p(\mathbf{e}) = p(e_1 ... e_{l_e}) = \prod_{i=1}^{l_e} p(e_i|e_1 ... e_{i-1}), \tag{4}$$

where l_e is the length of the sentence $\mathbf{e}$ and $p(e_i|e_1 ... e_{i-1})$ is probability of single words conditioned on their predecessors.

Estimating translation probability distributions for complete sentences is not feasible (no corpus can include all possible sentences). Therefore, the problem is decomposed and sentences broken down into smaller parts, which occur more frequently and their probabilities can be more reliably estimated. This can be approached in various ways: ranging from the trivial word-based models (Och and Ney, 2003), through the most commonly used phrase-based models (Koehn et al., 2003), to more linguistically-motivated hierarchical and syntax-based models (Chiang, 2005).

Word-based SMT

Word-based models of SMT were proposed by Brown et al. (1988) at IBM. They decompose sentences into words and assume that one input word is typically translated as a single output word; less frequently it can produce multiple words in the output or can be dropped out. The IBM "model" is, in fact, a series of models with increasing complexity, starting with a simple model only based on word (lexical) translation probability and further adding models for word reordering. The original IBM Models 1–5 are described in Brown et al. (1993). The most simple **Model 1** defines the following probability distribution:

$$p(\mathbf{f}, a|\mathbf{e}) = \frac{\epsilon}{(l_e + 1)^{l_f}} \prod_{j=1}^{l_f} t(f_j|e_{a(j)}), \tag{5}$$

where **f** is again the source sentence, **e** is the target sentence, a is the alignment function mapping (reordering) positions in the source sentence into the positions in the target sentence, l_f is the source sentence length, l_e is the target sentence length, ϵ is the normalization constant, and finally $t(f_j|e_i)$ is a probability distribution of lexical translations between the source and target language words on positions j and $i = a(j)$, respectively. The translation probability $p(\mathbf{f}|\mathbf{e})$ then considers all possible alignments:

$$p(\mathbf{f}|\mathbf{e}) = \sum_{a} p(\mathbf{f}, a|\mathbf{e}), \tag{6}$$

which are for simplicity assumed to be equally possible so the probability $p(\mathbf{f}|\mathbf{e})$ is defined as:

$$p(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l_e + 1)^{l_f}} \prod_{j=1}^{l_f} \sum_{i=1}^{l_e} t(f_j|e_i). \tag{7}$$

Model 2 extends Model 1 by adding an explicit model for word alignment $q(i|j, l_f, l_e)$:

$$p(\mathbf{f}, a|\mathbf{e}) = \epsilon \prod_{j=1}^{l_f} t(f_j|e_{a(j)}) q(a(j)|j, l_f, l_e), \tag{8}$$

and the final translation model $p(\mathbf{f}|\mathbf{e})$ is then formulated as:

$$p(\mathbf{f}|\mathbf{e}) = \epsilon \prod_{j=1}^{l_f} \sum_{i=1}^{l_e} t(f_j|e_i) q(i|j, l_f, l_e). \tag{9}$$

The further IBM models gradually extend the previous ones. **Model 3** adds a *fertility* model for allowing one-to-many translations and insertions of words, **Model 4** adds a relative alignment model (also called *reordering* or *distortion* model) and **Model 5** solves the problem when multiple words can appear on the same position (this is, in reality, not possible but was allowed in the previous models).

The IBM word-based models are **trained** from sentence-aligned parallel texts (sentences paired with their translations) by the Expectation Maximization algorithm with alignment as a hidden variable. The five models were implemented in a tool called GIZA (Al-onaizan et al., 1999) and later in GIZA++ (Och and Ney, 2003) – a open-source toolkit broadly used even nowadays for finding word-alignments in parallel texts, which is also useful in more complex phrase-based SMT. GIZA++ also implements Model 6, which combines Model 4 and a HMM alignment model in a log-linear way. Several other enhancements of the IBM models were proposed, such as the HMM model with relative distortion but not fertility (Vogel et al., 1996).

Decoding of word-based translation was implemented, e.g., in the ISI ReWrite Decoder (Germann et al., 2001; Germann, 2003) which employs IBM Model 3 and the CMU-Cambridge Statistical Language Modeling toolkit (Clarkson and Rosenfeld, 1997). The word-based models were, however, soon surpassed by more advanced approaches based on phrase-based SMT.

Phrase-based SMT

The key assumption in word-based models that words can be translated one by one is too simplifying in practice. Better candidates for units which can be translated one at a time are sequences of words. This is the approach adopted in phrase-based models proposed by F.J. Och (Och and Weber, 1998; Och et al., 1999) and well described in Koehn et al. (2003).

In phrase-based SMT, an input sentence **f** is segmented into I sequences of consecutive words $\bar{e}_i$ ($\mathbf{e} = \bar{e}_1^I$), called phrases (not necessarily linguistically adequate phrases). Each phrase $\bar{e}_i$ is then translated into a target-language phrase $\bar{f}_i$ which may be reordered with the other translated phrases to produce an output $\mathbf{f} = \bar{f}_1^I$. The original phrase-based model for $p(\mathbf{f}|\mathbf{e})$ is defined as:

$$p(\mathbf{f}|\mathbf{e}) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(\mathit{start}_i - \mathit{end}_{i-1} - 1), \tag{10}$$

where $\phi(\bar{f}_i|\bar{e}_i)$ is the **phrase translation model** estimated from relative counts of phrase pairs detected in sentence-aligned parallel texts by word-alignment (obtained, e.g., by GIZA++), alignment symmetrization, and phrase extraction algorithms (Och and Ney, 2003). $d(x)$ is the **distortion (reordering) model** based on relative distance: *start*$_i$ is the position of the first word of the source phrase that translates to the i-th target phrase and *end*$_i$ is the position of the last word of that target phrase. The reordering distance is then given as $x = \mathit{start}_i - \mathit{end}_{i-1} - 1$. The distortion model models probability distribution of that distance. It is not estimated from data but rather set as an exponential cost function $d(x) = \alpha^{|x|}$ with the parameter $\alpha \in [0, 1]$. Together with the language model, the best translation is selected based on three models:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}\in\mathbf{e}^*} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) \prod_{i=1}^{I} d(\mathit{start}_i - \mathit{end}_{i-1} - 1) \prod_{i=1}^{l_e} p(e_i|e_1...e_{i-1}) \tag{11}$$

Och and Ney (2002) reformulated the original phrase-based model as **log-linear model** which allows addition of new components (feature functions) and their weighting (by exponential parameters). The combination of the set of feature functions h_i and their weights λ_i is defined as:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}\in\mathbf{e}^*} \prod_{i=1}^{n} h_i(\mathbf{e},\mathbf{f})^{\lambda_i} = \arg\max_{\mathbf{e}\in\mathbf{e}^*} \sum_{i=1}^{n} \lambda_i \log h_i(\mathbf{e},\mathbf{f}). \tag{12}$$

The most popular implementation of this approach is called **Moses** (Koehn et al., 2007), also used in the experiments presented in this thesis. It includes the following feature functions:

- **reordering (distortion) model** (h_1–h_7) allowing the reordering of phrases in the input sentences (e.g., distance-based and lexicalized reordering),
- **language model** (h_8) ensuring that the translations are fluent,
- **phrase translation model** (h_9–h_{12}) ensuring that the source and target phrases are good translations of each other (e.g., direct and inverse phrase translation probability, direct and indirect lexical weighting, and phrase penalty),
- **phrase penalty** (h_{13}) controlling the number of phrases the translation consists of,
- **word penalty** (h_{14}) preventing the translations from being too long or too short.

The parameters (λ_i) of the log-linear model have a significant influence on the overall translation quality and require **tuning**. However, the optimal setting depends on the language pair and data used to train the model components. A common solution to optimize weights of the log-linear combination is *Minimum Error Rate Training* (MERT) proposed by Och (2003). It automatically searches for the values that minimize a given error measure (or maximize a given translation quality measure) on a development set of parallel sentences. Theoretically, any automatic measure can be used for this purpose. The most commonly used one is BLEU (*Bilingual Evaluation Understudy*, Papineni et al., 2002). In our work, we also experiment with PER (*Phrase Error Rate*, Tillmann et al., 1997). The search algorithm is a specific type of coordinate ascent method. It considers the n-best translation hypotheses for each input sentence, updates the feature weight which is most likely to improve the objective and iterates until convergence. Since the error surface is usually highly non-convex and the algorithm cannot explore the whole parameter space, it can converge to a local maximum. However, in practice, the obtained results are usually good (Bertoldi et al., 2009).

The entire search space for **decoding** in phrase-based SMT includes hypotheses formed by all possible segmentations of a source sentence and all possible translations of each phrase in these segmentations. For longer sentences, an exhaustive search through this space is not feasible and heuristic approaches which prune the search space must be applied, e.g., the greedy procedure employed by Germann et al. (2001) and Marcu and Wong (2002). The approach implemented in Moses is based on the beam-search algorithm originally proposed by Jelinek (1997) for decoding in speech recognition. The algorithm generates (partial) translation hypotheses from left to right by exploring the space represented as a graph by expanding the most promising nodes only (the cost is calculated for the already translated part of the input sentence plus an estimation of the future cost of the untranslated part). This procedure also allows to generate multiple best hypotheses (*n-best lists*).

Syntax-based SMT

Despite the word “phrase” in the title, the phrase-based SMT does not rely on linguistically defined phrases nor does it exploit any linguistic knowledge (except for information on word boundaries). The translation unit in phrase-based models is a phrase, treated as a flat sequence of words, translated into the target language also as a word sequence with no structure. Linguistically oriented approaches motivated by recursive structure (syntax) of language have been adopted in several SMT paradigms: hierarchical phrase-based SMT (Chiang, 2005, 2007) based on Synchronous Context-Free Grammar (SCFG), treelet-based SMT (Quirk and Menezes, 2006) with translations units as partial dependency trees, or deep-syntactic dependency trees (Žabokrtský et al., 2008). Hierarchical phrase-based models based on SCFG are also implemented in Moses (Koehn et al., 2007). Syntax-based models, however, do not usually outperform the phrase-based models.

2.3 Evaluation of machine translation

The problem of assessing the quality of MT output is crucial for measuring the quality of an individual system and for comparing the quality of multiple systems. MT evaluation can be done manually or automatically.

In **manual evaluation**, human annotators judge translation quality by scoring or ranking an MT output. This is a time consuming and expensive process, often suffering from low annotator agreement due to the subjectivity of the task. Manual MT evaluation campaigns have been organized within the WMT workshop series (e.g., Bojar et al., 2013, 2014, 2015).

In **automatic evaluation**, the MT output is automatically compared to a set of reference translations with the idea that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002). However, the quality of translation is highly subjective and automatic evaluation can penalize translations which are completely valid but differ from the reference. This can be addressed by providing multiple references for each test sentence, but in practice the variety of valid translations can be very high. Another way to diminish this problem is to increase the number of test cases and thus decrease the probability of penalizing valid translations. MT evaluation sets should therefore include several thousands of test sentences.

A number of automatic **evaluation metrics** have been proposed and evaluated with respect to correlation with human judgments, e.g., within the Metrics Task of the WMT and MetricsMATR workshop series (Callison-Burch et al., 2010; Macháček and Bojar, 2013, 2014; Stanojević et al., 2015). All of them are based on measuring some kind of similarity or overlap between the MT output and reference translations. The most commonly used ones include: PER (Phrase Error Rate, Tillmann et al., 1997) and TER (Translation Error Rate, Snover et al., 2006a) based on string edit distance, BLEU (Bilingual Evaluation Understudy, Papineni et al., 2002) and NIST (Doddington, 2002) based on exact matching of n-grams, and METEOR (Metric for Evaluation of Translation with Explicit ORdering, Banerjee and Lavie, 2005) and TERp (Snover et al., 2008) based on matching of synonyms and short paraphrases. In this work, due to the space limitation, BLEU is used as the main evaluation measure, which is broadly accepted as the standard in this field. Other measures are reported in the referred papers to provide a more detailed and complex view on the results.

2.4 Adaptation of statistical machine translation

Domain adaptation is one of the very active research field within SMT. As in any other machine learning tasks, the quality of SMT output strongly depends on training data. However, not only the quantity but also the quality of parallel and monolingual training data is important for development of an SMT system. Unless the system is trained on data of the same nature (distribution) as the test data, it is not guaranteed to perform optimally and produce good translations. The most extensive and publicly available resources of SMT training data include, for instance, parallel parliamentary proceedings (Europarl, Koehn, 2005 or Hansard, Roukos et al., 1995), legislation documents (the JRC corpus, Steinberger et al., 2006), or news stories (the Project Syndicate[11], Callison-Burch et al., 2012), which typically cover a wide range of various different topics and are typically understood as general-domain data. Training resources for specific domains (in-domain data) are usually much scarcer or often not available at all. Therefore, special domain adaptation techniques are applied to adapt an SMT system trained on general-domain data to improve translation of text within a specific domain.

Three main research directions can be identified in SMT domain-adaptation depending on the availability of domain-specific data. First, if any in-domain data is available, it can be directly used to improve the SMT system by combining the in-domain with out-of-domain resources for training. Second, if in-domain data exists but is not readily available, one may attempt to acquire domain-specific data from available domain-specific sources (e.g., from comparable corpora). Third, if in-domain data sources cannot be identified, one may attempt to extract pieces of in-domain data from larger general-domain (or mixed-domain) sources.

The initial attempts to perform domain adaptation were based on exploitation of existing in-domain parallel and monolingual data. The first such experiments were probably carried out by Langlais (2002), who integrated in-domain lexicons in the translation model. Wu and

[11] http://www.project-syndicate.org/

Wang (2004) then used in-domain data to improve word alignment in the training phase. Other work focused on mixture modeling where separate models were trained for individual data sets (e.g., in-domain and out-of-domain) and interpolated. This technique has been applied to language models (Koehn and Schroeder, 2007) as well as translation models (Nakov, 2008; Sanchis-Trilles and Casacuberta, 2010; Bisazza et al., 2011). The different models can be combined by linear or log-linear interpolation (Foster and Kuhn, 2007; Banerjee et al., 2011). The interpolation parameters can be optimized, e.g., by minimization of the model perplexity on a development set (Sennrich, 2012) or maximization of an evaluation metric (Haddow, 2013).

Several methods have been proposed for the situation when existing amounts of in-domain SMT training data are not sufficient or not existing at all. For example, (Munteanu and Marcu, 2005; Tanaka, 2002; Hewavitharana and Vogel, 2013) mined in-domain sentence pairs from **comparable corpora** (texts that are not strictly parallel but on the same topic). Daumé III and Jagarlamudi (2011) used a domain-specific comparable corpora to extract translation of Out-Of-Vocabulary (OOV) terms to reduce their rate. Dong et al. (2015) extracted parallel lexicons (dictionaries) from comparable corpora. In the work presented in this thesis, in-domain training data (parallel and monolingual) is mined from the web and used for parameter optimization to improve language and translation models (Pecina et al., 2011, 2012a). An interesting idea was explored by Bertoldi and Federico (2009), who created synthetic parallel training data from in-domain monolingual data.

The selection of **pseudo in-domain data** is another technique to obtain training data for domain-adaptation. It is based on the idea that a sufficiently broad general-domain (mixed-domain) corpus will include sentences that resemble the target domain. Eck et al. (2004b) presented such a technique for adapting the language model. Hildebrand et al. (2005) extended this approach to the translation model. Foster et al. (2010) weighted phrase pairs from out-of-domain corpora according to their relevance to the target domain. Moore and Lewis (2010) used difference of cross-entropy given an in-domain model and general-domain model to filter monolingual data for language modeling. Axelrod et al. (2011) applied this idea to filter parallel training data. Banerjee et al. (2013) extended the cross-entropy approach by combining this score with scores based on quality estimation. Toral (2013) then exploited linguistic units (lemmas and part-of-speech) instead of surface forms to perform the selection.

The 2012 JHU Summer Workshop (Carpuat et al., 2012) focused on the issues in domain-specific MT. They studied the use of phrase-sense disambiguation to model domain content in domain-specific SMT and found that it can successfully model lexical choice across domains.

Genre adaptation is related to domain adaptation. While domain adaptation mainly deals with the problem of lexical coverage (lack of domain-specific terminology), genre adaptation is concerned with changes in syntax and style, which have become very common and diverse in modern means of communication, such as SMS messages, Internet chats, discussion forums, and social network communication (e.g., unusual sentence length, ungrammatical constructions, missing punctuation, letter casing). Some recent work in this area has focused on SMT adaptation to genres such as patents and patent applications (Ceausu et al., 2011), short text messages (Callison-Burch et al., 2011), user-generated forum content (Banerjee et al., 2012), public conference talks (Bisazza and Federico, 2012), and movie subtitles (Fishel et al., 2012). The methods used in those works are generally similar to domain adaptation techniques.

2.4.1 SMT in the medical domain

From the MT point of view, medicine is a resource-rich domain, both in terms of available texts (document collections) and terminology lexicons (code sets, classifications), and MT has been applied to medical texts a number of times. We review some papers published recently.

Eck et al. (2004a) trained an SMT system for the translation of dialogues between doctors and patients and showed that a dictionary extracted from the Unified Medical Language System (UMLS) Metathesaurus and its semantic type classification (U.S. National Library of Medicine, 2009) can significantly improve translation quality from Spanish to English (measured by standard automatic evaluation metrics BLEU and NIST). Wu et al. (2011) analyzed MT quality of their SMT system and Google Translate applied to PubMed[12] document titles and studied whether it was sufficient for patients. The conclusions were very positive especially for languages with large training resources (English, Spanish, German). They manually evaluated the fluency and adequacy and the average scores were above four on a five-point scale. In automatic evaluation, their systems substantially outperform Google Translate. The findings of Costa-jussà et al. (2012) were not so positive regarding the quality of SMT in the medical domain. They analyzed and evaluated the quality of public web-based MT systems (such as Google Translate) and concluded that in both automatic and manual evaluation (reported for the total of 7 language pairs), the performance of these systems was still not good enough to be used in daily routines of medical doctors in hospitals. Jimeno Yepes et al. (2013) proposed a method for obtaining in-domain parallel corpora by extracting titles and abstracts of publications from the MEDLINE[13] database. The acquired data contained from 30,000 to 130,000 sentence pairs (depending on the language pair) and was used as additional training data for SMT training which significantly improved the translation quality compared to a baseline trained without these resources.

2.5 Domain-specific data acquisition

As discussed in the previous section, many of the domain adaptation techniques rely on availability of domain-specific data, parallel for training the translation model and monolingual for training the language model. Such data is often very scarce or not available at all. A unique source of such resources for many domains is the web. Its content is often publicly available to download and offers a great source of data (text). In this work, we are mainly interested in acquisition of monolingual and parallel data in specific domains. In this section, we review automatic acquisition methods of such resources.

Domain-focused web crawling for monolingual texts

Web crawling is usually defined as an automatic and repetitive process of traveling through the World Wide Web by extracting links from already fetched web pages and adding them to the list of pages to be visited. The initializations (setting seed URLs) and selection of the next link to be followed is a key challenge for the evolution of the crawl and is tied to the goal of the process. A crawler that aims to build domain-specific web collections (Qin and Chen, 2005) must prioritize which pages to visit (to discover domain-specific texts). Several generic algorithms have been exploited for selecting the most promising links. The Best-First algorithm (Cho et al., 1998) sorted the links with respect to their relevance scores and selects a predefined amount of them as the seeds for the next crawling cycle. The PageRank (Brin and Page, 1998) algorithm exploits the "popularity" of a web page, i.e., the probability that a random crawler will visit that page at any given time, instead of its relevance. Dziwiński and Rutkowska (2008); Gao et al. (2010) conditioned the selection of the next links to follow by the distance between relevant pages (i.e., the number of links the crawler must follow in order to visit a particular page starting from another relevant page). A general framework

[12] http://www.ncbi.nlm.nih.gov/pubmed/

[13] http://www.nlm.nih.gov/pubs/factsheets/medline.html

which defines crawling tasks of variable difficulty and fairly evaluates focused crawling algorithms under a number of performance metrics (precision and recall, relevance, algorithmic efficiency, etc.) was proposed by Srinivasan et al. (2005).

The basic assumption in domain-focused web crawling is that relevant pages are more likely to contain links to more pages in the same domain. Classification of web pages as relevant or not relevant is then the key to discover domain-relevant material. Qi and Davison (2009) reviewed various features and algorithms used previously to solve this task. Most of the reviewed algorithms apply supervised machine-learning methods on feature vectors consisting of on-page features, such as textual content and HTML tags (Yu et al., 2004). Many algorithms exploit additional information contained in web pages, including anchor text of hyperlinks. Some methods adopt the assumption that neighboring pages are likely to be in the same domain (Menczer, 2005).

A crucial step in producing good-quality language resources from the web is **boilerplate removal**, e.i. removal of parts of the web page which are of only limited or no value (Kilgarriff and Grefenstette, 2003). Boilerplate usually includes navigation links, advertisements, disclaimers, repeated headers and footers, etc. Several methods have been employed for this task. A review of cleaning methods is presented, e.g., in Spousta et al. (2008). Our own approach, based on sequence labeling with Conditional Random Fields (Marek et al., 2007), placed first in the Cleaneval competition organized in 2007 in Belgium[14].

To give a few examples of existing tools, one could mention, e.g., the WebBootCat toolkit (Baroni et al., 2006) which harvests domain-specific data from the web by querying search engines with tuples of in-domain terms and Combine[15] – an open-source focused crawler based on a combination of a general web crawler and a topic classifier.

Domain-focused web crawling for parallel texts

Parallel text acquisition from the web is even more challenging than crawling for monolingual data. Despite the fact that many websites are nowadays multilingual, it is difficult to discover such pages in an automatic fashion and mine parallel texts from their content. Several complete systems have been developed so far.

Systems such as PTMiner (Nie et al., 1999) and WeBiText (Désilets et al., 2008) exploited structural similarity of website pages and filtered the fetched web pages by keeping only those containing language markers in their URLs. Resnik and Smith (2003) presented the STRAND system where a search engine was used to search for multilingual websites and potentially parallel pages were identified based the similarity of the HTML structures of the fetched web pages. Parallel Text Identification System developed by Chen et al. (2004) incorporated a content analysis module using a predefined bilingual wordlist. Similarly, Zhang et al. (2006) adopted a naive aligner in order to estimate the content similarity of candidate parallel web pages. Bitextor developed by Esplà-Gomis and Forcada (2010) combined language identification with shallow features (file size, text length, tag structure, and list of numbers in a web page) to mine parallel pages from multilingual sites that have been already been stored locally with the HTTrack[16] website copier. Barbosa et al. (2012) crawled the web and examined the HTML DOM tree of visited web pages with the purpose of detecting multilingual websites based on the collation of links that are very likely to point to in-site pages in different languages. Once a multilingual site is detected, they use an intra-site crawler and alignment procedures to harvest parallel text for multiple pairs of languages.

[14] http://cleaneval.sigwac.org.uk/

[15] http://combine.it.lth.se/

[16] http://www.httrack.com/

2.6 Cross-lingual information retrieval

Information Retrieval (IR) is the task of finding material that satisfies an information need (Manning et al., 2008). Typically, the material is in a form of textual documents in one language (e.g., web pages), the collections are large (e.g., web-scale), and the information need is specified as a written query. This task has been studied for several decades now. The first attempts to design "auto-indexing" machines appeared in 1950's (Mooers, 1950), but the task has evolved into a broad research field dealing with a range of tasks and problems. **Cross-Lingual Information Retrieval** (CLIR) is a subfield of IR, where the documents are in a language different from the language of the user's query.

This "incompatibility" of languages is generally dealt with by either translating queries into the language of the documents, or translating the documents into the language of queries. An alternative approach is to translate both queries and documents into a common semantic representation which is language-independent (e.g., Ruiz et al., 1999; Blei et al., 2003). Our work presented in this thesis follows the option of translating queries rather than documents. For thorough and recent overviews of the research in this field we refer to Nie (2010), Peters et al. (2012), or Zhou et al. (2012).

Over the years, the translation methods for CLIR moved from dictionary-based techniques employing machine-readable bilingual dictionaries to map the query language onto the language of documents (e.g. Ballesteros and Croft, 1998; Maeda et al., 2000; Gao and Nie, 2006), over methods trying to mine query translations from parallel or comparable corpora (Nie et al., 1999; Talvensaari et al., 2007; Azarbonyad et al., 2012), to approaches based on proper SMT models (Zhou et al., 2012, pp. 23–24), often relying on third-party solutions or public on-line services.

The major issues in query translation are ambiguity (in both the source language and the target language) and low coverage (Zhou et al., 2012). To alleviate the problem of ambiguity, Pirkola (1998) proposed *structured query translation* with a synonym operator to group the translation alternatives for individual words. Darwish and Oard (2003) extended his work by weighting the translation candidates by translation probabilities. Federico and Bertoldi (2002) employed a query-translation model based on a Hidden Markov Model and a language-model based query-document model within a single statistical framework. Integration of the two models is ensured over the weighted n-best list of possible query translations.

Low coverage (i.e., handling OOV words during translation) has been addressed, e.g., by stemming (Oard et al., 2001), which is a standard method used in MT as well as in IR to effectively cluster words by removing their inflectional and derivational affixes, or lemmatization, which substitutes word forms by their canonical variants (see comparison in Hollink et al., 2004). Another approach is based on **query expansion**, which enriches the query with synonymous or related expressions. It can be easily achieved by a widely-used approach known as **pseudo-relevance** (Attar and Fraenkel, 1977), where the query in the source language is used to retrieve the top-ranked documents from the collection in the same language and highly weighted terms extracted from these documents (which are assumed to be related) are added to the original query. This technique was applied to CLIR, e.g., by Ballesteros and Croft (1997), where the expansion was performed also on the target side (which aimed at mitigating the effects caused by picking wrong translation alternatives).

Magdy and Jones (2011) proposed an interesting application of **stopword removal**, a technique commonly used in IR but rarely performed in MT. Prepositions, articles, pronouns, conjunctions, and other similar words are typically not indexed in IR and thus can be discarded in the queries. They removed stopwords from the MT training data (and performed stemming) and showed significant speed-up of the translation process in their CLIR setup.

IR and CLIR in the medical domain

The CLIR experiments presented in this thesis were conducted on data from the medical domain. Most previous works in this area take the advantage of the existence of the UMLS Metathesaurus U.S. National Library of Medicine (2009) of medical terminology, which is used as the main source of cross-lingual medical knowledge.

The earliest work (Eichmann et al., 1998) employed UMLS to translate queries in Spanish and French into English. Their MT approach was very trivial, based on full or partial phrase match, dictionary-based look-up, and adding the source language query terms. Pirkola (1998) evaluated his structured queries on data from the medical domain too. He built a Finnish–English health dictionary containing more than 60 thousand entries and showed that CLIR systems based on dictionary-based translation could achieve the performances of a monolingual system. Volk et al. (2002) matched UMLS terms and their semantic relations in queries and documents and reported a performance improvement in monolingual and cross-lingual setting. Rosemblat et al. (2003) used medical queries from the Clinical Trials website[17] to compare two main approaches in CLIR (query translation and document translation). The reported results favored the former approach. Tran et al. (2004) compared a simple UMLS-based translation and hybrid translation combining pattern-based module with morphological and syntactic conversion rules. They showed that a combination of the two systems outperformed these systems employed independently. Déjean et al. (2005) used bilingual lexicons automatically extracted from parallel and comparable corpora to enrich standard resources (such as UMLS) and reported that such improved lexicons also improve performance in the medical CLIR. Markó et al. (2005); Markó et al. (2007) in their multilingual retrieval system for medical documents followed the approach of translating both queries and documents into a morpho-semantic representation. They employed a dictionary constituting equivalence classes of morpho-semantically minimal units. The system outperformed other IR and CLIR approaches mainly for languages such as German, where decompouding words into smaller lexical units has a great potential to improve IR performance.

Both IR and CLIR in the medical domain have traditionally been a subject of various shared tasks and evaluation campaigns. The first such an activity was OHSUMED organized in 1994 (Hersh et al., 1994). Their test collection contained around 350,000 abstracts taken from 270 medical journals over a five-year period (1987–1991) and topics created in two ways: queries constructed manually without any formal restrictions and queries based on the controlled vocabulary thesaurus of the *Medical Subject Headings* (MeSH) (Rogers, 1963).

Another shared task related to the medical domain, the TREC Genomics Track (Roberts et al., 2009), ran between 2003 and 2007 and included several task (ranging from ad-hoc retrieval to document categorization, passage retrieval, and entity-based question-answering). The test collection comprised genomics-related publications from medical journals and clinical reports. The TREC Medical Records Track ran in 2011 and 2012 (Voorhees and Tong, 2011). The test collection contained anonymized medical records and the queries were resembled eligibility criteria of clinical studies. The goal of the task was to find patient cohorts that are relevant to the given criteria.

The CLEF eHealth series has been running since 2013 (Suominen et al., 2013; Goeuriot et al., 2014; Palotti et al., 2015). The test collection for 2013–2015 comprised about 1 million pages crawled from English medical websites. The IR tasks focused on queries posed by laypeople searching the web for medical information. In addition to the standard monolingual task, the campaign recently provided non-English queries to be used in a cross-lingual setting. This test collection was also used in the CLIR experiments described in this thesis.

[17] `http://www.clinicaltrials.gov/`

3 Overview of Results

This section provides an overview of our own contribution to the area of SMT adaptation to specific domains and applications. It is divided into subsections, each focusing on a particular piece of our research and referring (in bold) to the papers presented in Part II of this thesis. In each subsection, we describe the main content of a particular paper(s), put it into the context of our other work, present the main results, and summarize the findings and conclusions.

3.1 Web-crawling of resources for domain-adaptation (Pecina et al., 2011, 2012a)

Pecina et al. (2011) was our first contribution to SMT domain-adaptation, shortly followed by **Pecina et al. (2012a)**, which extended the work presented in the earlier paper. The research was conducted with cooperation with ILSP (Institute for Language and Speech Processing, Athens, Greece) within the Panacea project[18] (Bel et al., 2012) and focused on acquisition of domain-specific data from the web (monolingual as well as parallel). The evergrowing web provides vast amounts of texts in various languages and domains and as such, it is a convenient source of data for domain adaptation of SMT. Within the Panacea project, a method to web-crawl monolingual and parallel data from specific domains was developed and applied to the domains of labor legislation (*lab*) and environment (*env*) and two language pairs: English–French (EN–FR) and English–Greek (EN–EL).

The selection of the domains of labor legislation and environment was motivated by the Panacea's grant provider – the two domains are examples of important topics in the context of the European Union. The language selection was also well thought-out. It covers two conditions which can occur in a real-world situation. English and French are relatively similar languages (translation is easier) with a lot of existing language resources. For such a language pair, not only is translation easier *per se*, but the methods can also benefit from using larger data. English and Greek, on the contrary, are more "distant" (translation is harder) and less-resourced, which, in combination, poses a greater challenge to our task. The two domains and two language pairs formed the total of eight evaluation scenarios (translation in both directions, from English and to English) which can well cover possible use cases in practice.

The crawling procedure applied in the two papers operated with a queue initialized by several manually specified "seed" URLs for each domain and language. Such URLs can be identified either from manually maintained resources, such as the Open Directory Project[19] (which was exploited to get the list of relevant URLs for the *env* domain), or they can be obtained using web search engines to retrieve URLs of websites found for queries containing domain-specific keywords (which was done to collect the seed URLs for the *lab* domain). The crawling algorithm then retrieved the URLs from the queue and classified them to be in-domain or not, based on occurrence of predefined domain-specific keywords. The keywords were extracted from the multilingual Eurovoc thesaurus[20]. The in-domain pages were stored and links appearing in those pages were inserted in the queue. The algorithm continued until the queue was empty. The data was then cleaned and normalized. Duplicities were removed and parallel pages detected by Bitextor (Esplà-Gomis and Forcada, 2010) and parallel sentences extracted by Hunalign (Varga et al., 2005).

The entire data acquisition workflow (for parallel data) is visualized in Figure 1. It was implemented as an easy-to-use web service ready to be employed in industrial scenarios. It requires only limited human intervention for constructing the domain definition and the list

[18] `http://www.panacea-lr.eu/`

[19] `https://www.dmoz.org/`

[20] `http://eurovoc.europa.eu/`

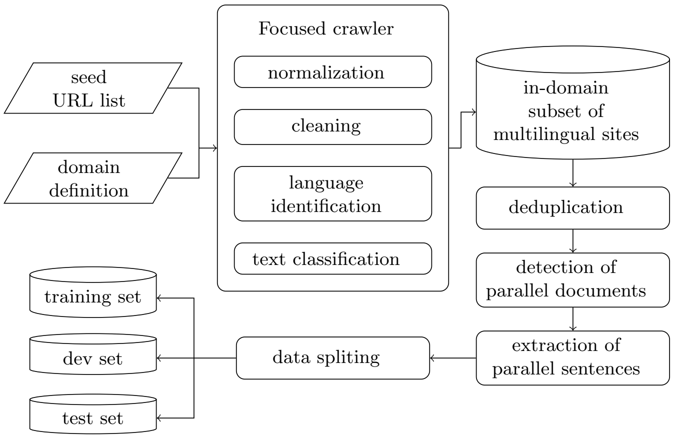

Figure 1: A schema of the workflow of parallel data web-crawling **(Pecina et al., 2015)**.

of seed URLs, which can easily be tweaked and tuned to acquire texts with a high accuracy of 94% (manually evaluated on a sample of data). The initial phase up to the detection of duplicate documents was designed by ISLP, the remaining parts by our team.

The acquired parallel data was initially used to produce development and test sets for the two domains and two language pairs. The procedure included automatic cleaning and alignment at sentence level. The data was then sampled for sentence pairs which were manually validated and corrected to obtain reliable data for tuning and testing translation quality of MT systems. The development data sets contained 500−1,400 sentence pairs (depending of language) and the test data set consisted of 2,000 sentence pairs for each language pair. The monolingual data obtained by the procedure was used as additional training data for adaptation of language model (LM). It consisted of about 1 million words per language.

In the initial experiments, the acquired development data was used for parameter optimization of an SMT system and the monolingual data for its language model adaptation. This was based on joining general-domain and domain-specific data into a single training set and alternatively to train an additional language model which was then combined with the baseline general-domain model using the log-linear combination during decoding. The baseline general-domain systems were trained on data extracted from the Europarl corpus (Koehn, 2005) and the baseline BLEU scores ranged from 20 to 31, depending on the domain and language pair (see Table 1, column denoted as *baseline*). The initial adaptation experiments (in-domain tuning, LM adaptation) showed a substantial improvement of the translation quality in all the evaluation scenarios (combination of the two domains and two language pairs in both directions created eight evaluation scenarios in total). In terms of automatic evaluation (measured by BLEU), the overall effect of using in-domain data was up to 48% relative compared to the unadapted baselines. Most of the improvement was caused by in-domain tuning, LM adaptation did not bring any large additional improvement, mostly because of the limited amounts of the in-domain monolingual data.

Based on those results, the data acquisition procedure was applied to acquire larger amounts of in-domain data to allow deeper analysis of its effect on LM and TM adaptation. The experiments published in **Pecina et al. (2012a)** were conducted on the same domains and language pairs using the same test sets as in **Pecina et al. (2011)**. The monolingual crawling procedure was running, on average, for 50 hours per language and domain. In total, it visited approximately 750 thousand pages, 1/4 of them were classified as in-domain and 1/3 out of those were discarded because of being duplicates. The remaining pages contained about 25 million pieces of text (paragraphs), out of which, 23% where removed because of being boilerplate and additional 14% because of being duplicates. Finally, the total amount of monolingual

direction	*dom*	*baseline*	*development*		*monolingual*		*parallel*		*both*	
		BLEU	BLEU	Δ%	BLEU	Δ%	BLEU	Δ%	BLEU	Δ%
English–French	*env*	28.03	35.81	*27.8*	39.23	*40.0*	40.53	*44.6*	40.72	*45.3*
	lab	22.26	30.84	*35.6*	34.00	*52.7*	39.55	*77.7*	39.35	*76.8*
French–English	*env*	31.79	39.04	*22.5*	40.57	*27.6*	42.23	*32.8*	42.17	*32.7*
	lab	27.00	33.52	*23.7*	38.07	*41.0*	44.14	*63.5*	43.85	*62.4*
English–Greek	*env*	20.20	26.18	*29.1*	32.06	*58.7*	33.83	*67.5*	34.50	*70.8*
	lab	22.92	28.79	*25.7*	33.59	*46.6*	33.54	*46.3*	33.71	*47.1*
Greek–English	*env*	29.23	34.15	*16.8*	36.93	*26.3*	39.13	*33.9*	39.18	*34.0*
	lab	31.71	37.55	*18.4*	40.17	*26.7*	40.44	*27.5*	40.33	*27.2*
Average				*25.5*		*40.0*		*49.2*		*49.5*

Table 1: BLEU scores obtained by domain adaptation of the baseline systems by exploiting in-domain resources: development data, monolingual training data, parallel training data, both monolingual and parallel training data. Δ% denotes relative improvements over the baseline.

data acquired ranged from 15 million to 45 million words depending on language and domain (English/French was richer, Greek less-resourced, as expected). As for parallel data, about 250 thousand to 700 thousand words per language pair was collected.

Those amounts of data allowed us to perform several interesting experiments. The main results are presented in Table 1. Language model adaptation using larger amounts of in-domain training data (LM adaptation only) can improve translation quality by a surprising difference (Table 1, column *monolingual*). Such data, of course, do not reduce OOV rates of the systems or introduce new translation options of known words, however, it can contribute to better estimations of language model probabilities of phrases consisting of known words which then help to select better translation variants during decoding. The average relative increase of BLEU obtained by LM adaptation via mixing the in-domain and general-domain data was about 14.5% relative compared to the in-domain tuned systems using the baseline models only. In comparison with the baseline, this gain was 40.0% relative on average. Not surprisingly, a larger improvement was also observed after exploiting the parallel data (TM adaptation only). Here, the trivial adaptation method which mixed the in-domain data with the general-domain sets were employed and single translation models were trained (Table 1, column *parallel*). The average increase of BLEU scores was 49.2% relative compared to the baseline and about 23.8% relative compared to the in-domain tuned systems using the baseline models only. To provide a complete picture, a fully adapted system was also trained using both general-domain and domain-specific sets of parallel and monolingual data (mixed) and tuned on the corrected in-domain development sets (Table 1, column *both*). In most scenarios, the difference of results of these systems compared to the TM-adapted systems were not statistically significant (measured by the test described in Koehn, 2004b, p=0.05).

Conclusions

Overall, the experiments proved that domain-focused web-crawling is an efficient way of acquisition of domain-specific data which can be effectively used for domain adaptation of SMT. Even very small amounts of in-domain parallel data (as few as 500 sentence pairs) can be used as development data to tune the system parameters. Additional parallel training data can improve the translation models. If in-domain parallel data is not available at all, a general-domain system can benefit from using additional in-domain monolingual data but larger amounts (tens of millions of words) are needed to obtain a substantial improvement. The effect of LM adaptation and TM adaptation, however, did not add up in single systems combining both types of in-domain resources (parallel and monolingual data).

dev	*test*	English–French		French–English		English–Greek		Greek–English	
		BLEU	Δ%	BLEU	Δ%	BLEU	Δ%	BLEU	Δ%
gen	*gen*	49.12	*0.00*	57.00	*0.00*	42.24	*0.00*	44.15	*0.00*
gen	*env*	28.03	*−42.94*	31.79	*−44.23*	20.20	*−52.18*	29.23	*−33.79*
gen	*lab*	22.26	*−54.68*	27.00	*−52.63*	22.92	*−45.74*	31.71	*−28.18*
gen	*med*	12.32	*−74.92*	15.33	*−73.11*	8.96	*−78.79*	14.79	*−66.50*
Average			*−57.51*		*−56.65*		*−58.90*		*−42.82*

Table 2: Translation quality (in BLEU) of systems tuned on general-domain and tested on specific domains (*env*, *lab*, *med*) compared with the test results on general domain (*gen*). Δ% indicates the relative change with respect to the generel-domain test set.

3.2 Parameter tuning for domain-adaptation (Pecina et al., 2012b)

The follow-up paper **(Pecina et al., 2012b)** concentrated on the scenario where a general-domain MT system needs to be adapted to a specific domain for which the only available in-domain resource is very limited amounts of parallel data or no in-domain data at all. In such a situation, the system can be adapted by tuning its parameters (i.e., weights of the underlying log-linear model) which can be realized in several approaches: proper in-domain tuning (using available data as development sets), cross-domain tuning (using development data from other domains) or no tuning at all. The experiments described in the paper were performed on the two domains as in the previous publications (labor legislation and environment) and for the first time, the experiments were conducted on the medical domain. The language pairs were the same (English–French and English–Greek). This formed 12 evaluation scenarios in total: three domains and four translation directions.

First, we confirmed the results from the previous papers, that systems trained and tuned on general domain perform poorly on specific domains, also on the medical domain (even in a larger extent). In general, this observation was not very surprising, similar results were already reported before (e.g., by Wu et al., 2008; Banerjee et al., 2010) but our results were very consistent in all evaluation scenarios and the amount of loss was unexpected. Overall, on average, the relative decrease was 54% (in terms of BLEU). In some scenarios, the relative decrease was as high as 75% and 78% (English–French and English–Greek translation in the medical domain, respectively). In absolute figures, the BLEU scores dropped from 49.12 to 12.32 for English–French and from 42.24 to 8.96 for English–Greek (the complete results for all the 12 scenarios are shown in Table 2.) This clearly confirmed certain overfitting to the domain of the training and tuning data. The magnitude of the drop correlates with the perplexity of the source side of the test data given the source side of the training data. The lower perplexity (better fit), the higher BLEU (better translation), see Figure 2. This can be potentially used to predict translation quality. The experiments also confirmed findings from the previous papers on the new domain (medicine): in-domain tuning (i.e., when parameters of systems trained on general domain are optimized on specific target domain data) can recover a great amount of the loss. The BLEU scores raised by 33% relative on average (comparing general-domain tuned vs. in-domain tuned systems tested on the specific domain test sets). On the medical domain, the scores raised from 12.32 to 18.47 for English–French and from 8.96 to 14.57 for English–Greek, for instance.

The observed improvements are in MT experimentation immense, especially given the fact that, the only change of the system is in parameter setting, which is obtained using a very small data set. In the paper **(Pecina et al., 2012b)**, we provided an explanation how and why this is effective: A system trained, tuned and tested on general domain tends to prefer long and few phrases in the output translations and therefore underperforms when tested on specific domains where the (longer) test set phrases do not occur in the training data

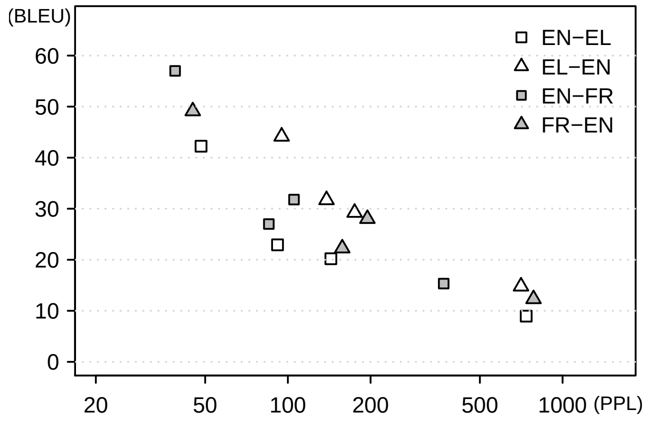

Figure 2: Perplexity (PPL) of the source side of the test data versus BLEU scores of the corresponding systems tuned on general-domain development data.

frequently (the average phrase length in these scenarios was 2.3 words and the distribution almost uniform). By contrast, the same systems, i.e., trained on general domain, but tuned on specific-domain data form output translations from shorter and larger number of phrases (the average phrase length was 1.8 words, and extremely skewed towards the short end), allow more reordering (evident from the higher weights of the reordering model features) and perform significantly and consistently better on the specific-domain data. In a sense, this is natural: substantial divergence between test and training data means that in particular long and potentially high quality phrase pairs obtained in training may no longer be applicable to the test data and that this divergence can only be bridged by smaller translation units and more flexible combination.

In another (also common) scenario when no in-domain data is available for parameter tuning, a possible solution is skipping tuning, i.e., using the default model parameters. This performed surprisingly well. This approach is recommended to be preferred over general-domain tuning to avoid overfitting, especially if the training and test domain differ substantially (which can be measured again by cross-perplexity of the test data and the training data). Another possible solution is cross-domain tuning, i.e., using development data from a different domain. For instance, for an English–Greek translation in the medical domain there was no difference when the system was tuned on either medical, labor-legislation or environment data (always around 37.55 BLEU). This approach has the effect of disassembling the original general-domain system towards shorter phrases and it does not make matter much which different development set to use.

Further, analysis of learning curves of various tuning processes was conducted. In the previous experiments, the size of the development sets ranged between 500 and 2,000 depending on their availability for each language pair. In learning curve analysis experiments, the size of the development sets varied from zero (this corresponds to skipping tuning) up to the maximum and measured the translation quality of the tuned systems (again in terms of BLEU). The results showed that decent-quality in-domain tuning of the general-domain-trained systems required about 100–200 sentence pairs only, the gain obtained from using more data was negligible. The results are depicted in Figure 3 and are quite encouraging, as in-domain tuning yields the best results and requires relatively small amounts of parallel data. The development sets of more than 400–600 sentence pairs do not improve translation quality at all but at the same time additional tuning data does not actively degrade performance so there is no need to reduce the size of the tuning set either.

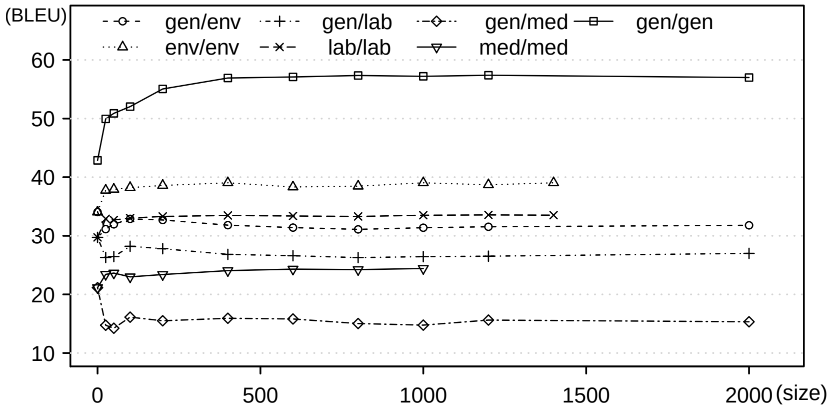

Figure 3: Translation quality of French–English MT systems tuned/tested on data from different domains and of varying size.

Conclusions

The main findings of the paper **(Pecina et al., 2012b)** came from the analysis of parameter tuning for specific domains and were formulated as guidelines for its effective solutions, including in-domain tuning, cross-domain tuning and skipping tuning to avoid overfitting. The acquired corpora are available from ELRA[21] under reference numbers ELRA-W0057 – ELRA-W0058 and ELRA-W0063 – ELRA-W0068.

3.3 Domain-adaptation using web-crawled data (Pecina et al., 2015)

Our overall effort in SMT domain-adaptation using web-crawled data culminated in **Pecina et al. (2015)** which reviewed the data acquisition methods and significantly extended the MT adaptation part. It provided more details of the experiments, full results, and a more thorough analysis and description of our findings. The test data scenarios included the environment and labor legislation domains and the English–French and English–Greek language pairs in both translation directions.

In the first part of the paper, we reviewed the web-crawling procedure and provided some additional measurements and analysis. (e.g., precision of visited pages classified as in-domain during crawling, which was estimated around 20%, i.e., 20% of visited pages were classified as in-domain, stored, and the outgoing hyperlinks followed). The second part was focused on the domain adaptation methods and compared various approaches including the state-of-the-art methods for adaptation of language models and translation models which further improved our best results from the previous work. Compared to the previous papers, the translation quality evaluation in this work is conducted on tokenized and lowercased translations to avoid any bias caused by recasing and detokenization. The main contribution can be split into the areas described below.

Correction of development data

In the initial experiment, the practical need to correct development data acquired by automatic webcrawling was assessed. In all the experiments presented earlier, the in-domain development data sets were automatically extracted from the webcrawls and then underwent manual

[21] `http://catalog.elra.info/`

	EN–EL/*env*	EN–FR/*lab*
1. perfect translation	53.49	72.23
2. minor corrections done	34.15	21.99
3. major corrections needed	3.00	0.33
4. misaligned sentence pair	5.09	1.58
5. wrong domain	4.28	3.86
Total	100.00	100.00

Table 3: Statistics (%) of manual correction of sentences from the web-crawled parallel data.

checking and corrections. The automatic acquisition procedure can not guarantee a good translation quality in any sense and if it is suboptimal it might have a negative impact on the tuning procedure. During the manual corrections, it was observed that 53–72% of sentence pairs were accurate translations (no corrections needed), 22–34% pairs needed only minor corrections, 1–3% would require major corrections, 2–5% of sentence pairs were misaligned and would have had to be translated completely, and about 4% of the sentence pairs were clearly from a different domain. See Table 3 for detailed figures.

To decrease the manual effort to create the development sets, some of the sentence pairs were not included, namely those which required major changes or complete translation from scratch and those which were out-of-domain. The amount of manual work was not trivial and we wanted to verify if it is necessary to perform such a step in order to create development data in real-world applications. Systems tuned on the manually corrected development sets were compared with systems tuned on raw development sets. This raw development data contained not only the sentences with imperfect translation, but also those that are misaligned and/or belong to other domains. As a consequence, the raw development sets contained about 10% more sentence pairs than the corrected ones. Surprisingly, in most experiments the results did not show any statistically significant difference which makes the manual correction of development data acquired by our procedure unnecessary in practice.

Analysis of model parameters

We have elaborated on the analysis of model parameter changes when switching from general-domain to domain-specific tuning. Values of the log-linear combination parameters in SMT are usually not investigated (they are not stable) but our experiments showed consistent results in all the evaluation scenarios and indicated interesting trends (see, for example, the visualization of feature weights of the English–French model in Figure 4).

In general, the systems trained and tuned on the general domain are characterized by the following observations: 1) the high weights assigned to h_{11} (*direct phrase translation probability*) indicate that the phrase pairs in their translation tables apply well to the matching-domain development data and translation hypotheses consisting of phrases with high translation probability are preferred (i.e., good general-domain translations); 2) the low negative weights assigned to h_{13} (*phrase penalty*) imply that the systems prefer hypotheses consisting of fewer but longer phrases; 3) the weights of the reordering models h_1–h_7 are assigned values around zero which implies that phrase reordering is not explicitly preferred. [22]

In comparison with the systems trained on the general domain and tuned on the specific domain, the following was observed: 1) the weights of h_{11} (*direct phrase translation probability*) decrease rapidly (in some scenarios even close to zero) which can be explained by a lack of good quality phrase translations for the specific domains; the best translations of

[22] The features (h_1–h_{14}) are explained in Section 2.2.

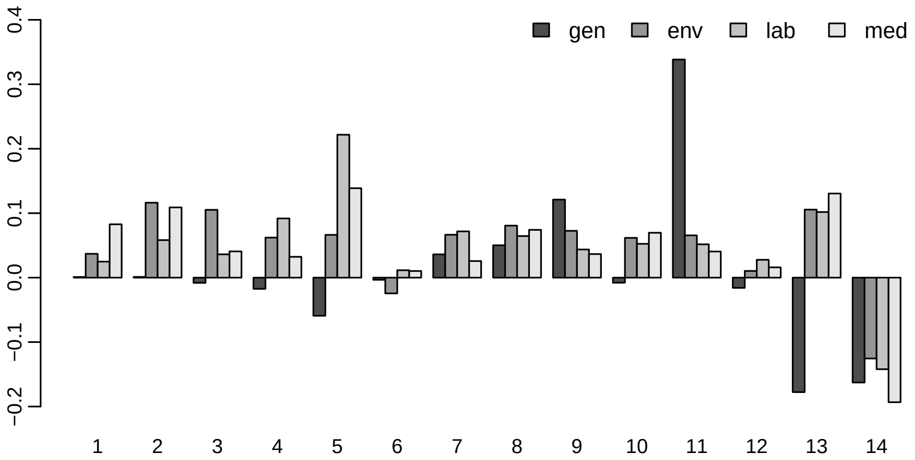

Figure 4: Visualization of the 14 feature weights of the English–French system tuned for four domains; the black bars refer to model weights of the system tuned on general-domain (*gen*) data while the gray bars refer to the model weights of the systems tuned on specific-domain data (*env*, *lab*, *med*). The features (1–14) are explained in Section 2.2.

the development set sentences are then formed by phrases of varying translation probability scores; 2) hypotheses segmented into a few (but longer) phrases are not preferred any more (weights of h_{13} are higher); instead, they are usually penalized and hypotheses segmented into more (and shorter) phrases are allowed or even preferred; 3) in almost all scenarios, the reordering model weights (features h_1–h_7) increased significantly, and the systems strongly prefer hypotheses with reordered words/phrases. More details can be found in **Pecina et al. (2015)**, Section 5. We also analyzed changes in phrase length distribution and found them consistent with our findings about feature weights.

Language model adaptation

As it was already pointed out, adapting an SMT system by adding in-domain monolingual training data can improve language model estimation and help to select better translation hypotheses based on their fluency. There are two principled ways of using monolingual data for adaptation of a language model: to replace the existing model by a new one trained on a simple concatenation of the original general-domain data and the new domain-specific data; or to build an additional language model from the domain-specific data and use it together with the original one. The fist approach is trivial, non-parametrized and not really possible to optimize. The second approach has the advantage of being optimizable. It can be realized in two ways (Foster and Kuhn, 2007): either, the two models are merged by linear interpolation into a single model or the two models are directly used as components of the log-linear feature combination. The two ways are similar but not identical. Both are parametrized to control relative importance of the two models: linear interpolation has a single coefficient weighting the two probability distributions in a linear manner, whereas in the second approach, the weighting happens in log-linear space. The first can be optimized by maximizing perplexity of some target-language data (e.g., the target side of the development set), the latter allows direct optimization towards MT quality (e.g., by MERT).

In the LM adaptation experiments, the general-domain data comprising 27–53 million tokens (per language) was combined with the in-domain data comprising 15–45 million tokens (per language) using the three approaches described above and all of them were proven to bring about better translation quality in all scenarios. The concatenation method is a special

direction	*dom*	*baseline*	*+tuning*		*+lang. model*		*+transl. model*		*specific only*	
		BLEU	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ
English–French	*env*	29.61	37.51	*7.90*	41.78	*4.27*	43.85	*2.07*	39.54	*−4.31*
	lab	23.94	32.15	*8.21*	38.54	*6.39*	48.31	*9.77*	43.05	*−5.26*
French–English	*env*	31.79	39.05	*7.26*	42.63	*3.58*	44.22	*1.59*	37.86	*−6.36*
	lab	26.96	33.48	*6.52*	41.11	*7.63*	50.56	*9.45*	43.74	*−6.82*
English–Greek	*env*	21.20	27.56	*6.36*	34.89	*7.33*	37.90	*3.01*	29.84	*−8.06*
	lab	24.04	30.07	*6.03*	34.15	*4.08*	34.76	*0.61*	26.19	*−8.57*
Greek–English	*env*	29.31	34.31	*5.00*	37.57	*3.26*	40.64	*3.07*	30.71	*−9.93*
	lab	31.73	37.57	*5.84*	40.09	*2.52*	40.75	*0.66*	29.54	*−11.21*
Average				*6.64*		*4.88*		*3.78*		*−7.57*

Table 4: Results of incremental adaptation: tuning, adapting the language models, and adapting the translation models. The last two columns refer to systems trained/tuned on domain-specific data only. Δ refers to absolute incremental improvement over the previous system.

case of linear interpolation (the coefficient is given by the relative size of the general-domain and domain-specific data) and it was no surprise that the linear interpolation approach was more flexible and effective, especially in the scenarios with larger development data available (allowing more reliable estimations of the interpolation coefficients). The average absolute improvement over all the evaluation scenarios was 4.88 BLEU points compared to the system with in-domain-tuned parameters, while the simple concatenation method achieved 3.78 BLEU points. Linear interpolation was also found to be superior to the log-linear combination method, which achieved 2.94 BLEU points (of absolute improvement on average), probably due to the tuning procedure (MERT) which operates in a complex feature space (model scores) with a complex objective function (BLEU). See Table 12 in **Pecina et al. (2015)** for details.

Translation model adaptation

Parallel data (especially from specific domains) is much scarcer than monolingual data but for training an MT system more important. While a good language model can improve an SMT system by better selection of phrase translation options in given contexts, it can not help if the translation model provides no translation option for a give phrase at all.

Methods for translation model adaptation to specific domains are analogous to those applicable to language models. General-domain and domain-specific data can be combined in three ways (e.g., Banerjee et al., 2011): 1) concatenation of the two types of data and retraining the translation model from scratch; 2) training a new translation model on the domain-specific data and its linear interpolation with the general-domain one in a linear fashion; 3) using the two independent translation models in log-linear combination during hypothesis scoring. The first approach does not require optimization of any additional parameters (the weight is given by the relative size of the data). In the second approach, four new coefficients must be set (one for each of the probability distributions provided by the Moses' translation model, i.e., h_9–h_{12}), usually based on optimization towards minimal perplexity of the development data (Sennrich, 2012). In the third approach, there are five new weights in the log-linear combination associated with the additional translation model (doubling h_9–h_{13}) which are then optimized together in the traditional way by maximizing translation quality on the development data (e.g., by using MERT).

All the alternative approaches were analyzed and compared in experiments exploiting in-domain data sets of 7,000–20,000 sentence pairs, depending on the language pair and domain, and substantial improvements in translation quality were observed in all the scenarios. In comparison with the systems trained on general domain and tuned for specific domains, the increase was 3.94 BLEU for the concatenation method, 4.35 BLEU for the linear interpolation method, and 4.56 BLEU for the log-linear combination method (all absolute, see **Pecina et al.**, **2015**, Table 13). Here, the differences between the approaches were smaller and often statistically indistinguishable. Overall, the most effective approach is the log-linear combination, but the improvement compared to the alternatives was statistically significant in three scenarios only. This is mainly caused by the limited amounts of the in-domain adaptation data and its match with the test data (see **Pecina et al.**, **2015**, Section 6.3 for details).

In combination with LM adaptation, the system improved even further. Compared to the LM adapted systems, the increase was 3.58 BLEU for the concatenation method, 3.78 BLEU for the linear interpolation method, and 3.69 BLEU for the log-liner combination method (all absolute, see **Pecina et al.**, **2015**, Table 14). Here the differences between linear interpolation and log-linear combination were even smaller. However, the important observation is that the effect of using in-domain monolingual and parallel data is largely independent and does not cancel out when these two types of resources are used at the same time. This was not confirmed in our earlier experiments **(Pecina et al.**, **2012a)** where only the trivial concatenation-based adaptation methods were employed.

The complete results of the incremental adaptation are shown in Table 4, which also reports the results of systems trained and tuned on domain-specific data only to illustrate the pure effect of such training data. In almost all scenarios, these systems outperformed the baseline (by 7.74 BLEU on average). However, the requirement of using general-domain data is unquestionable, the fully adapted systems were better by 7.57 BLEU on average (see the last column in Table 4).

Conclusions

The presented article **Pecina et al. (2015)** concluded our work on domain adaptation conducted within the Panacea project. This research was mostly applied with some new interesting findings. We provided guidelines to develop an MT system adapted to a specific domain based on techniques exploiting domain-specific data which can be easily crawled from the web. We developed the entire pipeline to crawl domain-specific monolingual and parallel data which is initiated by a small number of domain-relevant websites and short definition of the domain consisting of a set of (weighted) keywords. The pipeline proved to be very successful. It was applied to acquire adaptation data for two domains to allow translation for two language pairs in both directions. We conducted a large-scale comparison of the state-of-the-art adaptation techniques, including a deep analysis of the changes in the adapted systems. The findings have practical implications in real-word applications of MT to specific domains.

3.4 Medical text translation (Urešová et al., 2014; Dušek et al., 2014)

After concluding our work within the Panacea project (Bel et al., 2012), our research focus moved to another EU FP7 project called Khresmoi[23] (Aswani et al., 2012, 2013). While Panacea primarily focused on acquisition of domain-specific language resources and MT was used as a typical use-case for exploiting such resources, in Khresmoi, MT was a key component for providing multilingual capabilities in the multi-modal search and access system for biomedical information and documents developed as the main objective of the project.

The large area of medicine and health is an example of domain with a very specific and extensive vocabulary. It is widely used in the large community of medical professionals as

[23] `http://www.khresmoi.eu/`

general public queries	medical professional queries
cardiac phenotypes	*prostate*
hallucination	*lid hemangiomas AND avastin*
health	*cannabinoid pain*
cathartics	*trombocitopenia in pregnancy*
vitamin d	*nurse training*
AIDS serodiagnosis	*quality attachment treatment*
hydro	*delirium elderly*
ulceration	*nursing and hours*
atelectasis	*migraine disorders AND adrenergic beta agonist*
digital rectal examination	*terminally ill patients*
acute kidney injury	*gastric bypass*
development	*tarsal*
radioactive materials	*opinion*
agammaglobulinemia	*lesion neural*
neonatal	*training*

Table 5: Sample medical queries by general public and medical professionals taken from the Khresmoi Query Translation Test Data 1.0.[28]

well as in everyday life of laypeople. For example, Fox (2011) reported that 70% of search engine users in the US have conducted a web search for information about a specific disease or health problems. The fact that most medical content available on the Internet is written in English and that a large number of non-English speakers need to have content in their native language brings up the issue of better utilization of such information and make it accessible to a larger population (Cline and Haynes, 2001). Khresmoi targeted both groups of users, medical professionals (doctors) and laypeople (patients), and developed two systems providing access to medical information and documents across languages: Khresmoi Professional (Kelly et al., 2014) and Khresmoi for Everyone (Pletneva et al., 2014). Users of the systems were allowed to enter non-English queries to retrieve information in English documents.

The MT component of the two systems was designed to perform two distinct tasks: 1) translation of non-English search queries into English (to allow retrieval of English-written documents) and 2) translation of automatically generated summaries (snippets) of the retrieved documents into a user-specified language (to be presented to the user). Translating search queries is an inherently difficult task for current SMT systems due to several reasons: the queries are rather short expressions or ungrammatical sequences of terms with a lack of context, often mixing languages and containing words which should be treated as search operators (e.g., *swallowing in stroke patient*, *frequent nosebleeds in children*, *polycythemia AND stroke*). In contrast, the summary sentences are usually fluent and fully grammatical, but the fact that they were taken from summaries of medical documents implies that they are informatively "dense" (containing large number of content/terminological words) and rather long (compared to an average sentence) and thus more difficult to translate.

Test data preparation

One of our first goals in Khresmoi was to acquire test data sets for translation quality evaluation. The data sets were in a form of sample texts (search queries and summary sentences) in English with translations into other languages (Czech, German, and French). The data comprised two parts: one for final testing and one for development testing (system parameter tuning). For the query translation task, the entire process of test data acquisition was described in **Urešová et al. (2014)**. The English side of the data sets was extracted from two samples of real queries: queries posed by general public through the Health on the Net Foundation

summary sentences
In type II infection (also known as hemolytic streptococcal gangrene), group A streptococci are isolated alone or in combination with other species, most commonly S. aureus.
The amount of radiation used for a chest x-ray is very small.
Meningococcal Disease is a serious bacterial infection that can cause swelling of the brain and spinal cord, and infection of the blood and other organs.
Issues related to necrotizing cellulitis and necrotizing fasciitis will be reviewed here.
Solid organ transplant recipients have an increased risk for infection with NTM due to depressed cell-mediated immunity, but NTM infections are nevertheless rare in this population.
Osteoarthritis, rheumatoid arthritis and traumatic arthritis are three common causes of hip pain and immobility.
Exam revealed this ulcer (at 9:00 o'clock position) with a white, fibrinous base, and a dark, protruding visible vessel, signifying the site of recent bleeding.
When should I treat someone with COPD with a theophylline (methylxanthine)?

Table 6: Sample sentences extracted from summaries of medical articles taken from the Khresmoi Summary Translation Test Data 1.1.[29]

website[24] and queries posed by healthcare professionals in the Trip database[25] (Meats et al., 2007). The queries were manually checked, non-English and nonsensical items appearing in the texts (e.g., *asdfghj*) were removed. Spelling errors were corrected if the true meaning was identified unambiguously, otherwise the misspelled queries were discarded too. The final set consisted of 749 general public queries and 759 medical professional queries in English which were then human-translated into German, French, and Czech.

Initially, the translation was done by medical non-experts but quality of their output was found not to be sufficient. Therefore medical experts (doctors) must have been hired to carry out the work. All the translators were fluent in the target languages (but not necessarily native speakers). They were explicitly instructed to preserve the original (non-)syntax (translate as a phrase if the query appears to have syntax, otherwise translate the words one by one, not introducing any grammatical structure), e.g., *colon cancer* should be translated as *rakovina tlustého střeva/Dickdarmkrebs/cancer du côlon* (noun phrase), but *pain cancer* should result in *bolest rakovina/Schmerz Krebs/douleur cancer* (no syntax). Regarding abbreviations, the guidance was to keep the English original if it is used in the target language as well, (e.g., *EEG, CRP*) and use target language conventions for the meaning of the source abbreviation, i.e., use abbreviation in the target language if the abbreviation is commonly used, such as *JIP/ITS/USI* for *ICU* (meaning *Intensive Care Unit*), but use full text if that is the norm, e.g., *ultrazvukové vyšetření v reálném čase/Echtzeit-Ultraschall/ultasons en temps réel* for *RTU* (meaning *Real-Time Ultrasonography*). A special treatment of logical query operators was also required (*AND*, *OR*) which should have been identified (distinguished from conjunctions in their usual meaning) and left intact, as in *žíravý AND stent/caustique AND stent/kaustisch AND Stent* for *caustic AND stent*. The translations were then reviewed by independent experts, discrepancies resolved and approved by the translators. The resulting data set comprised a total of 1,508 queries which were split into a test set (1000 queries) and development set (508 queries), each with equal portions of general public queries and medical professional queries.

For the summary translation task, the test set acquisition and translation process was described in Bojar et al. (2014). The data contained sentences randomly sampled from automatically generated summaries (extracts) of English documents (web pages) containing medical information found to be relevant to 50 topics provided for the CLEF 2013 eHealth

[24] `http://www.hon.ch/`

[25] `http://www.tripdatabase.com/`

Task 3[26]. Manual processing of the data included removal of out-of-domain and ungrammatical sentences, translation by medical experts into Czech, German and French, and its revision. The final version of the data comprised 1,000 sentences in the test set and 500 sentences in the development set. Both the data sets were published and are available under the Creative Commons License from the LINDAT/CLARIN repository as Khresmoi Query Translation Test Data 1.0[27] and Khresmoi Summary Translation Test Data 1.1[28]. Within the KConnect project[29] in 2016, both the data sets were extended to four additional languages: Hungarian, Polish, Spanish, and Swedish.

WMT medical translation shared task

The two data sets were primarily created for MT development and evaluation within the Khresmoi project. In addition to that, these very unique resources were exploited in a shared task on translation in the medical domain. This shared task was organized by with a support from the Dublin City University as a part of the Ninth Workshop on Statistical Machine Translation (WMT) in 2014. The series of WMT workshops (since 2005) has a long tradition of annually organized shared tasks focused on translation between European languages. The medical translation task was accepted as the featured task for 2014. The workshop was collocated with the 52^{nd} Annual Meeting of the Association for Computational Linguistics in Baltimore, Maryland, USA in June 2014.

The goal of the shared task was to investigate the applicability of MT to translate domain-specific (medicine) and genre-specific (query and summary) texts between English and the Khresmoi languages (Czech, German, French) in both directions. Following the MT goals in Khresmoi, the shared task was split into two subtasks: 1) translation of sentences from summaries of medical articles, 2) translation of queries by users of medical search engines. In addition to the development and test data, we also provided access (in a form of URLs) to in-domain and out-of domain data for training (both monolingual and parallel). The shared task participants were asked to develop their systems using the specified resources (in the *constrained* task) or any additional resource (in the *unconstrained* task) and submit their translations of the test sets within 5 days. Our own system, developed within Khresmoi (called Khresmoi Translator) and described in **Dušek et al. (2014)** was used as a relatively strong baseline.

The results of the task were described in Section 5 of the task overview paper (Bojar et al., 2014). The total of eight teams participated in the shared task. The evaluation was carried out by automatic measure including BLEU (Papineni et al., 2002), TER (Snover et al., 2006b), PER (Tillmann et al., 1997), and CDER (Leusch et al., 2006). Human evaluation was not performed. The main reason was the lack of domain-specific expertise of prospective raters. Human evaluation of translation quality in this specific domain would have required a very good knowledge of the domain to provide reliable judgments and the raters with such an expertise (medical doctors and native speakers) were not available.

The results varied depending on the subtask and translation direction. Most of the systems were based on the teams' systems applied to the standard translation task but trained on the data provided for the medical task. In the query translation subtask, in most translation directions, our systems performed best according to the automatic measures ignoring letter casing. The only exception was the French—English translation direction, where the

[26] `https://sites.google.com/site/shareclefehealth/`

[27] `http://hdl.handle.net/11858/00-097C-0000-0022-D9BF-5`

[28] `http://hdl.handle.net/11858/00-097C-0000-0023-866E-1`

[29] `http://www.kconnect.eu/`

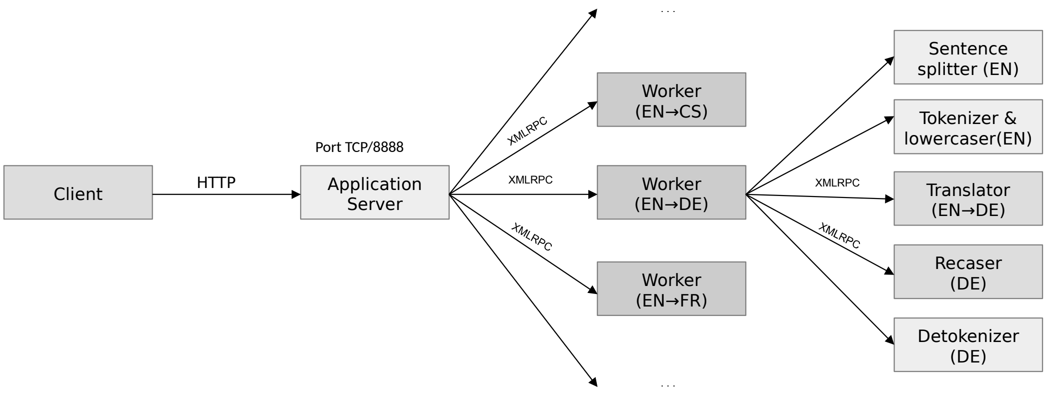

Figure 5: Schema of the MTMonkey architecture for deploying MT as a webservice.

best result was achieved by the team of the Dublin City University. The difference, however, was not statistically significant. In the summary translation subtask, the best overall results were achieved by the University of Edinburgh team, which won for German—English, English—Czech, and English—French, followed by the team of the University of Macao, which performed on par with Edinburgh in all the other translation directions. All these systems did outperform the baselines. The result analysis confirmed our previous observation that tuning on in-domain data is extremely important. The most successful systems employed techniques for linear interpolation of models trained on out-of-domain and in-domain data. Many of them performed selection of pseudo-in-domain data based on Moore and Lewis (2010).

Khresmoi Translator

Our participation in the WMT 2014 medical task mainly aimed to provide baseline results for all the language pairs, translation directions and both subtasks. The systems were developed as the first version of the Khresmoi translation service called **Khresmoi Translator**.

A straightforward phrase-based SMT setting was used together with a specific approach to data selection based on Moore and Lewis's (2010) idea of selecting "pseudo-in-domain data". In the original method, the pseudo-in-domain data is mined from a larger pool of general-domain training data based on sentence-level similarity to some in-domain data samples. After the selection is done, any remaining sentences are discarded (not used). In our approach, this data was not thrown away but used to train "out-of-domain" models to be used jointly in the system combined by linear interpolation. The interpolation coefficients were trained on the development data to maximize their likelihood. The same approach was applied to parallel and monolingual data, so all the models in the systems (language and translation) are linear interpolations of pseudo-in-domain and pseudo-out-of-domain models. This proved to be very successful. For some translations directions, the WMT medical translation baseline systems were not beaten by any of the participants. The method can utilize all available training data in an optimal combination. No training data is wasted and the relative importance of the two models is estimated on a real sample of data. The resulting models are large but the production systems can be easily pruned (by ignoring low-probability phrases) to reduce their size (memory requirements) and speed of translation (response time).

The Khresmoi Translator was deployed as a web service and used by several demonstrators within the project, e.g., Khresmoi Professional (Kelly et al., 2014) and Khresmoi for Everyone (Pletneva et al., 2014). However, since Moses, as an experimental toolkit, does not easily allow to be deployed in real-world applications, we developed **MTMonkey** – an infrastructure for a scalable web service providing MT in multiple languages to remote clients.

It was described and evaluated for efficiency (response time, scalability, etc.) in Tamchyna et al. (2013). It consists of an application server and a set of workers (see Figure 5). The application server receives translation requests and distributes them to the workers, which perform the translation. Each worker runs one instance of Moses to serve one translation direction (multiple workers for one translation direction are allowed). The system is designed to handle multiple simultaneous incoming requests by load balancing and queuing. The system is available as open source and is available through the LINDAT/CLARIN repository[30] and Github[31]. The system has been adopted by several other projects too.

3.5 Query translation for cross-lingual information retrieval (Pecina et al., 2014)

Our participation in the Khresmoi project resulted in a journal article **Pecina et al. (2014)**, a joint work of the teams from the Charles University in Prague and Dublin City University. This paper investigated machine translation of user search queries in the context of cross-lingual information retrieval in the medical domain. The word “adaptation” in the title of the paper is related to two concepts: a) adaptation of MT to a specific domain (medicine) and b) adaptation of MT to a specific application (information retrieval). The role of MT in this task is to translate (or *map*) an input user query to a space where it is used to search documents in a different language. Therefore, translation quality is not the main factor in assessment of MT effectiveness in this task.

The fundamental difference in this scenario is that the user is not a direct consumer of MT output, it is the IR system which performs the search step – the users usually do not care about MT quality per se, they are primarily concerned about the search results, i.e., whether they find what they are looking for. This fact has two consequences: first, since translation quality may not correlate with search quality, it is better to tune the MT system directly for search quality (i.e., extrinsically) rather than for translation quality (i.e., intrinsically) as usual; second, the MT output does not have to be in the traditional human-readable form (text). It can be completely hidden from the user and thus better comply with the design of the decoder.

In reality, assessing translation quality in the context of search queries is precarious. The traditional way of defining translation quality based on adequacy and fluency (see, e.g., White et al., 1994) cannot be straightforwardly applied here. Users of current search engines are biased by the fact that they got accustomed that such systems are typically based on bag-of-words models and ignore word order. Therefore, the queries are often ungrammatical sequences of terms rather than fluent natural language expressions. The issue is whether the translation should preserve such ungrammaticality and disfluency or not (cf., e.g., *skin cancer* vs. *cancer skin*). Assessment of adequacy is also questionable in this context. The queries are usually much shorter than a standard MT input (Spink et al., 2001) and often ambiguous (the underlying information need of the user is not known) and decision on adequacy of translation (i.e., how much information is transferred from the original query to the translation) is difficult. Moreover, neither adequacy nor fluency do reflect how the translation distinguishes between relevant and irrelevant documents which should be the key point in this task (cf., e.g., *skin* vs. *epidermis* vs. *dermis*). Another problem occurs when then original query suffers from “circumlocution” (Stanton et al., 2014), i.e., when a user lacks the knowledge to construct an optimal query directly specifying his/her information need and instead he/she uses a large number of less related terms describing the need indirectly and/or vaguely (cf., e.g., *white part of eye turned green* vs. *scleral icterus* vs. *jaundice*). Insisting on “literal” translation of all the words is probably not necessary if a more appropriate translation exists.

[30] `http://hdl.handle.net/11858/00-097C-0000-0022-AAF5-B`
[31] `https://github.com/ufal/mtmonkey`

All these issues suggest that focusing on extrinsic evaluation of MT in the CLIR setting is more reasonable. The paper **Pecina et al. (2014)** first focused on techniques to adapt MT to increase translation quality (measured by standard MT evaluation metrics) and then explored methods for MT adaptation to improve effectiveness of the entire pipeline for cross-lingual IR (measured by standard IR evaluation metrics). The MT experiments focused on in-domain training and tuning, intelligent training data selection (pseudo-in-doman), optimization of phrase table configuration, compound splitting, and exploiting synonyms as translation variants. In the IR part, we experimented with morphological normalization and with using multiple translation variants for query expansion. The experiments were performed and thoroughly evaluated on three language pairs: Czech–English, German–English, and French–English.

3.5.1 Machine translation of medical queries

First, we focused on the experiments evaluated for the traditional translation quality. The MT system was again based on Moses with technical details not very different from those used in our previous experiments, both in terms of Moses parameter setting and data selection for training. For development purposes and the (intrinsic) evaluation of MT quality, we exploited the Khresmoi data sets introduced in **Urešová et al. (2014)** and additional data was extracted from the translation memory produced by the European Centre for Disease Prevention and Control (ECDC)[32]. The details of all resources can be found in the paper **Pecina et al. (2014)**.

For a better comparison of the experiment results, the size of training data was limited to maximum of 10 million parallel sentence pairs for training translation tables and 30 million monolingual sentences for training the language models. The evaluation was performed on the Khresmoi query set and the main results are shown in Table 7. The baseline systems were trained on random samples taken from all available general-domain resources and tuned on a set of sentences from news articles (for a better comparison). The BLEU scores ranged from 23.03 for German-English, 26.59 for Czech-English to 32.67 for French-English translation (see Table 7, system *baseline*). The **in-domain tuning**, which turned to be very effective in our previous experiments, improved the BLEU scores by 6.80 on average (system *+tuning*).

We explored several possible ways for **parallel training data selection** and the optimal solution turned to be using all in-domain dictionary data plus an intelligent selection based on Moore and Lewis (2010) applied to all remaining data (in-domain and general-domain), system denoted as *+transl. model*). The lesson learned here is not to trust the explicit domain classification of each data source but rather to score each sentence for being from the domain or not. The same approach applied for **monolingual training data selection** pushed the BLEU scores by additional 7.61 absolute points on average. This is 14.77 in terms of absolute improvement (56% relative) compared to the baseline (system *+lang. model*).

A specific attention in our experiments was given to the English–Greek system and the specific problem of German to create long **compound nouns** which consists of multiple regular words joined together and forming a complex concept. For example, *Raucherentwöhnungsprogramm* consists of three individual parts (*Raucher*, *entwöhnung*, and *programm*). Such words pose a major problem not only for MT. They increase the vocabulary size and often are unrecognized (treated as out-of-vocabulary) which harms the overall performance of the methods. Increasing the size of the training data, in principle, cannot solve the problem because the compounds are created ad-hoc as needed. The rich medical terminology makes this issue even stronger (especially in query translation). The problem was tackled using a tool which automatically splits compound words into components based on occurrence statistics of the compounds and their components. This step is applied on the training data prior

[32] http://ipsc.jrc.ec.europa.eu/

system	Czech–English			German–English			French–English		
	BLEU	1-PER	HUM	BLEU	1-PER	HUM	BLEU	1-PER	HUM
baseline	26.59	55.25	23.91	23.03	54.76	29.31	32.67	65.73	17.05
+tuning	35.73	66.21	–	29.50	60.40	–	37.84	71.78	
+transl. model	36.65	68.23	–	34.67	**64.03**	–	**42.74**	**76.47**	–
+lang. model	⋆**41.45**	⋆**71.61**	45.83	**40.65**	**65.43**	37.63	⋆**44.50**	⋆**77.24**	⋆56.06
+compounds	–	–	–	⋆**40.82**	⋆**67.75**	37.78	–	–	–
+synonyms	**40.38**	**70.86**	–	**37.52**	**65.52**	–	**42.53**	**76.61**	–
Google	**40.65**	**70.50**	⋆56.47	**38.64**	**66.13**	⋆54.39	**42.95**	**76.01**	45.45
Bing	27.54	51.25	–	35.25	61.88	–	36.44	71.39	–
reference	–	–	74.16	–	–	80.29	–	–	84.34

Table 7: Intrinsic evaluation of translation quality. Outputs of selected systems are compared with translations by public web-based systems (Google, Bing) using automatic evaluation measures (BLEU, 1-PER) and human evaluation (HUM). The figures in bold are those which are not statistically significantly different from the best score in each column (denoted by ⋆).

to training and on any text to be translated during the test phase. This method lead to a minor improvement only (BLEU increased by 0.17 points absolute, which is not statistically significant), the resulting scores are shown in Table 7 (system denoted as *+compounds*).

Some of the experiments described in **Pecina et al. (2014)** were not successful at all. For example, we attempted to decrease the morphological variance on the source side by replacing the word forms by their lemmas, stemms, or prefixes (both in the training and test phases). Such **morphological preprocessing** reduces the data sparsity problem but it also brings some ambiguity to the translated text which might be difficult (or impossible) to resolve by the MT system especially for longer sentences. Since our data was short queries, this might not have been an issue. However, none of the mentioned methods brought any improvement. Moreover, for the two morphologically richer languages, the translation quality degraded significantly. In another experiment, the target side of the translation tables was enriched by **adding synonyms for morphological terms** extracted from the UMLS metathesaurus and hoped that the language model would better decide on the optimal variant. This also turned out not to be useful. This step probably brought too much variance into the translation process and the automatic evaluation with respect to one reference translation test set did not show any improvement. Results of these systems are denoted as *+synonyms* in Table 7.

Result analysis

In Table 7, the results of our experiments are compared with results of two freely available MT services on the web: Google Translate[33] and Microsoft Bing Translator[34]. In terms of BLEU, for all language pairs our systems performed better – although not by a statistically significant difference (ranging between 0.80–2.18 points). In terms of PER, the results are similar. The (inverse) PER scores of the baseline systems have increased by 13.62 points absolute on average. The respective relative improvement was 23.61% and Google Translate was outperformed by 1.11–1.62 points.

In addition to the standard automatic evaluation of MT quality, we also compared the results by two kinds of manual evaluation. First, we wanted to compare the results of automatic evaluation (in terms of BLEU and PER) with human judgments, which may not correlate well (e.g., Callison-Burch et al., 2012). Based on the previous results, we selected several systems and for each language pair, human experts were asked to rank outputs of these systems and

[33] `http://translate.google.com/`

[34] `http://www.bing.com/translator/`

C	4	3	2	1	0	Σ_{base}
4	38.8	*3.6*	*1.7*	*0.4*	*0.1*	44.6
3	**4.4**	14.5	*1.6*	*1.6*	*0.5*	22.6
2	**4.9**	**3.7**	5.6	*1.2*	*0.2*	15.7
1	**1.1**	**1.4**	**0.0**	1.1	*0.0*	3.6
0	**3.7**	**2.1**	**0.5**	**0.5**	4.7	11.4
Σ_{best}	52.9	25.3	9.4	4.7	5.5	

Table 8: Results of manual translation-error analysis for the Czech–English translation. The figures (in %) represent joint and marginal (Σ) distributions of the translation quality categories 0–4 (C) in the baseline (rows) and best system (columns) translations.

the reference translations for 100 randomly sampled queries (presented in a random order) according to a descending translation quality (ties allowed). The output was transformed to pairwise comparison and presented as a percentage of cases when the translation of a particular system was judged as better than outputs of the other systems with ties ignored (this approach was proposed by Bojar et al., 2011). The results are shown in Table 7, in the columns denoted as HUM. The scores did not correlate well with the scores of the automatic measures. Google Translate was outperformed by the system denoted as *+lang. model*) only for FR–EN. However, these differences are not really significant, since the proportion of ties in all pairwise contests is more than 72% (i.e., for each pair of systems, the two systems were judged as beeing of equal quality in more than 72% of the queries on average).

The second kind of manual evaluation involved a detailed manual analysis of the results of the best-performing systems in comparison with the baselines. For each language pair, medical experts evaluated the translations produced by the two systems using this scale:

4 – perfect translation, identical to the reference;
3 – perfect translation, different from the reference;
2 – acceptable translation, errors allowed in morphology, word order, and stopwords;
1 – bad translation, no untranslated words;
0 – bad translation, some words remain untranslated.

This scale allowed to assess the overall translation quality and to analyze the improvement of the best systems over the baselines. Table 8 shows the results for CS–EN, results for the other languages are similar and can be found in the paper **(Pecina et al., 2014)**.

The baseline system did not performed badly, almost one half (44.6%) of the Czech test queries were translated to fully match the reference translations (category 4). An additional quarter (22.6%) was also judged as perfect but different from the reference (category 3). Such cases are not completely matched by the automatic evaluation measures (such as BLEU) and their scores are therefore undervalued. Further 15.7% of the translations were not judged as fully correct, but were considered adequate for querying in IR (category 2), where stopwords and word order are typically ignored and morphological variants neglected (through lemmatization or stemming). Such cases included, e.g.: *potravinová alergie* (*food allergy*) translated as *food allergies* (error in number), *chirurgické odnětí dělohy* (*hysterectomy*) translated as *surgical womb removed* (error in syntax), *rodičovský* (*parental*) translated as *parent* (error in part-of-speech), *růstový faktor hepatocytů* (*hepatocyte growth factor*) translated as *growth factor hepatocytes* (error in word order). The remaining 15% translations were completely wrong (category 1 and 0) and 3/4 of them contained one or more untranslated words (category 0).

In comparison with the best system, an improvement was observed in 22% of the translations (sum of the figures in bold in Table 8) and a degradation in 11% (sum of the figures in italics). Considering the bad baseline translations (category 1 and 0), 55% of them were improved and translated as perfect (category 4 and 3). In less than 3% of cases, the opposite behavior was observed. The best system for CS–EN was estimated to produce perfect results (category 4 and 3) in 78% and only 11% are judged as bad (category 1 and 0). This seems quite promising for the application in IR discussed in the next section.

3.5.2 Adapting query translation for cross-lingual information retrieval

The traditional notion of translation quality based on adequacy and fluency was the main evaluation criterion in the experiments described so far. The SMT systems were tuned towards such quality measured by BLEU with respect to a single reference translation. In the experiments presented in this section, the translation quality was tested extrinsically in the context of cross-lingual information retrieval where the evaluation criteria focused on retrieval quality rather than translation quality.

An CLIR system has typically two components: an MT component translates an input query to the language of documents and the translated query enters an IR component which retrieves the most relevant documents. In the most trivial approach, the MT component is a standard MT system, blindly adopted for this purpose and treated as a black box. However, MT can be adapted for this specific purpose (application), modified and more tightly integrated with the IR component. In practice, most IR systems operate on "term level", ignore word order and also other linguistic properties (letter casing, morphological variants, etc.) and over the years, people have become accustomed to communicating with IR systems in a "keyword language", which lacks these properties. Queries are generally not fluent and grammatically correct, they do not form complete sentences which a standard MT system is trained to translate.

Adapting MT for CLIR can be done on various levels. One possibility is to take this into account and eliminate the information in translations which is not really important for the IR component (e.g., word order, morphological variants). Another possibility is to enrich the translations by information which can the IR component benefit from. For example, inclusion of words which are synonymous or semantically related to the query or assigning weights to query terms. In this section, we present several methods to integrate MT and IR components in a CLIR pipeline and evaluate their effectiveness. First, we will describe the experimental data and the IR system and then present the experiments and their evaluation.

Experimental settings

The IR experiments were carried out on the CLEF 2013 eHealth Task 3 test collection which contains around one million documents (web pages) related to medicine and 50 English medical queries with corresponding relevance assessments (Suominen et al., 2013). The documents were crawled by the Khresmoi project mostly from health and medicine websites certified by the Health on the Net Foundation[35] and from other commonly used health and medicine websites.

The queries aimed to model typical queries by laypeople (e.g., patients) formulated to find out more about their disorders after they have examined their discharge summary. The discharge summaries were taken from the anonymized clinical free-text notes of the MIMIC II database[36]. The queries were relatively short with an average length of about two or three terms (words) in the titles, for example: *facial cuts and scar tissue*; *asystolic arrest*; *nausea and*

[35] `http://www.hon.ch`

[36] `http://mimic.physionet.org/`

vomiting and hematemesis; *sinus tachycardia*; *chills and gallstones*). The complete specification also included a description (details of what the query means) and narrative (expected content of the relevant documents), but those were used for relevance assessment only. The titles (originally in English) were manually translated for the purpose of our work by medical professionals into German, French, and Czech and the translations were used to query the CLIR system. This set of parallel data will be further denoted as CLEF test data.

The **relevance assessment** was performed within the shared task by experts. Top 10 documents for each query and each evaluated run were assessed which resulted in 6,391 judged documents out of which 1,878 documents were found to be relevant. It should be emphasized that the resulting pools are rather shallow and other relevant documents may have been missed and not assessed. In general, such documents are treated as not relevant and might bias the evaluation. Details can be found in Goeuriot et al. (2013).

The **IR system** employed in our experiments was based on the BM25 retrieval model (Robertson et al., 1998) implemented in Lucene 3[37] which had been previously demonstrated to perform well in medical IR (Leveling et al., 2012). The documents from the test collection were preprocessed to extract textual content of the web pages which removed HTML markup and JavaScript, leaving raw textual content which was flattened into single index. The text was transformed to lower case and Salton's (1971) stopword list was used to identify words which were not indexed. Stemming of queries and documents was performed using the English Snowball stemmer provided in Lucene, which is based on the Porter algorithm (Porter, 2001).

MT evaluation

Most of the translation experiments described in Section 3.5.1 improved the baseline results on the Khresmoi query test sets. Our best systems even outperformed the state-of-the-art publicly available on-line services (in automatic evaluation, and for FR–EN in human evaluation as well). Even if we now want to focus on retrieval quality of the entire CLIR pipeline rather than translation quality of the MT component (extrinsic evaluation), we first analyze the **translation quality on the CLEF test sets** (the 50 CLEF 2013 eHealth Task 3 queries).

The results are tabulated in Table 9 using the same evaluation measures as in Table 7. In contrast, the bold font now indicates the scores that are significantly better than the baseline (in Table 7, the bold font referred to the results statistically indistinguishable from the best ones). The best scores for each language pair are again indicated by the ⋆ symbol. The baseline systems are the same as in the previous section, trained and tuned on general-domain data. The adapted systems are those trained and tuned on domain-specific data using the optimal configuration as in the previous section. The label *+compounds* denotes the systems employing the German decompounding method, *+synonyms* denotes the systems with target side enriched by synonyms, and *+PER* refers to a new configuration which is based on the "adapted" system tuned on the query development sets by MERT (Och, 2003) but optimizing PER instead of BLEU (this experiment will be discussed later).

The first important note is that these results (obtained on the CLEF test set) are not as reliable as those presented in Table 7 (Khresmoi test set). Due to the size of the CLEF test set (50 queries), the sample variance and confidence intervals for BLEU and PER are much larger and it is not possible to reliably measure difference of the methods under comparison. Also, the Khresmoi and CLEF test sets cannot be considered completely comparable for another reason – although both comprise medical queries, the CLEF test sets were created in a more controlled setting and the queries are more thought-out and fluent.

[37] `http://lucene.apache.org/`

system	Czech–English			German–English			French–English		
	BLEU	1-PER	HUM	BLEU	1-PER	HUM	BLEU	1-PER	HUM
baseline	47.01	66.41	26.56	39.52	62.47	17.95	39.20	71.48	12.09
adapted	40.91	70.26	28.57	42.95	**64.19**	36.36	**52.96**	**76.69**	55.56
adapted+compounds	–	–	–	43.42	**66.41**	–	–	–	–
adapted+synonyms	47.60	71.66	–	40.19	**62.58**	–	**54.50**	**77.19**	–
adapted+PER	52.28	**75.06**	–	**50.24**	**68.91**	–	**51.01**	**80.05**	–
Google	⋆**56.02**	⋆**77.30**	⋆75.93	**54.53**	**75.78**	⋆71.43	⋆**61.99**	⋆**83.02**	⋆66.67
Bing	47.46	67.06	–	⋆**59.72**	⋆**76.54**	–	**58.34**	80.79	–
reference	–	–	72.31	–	–	78.87	–	–	76.92

Table 9: Translation quality of selected MT systems measured on the CLEF test data. The bold font denotes results that are significantly better than the baseline. The ⋆ symbol indicates the best score for each language pair and measure. Reference is the original version in English.

The overall translation results on the CLEF test sets did not really confirm the results obtained on the Khresmoi test sets. The basic adapted system outperformed the baselines for DE–EN and FR–EN only. In case of FR–EN, the improvement was statistically significant in terms of both BLEU and PER. In case of DE–EN, the improvement was significant measured by PER only. For CS–EN, the baseline scores were surprisingly high and performance of the adapted system dropped. Decompounding applied to German helped. Exploiting synonyms helped for CS–EN and FR–EN, but in all experiments, the increase was not statistically significant. Google Translate, however, dominated all our systems for all language pairs measured by all three measures. Although for DE–EN, the absolute winner is Microsoft Bing Translator, none of these results are significantly better than those achieved by our best systems. Interestingly, the HUM score for CS–EN suggests that the output of Google Translate is better than the reference. However, this is caused by including comparison with the other systems (adapted), in which Google Translate is judged as better more often than the reference.

IR evaluation

In IR, evaluation is based on ability to rank documents with respect to their relevance to a given query (information need). The system which ranks relevant documents higher than the irrelevant is scored better. We evaluated IR performance by standard IR evaluation measures computed with the TREC evaluation tool[38]. The measures include Precision at a cut-off of 10 documents (P@10) and Mean Average Precision (MAP) (Voorhees and Harman, 2005). The cross-lingual MAP scores are also compared with the monolingual ones, i.e., those obtained by using the reference (English) translations (MAP$^{rel}_{EN}$) to see how the system would perform if the queries were translated perfectly. P@10 is taken as the main evaluation measure, since this task is more oriented towards precision than towards recall. Scores of MAP are probably underestimated due to the rather shallow relevance assessments. The paper also presents scores of Normalized Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002) at top 10 documents.

The main results of the IR evaluation are tabulated in Table 10. Significance of the results is indicated with respect to the baseline using the standard Wilcoxon signed rank test (Hull, 1993, $p < 0.05$); those which are significantly better are typed in bold and those which are significantly worse are typed in italics. The best cross-lingual results are marked with the ⋆ symbol. The monolingual scores are quite comparable to the best results achieved by the CLEF eHealth 2013 task participants (Goeuriot et al., 2013) and the cross-lingual ones achieved as much as 86% of the monolingual ones (in terms of MAP) which is a very good result, although

[38] `http://trec.nist.gov/trec_eval/`

system	Czech–English			German–English			French–English		
	P@10	MAP	MAP^{rel}_{EN}	P@10	MAP	MAP^{rel}_{EN}	P@10	MAP	MAP^{rel}_{EN}
baseline	34.8	24.28	80.00	29.4	19.02	62.67	31.0	21.87	72.06
adapted	37.2	23.67	77.99	32.8	21.85	71.99	**38.4**	**26.33**	86.75
adapted+PER	35.4	23.16	76.31	35.0	22.62	74.53	**38.4**	**25.94**	**85.47**
adapted+stemming	28.2	20.27	66.79	23.4	16.29	53.67	32.0	20.33	66.99
adapted+n-best	31.4	22.71	74.83	21.8	*16.19*	*53.34*	29.6	21.44	70.64
Google	⋆38.4	⋆**25.97**	⋆**85.57**	**37.0**	**23.22**	**76.51**	⋆**40.6**	**26.74**	**88.11**
Bing	32.6	22.76	74.99	⋆**38.8**	⋆**25.09**	⋆**82.67**	**40.2**	⋆**27.57**	⋆**90.84**
reference	**47.0**	**30.35**	100.00	**47.0**	**30.35**	100.00	**47.0**	**30.35**	100.00

Table 10: Extrinsic evaluation of translation quality. IR results for query translations produced by various MT systems compared with the original (reference) queries in English. The scores typed in bold, normal, and italics are significantly better, equal, and worse (respectively) than the baseline. MAP^{rel}_{EN}refers to MAP relative to the monolingual performance (reference).

the commercial systems reached as much as 90%. Google Translate outperformed Microsoft Bing Translator on CS–EN and FR–EN and is on par with it on DE–EN. In our experiments, the highest baseline scores were interestingly observed for the CS–EN translation direction although Czech is typically harder to translate than German and especially French. On this particular test set, however, we experienced the opposite which is probably due to the randomness of training data selection for the baseline system. For the CS–EN language pair, it must have contained more material relevant to this data set than for the other language pairs.

In addition to the MT systems described so far, we also investigated three new configurations aiming to improve retrieval quality – one producing translations focused on adequacy rather than fluency (*adapted+PER*), one for producing stemmed translations (*adapted+stemming*), and one for exploiting multiple translation options (*adapted+n-bets*):

Systems *adapted+PER* are based on the *adapted* systems with **tuning towards PER** instead of BLEU. As it was discussed earlier, PER ignores word order when comparing an MT output with its reference (any permutation of words in the translation are scored equally) which implies more focus on translation adequacy and less focus on fluency (compared to BLEU). This indeed confirmed to be positive for query translation in a CLIR pipeline, where word order does not matter. As shown in Table 9, in MT-focused evaluation by PER, these systems outperformed all our other systems. In IR-focused evaluation, with the exception of CS–EN (explained above), the PER-tuned systems are also superior to other systems which confirms our earlier hypothesis of a better fit of PER for tuning MT for query translation in CLIR.

Stemming as a standard technique was used in all our IR experiments for indexing the documents. This required the translated queries to be also stemmed (before entering the IR component). As an alternative, the stemmed queries were produced directly during translation (systems *adapted+stemming*). This was achieved by stemming the target language side of the parallel training data as well as the monolingual data for language models. In this experiment, stemming was also performed on the source language side to reduce the morphological complexity (which is important especially for Czech). The Porter's Snowball stemmer was employed for the source side languages (Czech, German, French) and the original Porter's stemmer for English (Porter, 1980). The results, however, were not affirmative since degradation in the retrieval performance was also observed when stems were used instead of full word forms on both source and target side. Training MT on stemmed words probably introduced too much ambiguity, which hurts not only the MT quality but also the IR performance.

In the final experiments, the advantage of SMT to produce multiple translation options for a given source text was exploited. Usually, only the highest-scored option is considered as translation. However, from the IR point of view, the other (lower-scored) hypotheses might contain useful information too. We explored an idea where the final query is constructed as a **merging multiple translation options** (the top n ones where n is a parameter of the method). This was first presented by Nikoulina et al. (2012) as a way of query expansion and was claimed to be successful. Several values of n were tested but none of them confirmed the previously reported findings and did not improve the results. A possible reason for the decrease is that the translation variants differ to such an extent that they cannot be considered good translations and therefore the queries are expanded by non-relevant terms.

Conclusions

The discussed publication **(Pecina et al., 2014)** concluded our work on the Khresmoi project. It provided the first comprehensive and rigorous assessment of the contribution of domain-based MT adaptation as well as IR-targeted MT adaptation to real user queries (focused on health and health-related problems). Even though some of our findings illustrate that certain traditional techniques for MT do not improve translation quality in this specific domain and for this specific purpose, our overall results are very positive.

In terms of translation quality, most of the adaptation techniques substantially outperformed our baseline and even the state-of-the-art on-line MT systems. However, when applied to the cross-language IR task, the positive MT results did not directly translate to improvements over the state-of-the-art in MT and one of the main findings is that MT quality and IR quality do not correlate in a straightforward way. However, the size of the CLEF test data for IR is much smaller than the Khresmoi test data for MT and 50 queries are probably not sufficient to reflect the qualitative changes in the MT components.

The most promising results were obtained by SMT tuning for PER (Position-independent word Error Rate) which confirmed the hypothesis that tuning towards the standard translation quality is not optimal. However, the IR results based on the baseline translations were significantly improved by using the adapted translations for the FR–EN direction only. However, none of our translations did outperform the system using Google Translate's translation. The CS–EN system did outperform Microsoft Bing Translator, though.

3.6 Reranking query translations for cross-lingual information retrieval (Saleh and Pecina, 2016c)

Our recent activity in the area of adaptation of machine translation to cross-lingual information retrieval has focused on reranking of query translation hypotheses. This research was mainly conducted within the KConnect project[39] with the goal to further improve the query translation process for CLIR in the medical domain. As it was stated earlier in this thesis, the decoding process in SMT is optimized towards translation quality (measured by, e.g., BLEU or PER) which is not be optimal from the retrieval point of view. Our recent work exploits multiple translations produced by SMT which are then reranked using a discriminative machine-learning method trained to optimize the retrieval quality directly. Results of this research were presented at the main session of CLEF 2016 **(Saleh and Pecina, 2016c)**.

Our approach is motivated by the work of Nikoulina et al. (2012). They reranked 1 000 best translations hypotheses obtained by a phrase-based SMT system for each non-English query to select the best (English) translation which was then searched for in an English document collection. The reranking model was based on a (weighted) linear combination of features

[39] `http://www.kconnect.eu/`

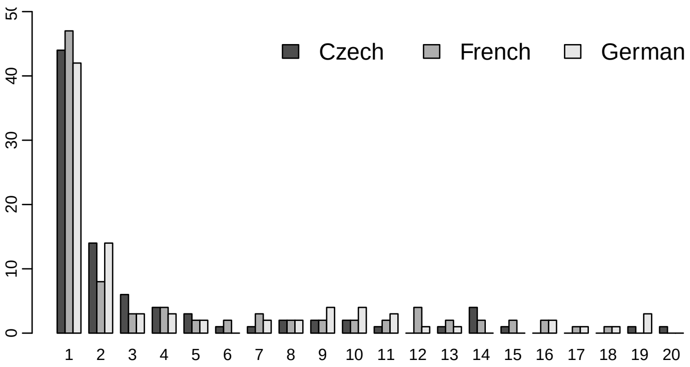

Figure 6: Histogram of ranks of translation hypotheses with the highest P@10 for each training query.

which included scores of the SMT model features and some syntactic and morphological features obtained from a linguistic analysis of the translation. The parameters of the model were estimated by MIRA (Margin Infused Relaxed Algorithm, Crammer and Singer, 2003) on training data which contained translation hypotheses of training queries each scored using MAP obtained as from evaluation retrieval results of the translation used as a query on the document collection with relevance assessments. Nikoulina et al. (2012) tested their approach on the CLEF AdHocTEL 2009 task (50 topics in German and French) and showed only a moderate improvement between 0.010–0.025 MAP points (not tested for statistical significance).

Our experiments were conducted on the data from the CLEF eHealth Lab series in 2013–2015, which included a total of 166 queries in English manually translated to Czech, German, and French. The queries were mixed and randomly split into 100 queries for training and 66 queries for testing to avoid overfitting. Additional relevance assessment was performed on all unjudged documents appearing among top 10 documents in each experiment (i.e., all the results are fully assessed). The oracle experiment (see Figure 6) performed on the training set showed that, from the IR perspective, the best translation often appears among 20 best hypotheses provided by the SMT system (the Khresmoi Translator). Therefore, we rerank only 20 best translations for each query. Each translation hypothesis in the training data is scored using the difference between its P@10 score and the best possible (oracle) P@10 score for that particular query (P@10 is used as the main evaluation measure as in the official CLEF eHealth tasks). The reranking model is a linear regression model implemented in R[40] to predict the P@10 score of each query translation hypothesis in the test data. The highest-scored hypothesis is then used to perform the retrieval step. The training data for all three languages were joined together, i.e., we built one single reranking model for all the languages. This approach proved better than building three independent language-specific models (it increased the training data size which probably helped to train a more robust model).

The feature set consists of the SMT features provided by Moses (see Section 2.2) and various additional features including: document collection features (term weights extracted from the test collection, e.g., inverse document frequency), blind-relevance feedback features (term weights extracted from documents retrieved by the translation hypothesis used as a query from the test collection), translation pool features (term weights extracted from all translation hypotheses of the query), Wikipedia features (term weights obtained from titles and abstracts of Wikipedia articles retrieved by the translation hypothesis used as a query from the English Wikipedia dump), UMLS features (number of UMLS concepts found in the translation hypothesis), retrieval status value (the value of the highest scored document retrieved from the test collection by the translation hypothesis).

[40] https://www.r-project.org

system	Czech–English			French–English			German–English		
	P@10	NDCG@10	MAP	P@10	NDCG@10	MAP	P@10	NDCG@10	MAP
Mono	0.5030	0.4995	0.2997	0.5030	0.4995	0.2997	0.5030	0.4995	0.2997
Baseline	0.4561	0.3857	0.2358	0.4773	0.4111	0.2572	0.4242	0.3647	0.2274
RR-SMT	0.4470	0.3792	0.2477	0.4879	0.4285	0.2581	0.4273	0.3788	0.2265
RR-ALL	***0.5015***	**0.4072**	***0.2573***	***0.5106***	***0.4649***	***0.2786***	***0.4530***	**0.3947**	**0.2371**
Google	0.5091	0.3998	0.2693	0.4970	0.4388	0.2636	0.4939	0.4277	0.2687
Bing	0.4788	0.4051	0.2522	0.4864	0.4275	0.2643	0.4652	0.4169	0.2504

Table 11: Results of reranking applied to SMT features (*RR-SMT*) and to all available features (*RR-all*) compared to the results of systems based on single best SMT translations (*Baseline*), translations by free MT services (*Google*, *Bing*), and results of monolingual retrieval (*Mono*).

The main results of our experiments (obtained on the test set of 66 queries and all three source languages) are displayed and compared to several other systems in Table 11. The baseline system employed the highest-scored hypothesis provided by the SMT system. Google Translate and Microsoft Bing Translator denote systems exploiting translations from two public MT services. *Mono* refers to the monolingual system which uses the reference (English) translations. First, we tested the effect of our method applied to the SMT features only (system *RR-SMT*). This system outperformed the baseline system only for French but the difference was not statistically significant. However, the system employing all the features sketched above (*RR-all*) performed very well. Measured by P@10, not only did it outperform the baseline for all languages by a statistically significant difference but it also outperformed the system which used query translations from French obtained by Google (though this difference was not statistically significant).

More detailed results are presented in Figure 7. The plot shows how P@10 changes for each query in the test set if the best system is compared with the baseline. Positive values refer to improvement (increase of P@10) which was observed for approximately 20% of queries. Negative values denote degradation, which was observed in about 3 cases for each language, on average. A good example of a query whose translation was improved is query 2015.11 (reference translation: *white patchiness in mouth*). The baseline translation from Czech *white coating mouth* improved to *white coating in oral cavity* (P@10 increased from 0.10 to 0.80) and the baseline translation from French *white spots in the mouth* improved to *white patches in the mouth* (P@10 increased from 0.10 to 0.70).

The presented approach confirmed to be very successful. Its main advantages are the ability to combine various kinds of features and the fact that the trained models can be applied to new languages. A disadvantage is that it requires training data which might be difficult to obtain. For medicine, however, the queries from the CLEF eHealth tasks seem sufficient.

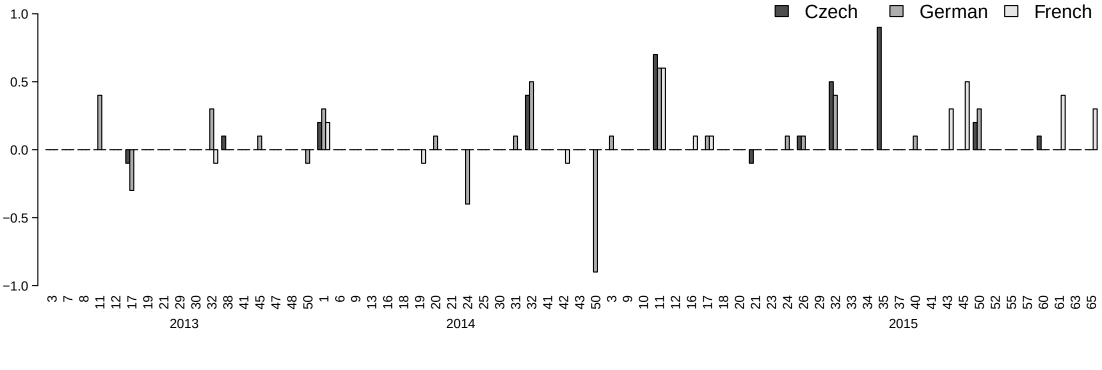

Figure 7: Per-query results on the test set. The bars represent absolute difference of P@10 of the best system (ALL) and the baseline system for each query and each language.

4 Conclusions and Final Remarks

In this part of the thesis, we attempted to provide a concise and coherent overview of our research in the area of SMT adaptation. We presented complete background information including an overview or related work, reviewed our achievements in this field, and put them into the context of each other and the three research projects which mainly funded this research.

The selected publications form Part II of this thesis. However, there are also other important publications which are related to our work in this area. Already in 2007, we participated in the shared task competition on boilerplate removal from HTML pages organized by SIGWAC (The Special Interest Group of the Association for Computational Linguistics on Web as Corpus) and placed first with our system based on conditional random fields (Marek et al., 2007) which was further enhanced and described in Spousta et al. (2008). In Murray et al. (2006a,b) and Lin et al. (2009), we described our effort to translate large-scale ontology from the Malach project[41] to allow cross-lingual search in the Malach archive of Holocaust survivors' testimonies. In Homola et al. (2009) and later in Libovický and Pecina (2014), we proposed two MT quality metrics, based on the standard BLEU modified to better reflect quality of translation into morphologically richer languages (such as Czech). In Spoustová et al. (2010), we described our effort towards acquisition of large web-based corpus of Czech. In Tamchyna et al. (2013), we presented the MTMonkey framework for providing MT as a web service which has been already adopted by a number of other projects. Saleh and Pecina (2014), Saleh et al. (2015), and Saleh and Pecina (2016b) are working notes describing our participation during three years of the CLEF eHealth evaluation lab focused on multilingual user-centred health information retrieval which we also helped to organize. In Saleh and Pecina (2016a), we presented experiments with the query reraning method and new languages. In Bojar et al. (2014, Section 5), we presented the results of the medical translation shared task, a part of the WMT evaluation campaign which we organized in 2014. Finally, in a joint work Toral et al. (2015), we proposed a modification of the state-of-the-art method for training data selection based on word-level linguistic features, such as lemmas, named entities, and part-of-speech tags which was shown to improve language model perplexity especially for morphologically rich languages.

Adaptation of SMT is a large research area actively studied in the recent years with an increasing number of real-world applications. Domain adaptation focuses mainly on effective exploitation of in-domain resources, whilst application adaptation goes usually deeper into the system and changes its functioning to provide optimal output for downstream applications. Our research in the first strand focused on data availability and analysis of various adaptation methods exploiting the acquired resources. Our future research in this area will focus on on-line techniques which change system parameters or switch component models (language and translation models) dynamically for each document or sentence to translate. In the second strand, our effort focused on modifying MT systems to provide optimal translation of search queries in cross-lingual information retrieval. Here, a lot of open questions and ideas to explore certainly remain. The most promising ones include SMT tuning directly towards retrieval quality and a better way of exploiting translation variants and the modern neural network methods for SMT will also provide new opportunities of research in this area.

[41] http://malach.umiacs.umd.edu/

Acknowledgements

The work presented in this thesis was mainly supported through three projects funded by the European Union in the 7th Framework Programme and Horizon 2020: Panacea (grant agreement no. 248 064), Khresmoi (grant agreement no. 257 528), and KConnect (grant agreement no. 644 753). Panacea focused on language resources for machine translation, including acquisition, production, and maintenance and developed a set of tools for easy acquisition of domain-specific textual data from the web. Our part in this project was conducted at the Centre for Next Generation Localisation, Dublin City University, Ireland in 2010–2012. Khresmoi and its successor KConnect focused on information retrieval and analysis in the medical domain and developed a multi-lingual multi-modal search and access systems for biomedical information and documents. Our work on these two projects was carried out at the Institute of Formal and Applied Linguistics, Charles University in Prague, Czech Republic during 2012–2016. An additional support came from the Czech Science Foundation through the Center for large-scale multi-modal data interpretation (CEMI, project no. P103/12/G084). Some of the language resources used in this work are hosted in the LINDAT/CLARIN repository, funded by the Ministry of Education Youth and Sports the Czech Republic (project no. LM2010013).

References

Al-onaizan, Y., Cuřín, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, Final Report, JHU Summer Workshop.

ALPAC (1966). Language and machines: Computers in translation and linguistics. Technical report, National Academy of Sciences, Washington, DC, USA. Publication 1416.

Appelo, L. and Landsbergen, J. (1986). The machine translation project Rosetta. In Gerhardt, T., editor, *I. International Conference on the State of the Art in Machine Translation in America, Asia and Europe: Proceedings of IAI-MT86*, page 34–51. IAI/EUROTRA-D.

Aswani, N., Beckers, T., Birngruber, E., Boyer, C., Burner, A., Bystroň, J., et al. (2012). Khresmoi: Multimodal multilingual medical information search. In Mantas, J., Andersen, S. K., Mazzoleni, M. C., Blobel, B., Quaglini, S., and Moen, A., editors, *Proceedings of the 24th International Conference of the European Federation for Medical Informatics, Quality of Life through Quality of Information, Village of the future*, Pisa, Italy.

Aswani, N., Beckers, T., Birngruber, E., Boyer, C., Burner, A., Bystroň, J., et al. (2013). Khresmoi – multilingual semantic search of medical text and images. In *MEDINFO 2013: Proceedings of the 14th World Congress on Medical Informatics*, volume 192 of *Studies in health technology and informatics*, Copenhagen, Denmark.

Attar, R. and Fraenkel, A. S. (1977). Local feedback in full-text retrieval systems. *Journal of Association for Computing Machinery*, 24(3):397–417, Association for Computing Machinery.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 355–362, Edinburgh, United Kingdom.

Axelrod, A. E. (2006). *Factored Language Models for Statistical Machine Translation*. PhD thesis, University of Edinburgh.

Azarbonyad, H., Shakery, A., and Faili, H. (2012). Using learning to rank approach for parallel corpora based cross language information retrieval. In *ECAI 2012: 20th European Conference on Artificial Intelligence*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 79–84.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. *SIGIR Forum*, 31:84–91, Association for Computing Machinery.

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71, Melbourne, Australia.

Banerjee, P., Du, J., Li, B., Naskar, S., Way, A., and van Genabith, J. (2010). Combining multi-domain statistical machine translation models using automatic classifiers. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, page 141–150, Denver, Colorado, USA.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2011). Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. In *Proceedings of the Machine Translation Summit XIII*, page 285–292, Xiamen, China.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2012). Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data? In *Proceedings of the 16th Annual Meeting of the European Association for Machine Translation*, page 169–176, Trento, Italy.

Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J. (2013). Quality estimation-guided data selection for domain adaptation of smt. In *Proceedings of the XIV Machine Translation Summit*, page 101–108, Nice, France.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65–72, Ann Arbor, Michigan.

Barbosa, L., Rangarajan Sridhar, V. K., Yarmohammadi, M., and Bangalore, S. (2012). Harvesting parallel text in multiple languages with limited supervision. In *Proceedings of the 24th International Conference on Computational Linguistics*, page 201–214, Mumbai, India.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: Instant domain-specific corpora to support human translators. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, page 47–252, Oslo, Norway.

Bel, N., Poch, M., and Toral, A. (2012). PANACEA (Platform for Automatic, Normalised Annotation and Cost-Effective Acquisition of language resources for human language technologies). In *Proceedins of the 16th Annual Conference of the European Association for Machine Translation EAMT'2012*, page 90, Trento, Italy.

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., and Ureš, L. (1994). The candide system for machine translation. In *Proceedings of the workshop on Human Language Technology*, page 157–162, Plainsboro, New Jerey, USA.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, page 182–189, Athens, Greece.

Bertoldi, N., Haddow, B., and Fouet, J.-B. (2009). Improved minimum error rate training in Moses. *Prague Bulletin of Mathematical Linguistics*, 91:7–16, Charles University in Prague.

Bisazza, A. and Federico, M. (2012). Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 439–448, Avignon, France.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, page 136–143, San Francisco, CA, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, JMLR.org.

Boitet, C. (1989). Geta project. In Nagao, M., editor, *Translation Summit*, page 54–65, Tokyo, Japan.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. F. (2011). A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, Elsevier Science Publishers B. V.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., and Roossin, P. (1988). A statistical approach to language translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 71–76.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, MIT Press.

Bruderer, H. E. (1977). The present state of machine and machine-assisted translation. In *Commission of the European Communities. Third European Congress on Information Systems and Networks: Overcoming the Language Barrier*, volume 1, page 529–556, Munich. Verlag Dokumentation.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and 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*, pages 17–53, Uppsala, Sweden. Revised August 2010.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (2011). Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, page 22–64, Edinburgh, Scotland.

Carl, M., Way, A., and Daelemans, W. (2004). Recent advances in example-based machine translation. *Computational Linguistics*, 30(4):516–520, MIT Press.

Carpuat, M., Daumé III, H., Fraser, A., Quirk, C., Braune, F., Clifton, A., Irvine, A., Jagarlamudi, J., Morgan, J., Razmara, M., Tamchyna, A., Henry, K., and Rudinger, R. (2012). Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*. Johns Hopkins University, Baltimore, MD.

Ceausu, A., Tinsley, J., Zhang, J., and Way, A. (2011). Experiments on domain adaptation for patent machine translation in the PLuTO project. In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation*, page 21–28.

Chen, J., Chau, R., and Yeh, C.-H. (2004). Discovering parallel text from the World Wide Web. In *Proceedings of the 2nd workshop on Australasian information security, Data Mining and*

Web Intelligence, and Software Internationalisation, volume 32, page 157–161, Darlinghurst, Australia.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Santa Cruz, California.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Ann Arbor, Michigan.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistic*, 33(2):201–228, MIT Press.

Cho, J., Garcia-Molina, H., and Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30:161–172, Elsevier Science Publishers B. V.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Clarkson, P. and Rosenfeld, R. (1997). Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of Eurospeech*, pages 2707–2710, Rhodes, Gereece.

Cline, R. J. W. and Haynes, K. M. (2001). Consumer health information seeking on the internet: the state of the art. *Health Education Research*, 16(6):671–692.

Costa-jussà, M. R., Farrús, M., and Pons, J. S. (2012). Machine translation in medicine. A quality analysis of statistical machine translation in the medical domain. In *Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas*, page 1995–1998, Žilina, Slovakia.

Crammer, K. and Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

Darwish, K. and Oard, D. W. (2003). Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 338–344, Toronto, Canada. Association for Computing Machinery.

Daumé III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics and Human Language Technologies, Short Papers*, page 407–412, Portland, Oregon, USA.

Déjean, H., Gaussier, E., Renders, J.-M., and Sadat, F. (2005). Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *Artificial Intelligence in Medicine*, 33(2):111–124, Elsevier Science Publishers Ltd.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. (2008). WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, volume 30, page 27–28, London, UK.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, page 138–145, San Diego, CA, USA.

Dong, M., Liu, Y., Luan, H., Sun, M., Izuha, T., and Zhang, D. (2015). Iterative learning of parallel lexicons and phrases from non-parallel corpora. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1250–1256. AAAI Press.

Du, J. and Way, A. (2009). A three-pass system combination framework by combining multiple hypothesis alignment methods. In *International Conference on Asian Language Processing, IALP '09*, pages 172–176, Singapore.

Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., Tamchyna, A., Urešová, Z., and Zeman, D. (2014). Machine translation of medical texts in the Khresmoi project. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, page 221–228, Baltimore, Maryland, USA.

Dziwiński, P. and Rutkowska, D. (2008). Ant focused crawling algorithm. In *Proceedings of the 9th international conference on Artificial Intelligence and Soft Computing*, page 1018–1028, Zakopane, Poland. Springer-Verlag.

Eck, M., Vogel, S., and Waibel, A. (2004a). Improving statistical machine translation in the medical domain using the Unified Medical Language System. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, page 792–798, Geneva, Switzerland.

Eck, M., Vogel, S., and Waibel, A. (2004b). Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of the International Conference on Language Resources and Evaluation*, page 327–330, Lisbon, Portugal.

Eichmann, D., Ruiz, M. E., and Srinivasan, P. (1998). Cross-language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 72–80, Melbourne, Australia.

Esplà-Gomis, M. and Forcada, M. L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. *The Prague Bulletin of Mathemathical Lingustics*, 93:77–86.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174, Tampere, Finland.

Fishel, M., Georgakopoulou, Y., Penkale, S., Petukhova, V., Rojc, M., Volk, M., and Way, A. (2012). From subtitles to parallel corpora. In *Proceedins of the 16th Annual Conference of the European Association for Machine Translation EAMT'2012*, page 3–6, Trento, Italy.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, page 451–459, Cambridge, Massachusetts, USA.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, page 128–135, Prague, Czech Republic.

Fox, S. (2011). Health Topics: 80% of internet users look for health information online. Technical report, Pew Research Center.

Gao, J. and Nie, J.-Y. (2006). A study of statistical models for query translation: finding a good unit of translation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201, Seattle, Washington, USA.

Gao, Z., Du, Y., Yi, L., Yang, Y., and Peng, Q. (2010). Focused web crawling based on incremental learning. *Journal of Computational Information Systems*, 6:9–16.

Germann, U. (2003). Greedy decoding for statistical machine translation in almost linear time. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 1–8, Edmonton, Canada.

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 228–235, Toulouse, France.

Goeuriot, L., Jones, G. J. F., Kelly, L., Leveling, J., Hanbury, A., Müller, H., et al. (2013). ShARe/CLEF eHealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In Pamela Forner, Roberto Navigli, D. T., editor, *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, Valencia, Spain.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G. J., and Müller, H. (2014). ShARe/CLEF eHealth evaluation lab 2014, task 3: User-centred health information retrieval. In *CLEF Online Working Notes, CEUR Workshop Proceedings*, volume 1180, pages 43–61, Sheffield, UK.

Goldstine, H. H. and Goldstine, A. (1946). The electronic numerical integrator and computer (eniac). *Mathematical Tables and Other Aids to Computation*, pages 97–110, JSTOR.

Guilbaud, J.-P. (1987). Principles and results of a German–French MT system at grenoble university. In King, M., editor, *Machine translation today: the state of the art*, page 278–318. Edinburgh University Press.

Haddow, B. (2013). Applying pairwise ranked optimisation to improve the interpolation of translation models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 342–347, Atlanta, Georgia.

Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201, Dublin, Ireland.

Hewavitharana, S. and Vogel, S. (2013). Extracting parallel phrases from comparable data. In *Building and Using Comparable Corpora*, pages 191–204. Springer.

Hildebrand, A. S., Eck, M., Vogel, S., and Waibel, A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, page 133–142, Budapest, Hungary.

Hollink, V., Kamps, J., Monz, C., and de Rijke, M. (2004). Monolingual document retrieval for European languages. *Information Retrieval*, 7(1-2):33–52, Kluwer Academic Publishers.

Homola, P., Kuboň, V., and Pecina, P. (2009). A simple automatic MT evaluation metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 33–36, Athens, Greece. Association for Computational Linguistics.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 329–338, Pittsburgh, PA, USA.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, Association for Computing Machinery.

Jelinek, F. (1997). *Statistical methods for speech recognition.* MIT Press, Cambridge, MA, USA.

Jimeno Yepes, A., Prieur-Gaston, É., and Névéol, A. (2013). Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):1–10, BioMed Central.

Kay, M., Norvig, P., and Gawron, M. (1992). *Verbmobil: A Translation System for Face-to-Face Dialog.* University of Chicago Press, Chicago, IL, USA.

Kelly, L., Dungs, S., Kriewel, S., Hanbury, A., Goeuriot, L., Jones, G. J., Langs, G., and Mueller, H. (2014). Khresmoi professional: multilingual, multimodal professional medical search. In *36th European Conference on Information Retrieval (ECIR 2014)*, page 13—16, Amsterdam, Netherlands.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29(3):333–348.

King, M. (1982). EUROTRA: An attempt to achieve multilingual mt. In 1982, L., editor, *Practical experience of machine translation*, page 139–147. North-Holland Publishing Company.

Koehn, P. (2004a). Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In Frederking, R. E. and Taylor, K. B., editors, *Machine Translation: From Real Users to Research*, volume 3265 of *Lecture Notes in Computer Science*, pages 115–124. Springer Berlin Heidelberg.

Koehn, P. (2004b). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, page 388–395, Barcelona, Spain.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings of the Tenth Machine Translation Summit*, page 79–86, Phuket, Thailand.

Koehn, P. (2010). *Statistical Machine Translation.* Cambridge University Press, New York, NY, USA, 1st edition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, page 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, page 48–54, Edmonton, Canada.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, page 224–227, Prague, Czech Republic.

Langlais, P. (2002). Improving a general-purpose statistical translation engine by terminological lexicons. In *COMPUTERM 2002: Second International Workshop on Computational Terminology*, page 1–7, Taipei, Taiwan.

Lehmann, W., Bennet, W., J., S., Smith, H., Fluger, S., and Eveland, S. (1981). The metal system. final technical report. Technical Report RADC-TR-80-374, Linguistics Research Center, University of Texas at Austin. NTIS report AO-97896.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: efficient MT evaluation using block movements. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 241–248, Trento, Italy.

Leveling, J., Goeuriot, L., Kelly, L., and Jones, G. J. (2012). DCU@TRECMed 2012: Using adhoc baselines for domain-specific retrieval. In *Text Retrieval Conference (TREC) 2012*, pages 1–9, Gaithersburg, MD, USA.

Libovický, J. and Pecina, P. (2014). Tolerant bleu: a submission to the wmt14 metrics task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 409–413, Baltimore, Maryland, USA.

Lin, J., Murray, C. G., Dorr, B., Hajič, J., and Pecina, P. (2009). A cost-effective lexical acquisition process for large-scale thesaurus translation. *Language Resources and Evaluation*, 43:27–40, Springer Netherlands.

Maas, H. (1987). The MT system SUSY. In King, M., editor, *Machine translation today: the state of the art*, page 209–246. Edinburgh University Press.

Macdonald, N. (1954). Language translation by machine - a report of the first successful trial. *Computers and Automation*, 3(2):6–10.

Macháček, M. and Bojar, O. (2013). Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria.

Macháček, M. and Bojar, O. (2014). Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA.

Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000). Query term disambiguation for Web cross-language information retrieval using a search engine. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 25–32, Hong Kong, China.

Magdy, W. and Jones, G. J. F. (2011). An efficient method for using machine translation technologies in cross-language patent search. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1925–1928, Glasgow, United Kingdom.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 133–139, Philadelphia, PA, USA.

Marek, M., Pecina, P., and Spousta, M. (2007). Web page cleaning with conditional random fields. In Fairon, C., Naets, H., Kilgarriff, A., and de Schryver, G.-M., editors, *Proceedings of the 3rd Web As a Corpus Workshop, Incorporating CLEANEVAL*, pages 155–162, Louvain-la-Neuve, Belgium.

Markó, K., Schulz, S., and Hahn, U. (2005). MorphoSaurus–design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of information in medicine*, 44(4):9.

Markó, K. G., Daumke, P., Schulz, S., Klar, R., and Hahn, U. (2007). Large-scale evaluation of a medical cross-language information retrieval system. In Kuhn, K. A., Warren, J. R., and

Leong, T.-Y., editors, *Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems*, volume 129 of *Studies in Health Technology and Informatics*, pages 392–396, Brisbane, Australia.

Meats, E., Brassey, J., Heneghan, C., and Glasziou, P. (2007). Using the Turning Research Into Practice (TRIP) database: how do clinicians really search? *Journal of the Medical Library Association*, 95(2):156–163, Medical Library Association.

Menczer, F. (2005). Mapping the semantics of Web text and links. *IEEE Internet Computing*, 9:27–36, IEEE Educational Activities Department.

Mooers, C. E. (1950). Coding, information retrieval, and the rapid selector. *American Documentation*, 1:225–229.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, page 220–224, Uppsala, Sweden.

Munteanu, D. S. and Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 289–295, Philadelphia, PA, USA.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Murray, C., Dorr, B., Lin, J., Pecina, P., and Hajič, J. (2006a). Leveraging recurrent phrase structure in large-scale ontology translation. In *Proceedings of the 11th Annual conference of the European Association for Machine Translation*, pages 1–10, Oslo, Norway.

Murray, C. G., Dorr, B. J., Lin, J., Hajič, J., and Pecina, P. (2006b). Leveraging reusability: Cost-effective lexical acquisition for large-scale ontology translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 945–952, Sydney, Australia.

Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In *Proceedings Of the International NATO Symposium on Artificial and Human Intelligence*, page 173–180, Lyon, France.

Nakov, P. (2008). Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, page 147–150, Columbus, Ohio, USA.

Nie, J.-Y. (2010). *Cross-language information retrieval.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, page 74–81, New York, New York, USA.

Nikoulina, V., Kovachev, B., Lagos, N., and Monz, C. (2012). Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 109–119, Avignon, France.

Oard, D. W., Levow, G.-A., and Cabezas, C. I. (2001). CLEF experiments at Maryland: Statistical stemming and backoff translation. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation*, volume 2069 of *Lecture Notes in Computer Science*, pages 176–187. Springer Berlin Heidelberg.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, page 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, USA.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, MIT Press.

Och, F. J., Tillmann, C., Ney, H., et al. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.

Och, F. J. and Weber, H. (1998). Improving statistical natural language translation with categories and rules. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 985–989.

Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanbury, A., Jones, G. J., Lupu, M., and Pecina, P. (2015). CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, CEUR Workshop Proceedings*, volume 1391, Toulouse, France.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311–318, Philadelphia, Pennsylvania, USA.

Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlavářová, J., Jones, G. J., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., and Urešová, Z. (2014). Adaptation of machine translation for multilingual information retrieval in the medical domain. *Artificial Intelligence in Medicine*, 61(3):165–185.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., Tamchyna, A., Way, A., and van Genabith, J. (2015). Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193, Springer Netherlands.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., and van Genabith, J. (2012a). Domain adaptation of statistical machine translation using Web-crawled resources: a case study. In Cettolo, M., Federico, M., Specia, L., and Way, A., editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, page 145–152, Trento, Italy.

Pecina, P., Toral, A., and van Genabith, J. (2012b). Simple and effective parameter tuning for domain adaptation of statistical machine translation. In *Proceedings of the 24th International Conference on Computational Linguistics*, page 2209–2224, Mumbai, India.

Pecina, P., Toral, A., Way, A., Papavassiliou, V., Prokopidis, P., and Giagkou, M. (2011). Towards using web-crawled data for domain adaptation in statistical machine translation. In *Proceedings of the 15th Annual Conference of the European Associtation for Machine Translation*, page 297–304, Leuven, Belgium.

Peters, C., Braschler, M., and Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Springer Berlin Heidelberg.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63, Melbourne, Australia.

Pletneva, N., Ruiz de Castaneda, R., Baroz, F., and Boyer, C. (2014). General vs health specialized search engine: a blind comparative evaluation of top search results. *Studies in health technology and informatics*, 205:201—205.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, MCB UP Ltd.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. http://snowball.tartarus.org/ (accessed 1 January, 2014).

Qi, X. and Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM Computing Surveys*, 41:12:1–12:31, ACM.

Qin, J. and Chen, H. (2005). Using genetic algorithm in building domain-specific collections: An experiment in the nanotechnology domain. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, volume 4, page 102b, Big Island, Hawaii, USA.

Quirk, C. and Menezes, A. (2006). Dependency treelet translation: the convergence of statistical and example-based machine-translation? *Machine Translation*, 20(1):43–65, Springer.

Resnik, P. and Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics, Special Issue on the Web as Corpus*, 29:349–380, MIT Press.

Roberts, P. M., Cohen, A. M., and Hersh, W. R. (2009). Tasks, topics and relevance judging for the TREC Genomics Track: five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, 12:81–97.

Robertson, S. E., Walker, S., and Beaulieu, M. (1998). Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track. In Harman, D. K., editor, *The Seventh Text REtrieval Conference (TREC-7)*, pages 253–264, Gaithersburg, MD, USA.

Rogers, F. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116, Medical Library Association.

Rosa, R. (2014). Depfix, a tool for automatic rule-based post-editing of SMT. *The Prague Bulletin of Mathematical Linguistics*, 102:47–56.

Rosemblat, G., Gemoets, D., Browne, A. C., and Tse, T. (2003). Machine translation-supported cross-language information retrieval for a consumer health resource. *AMIA Annual Symposium proceedings*, pages 564–568, American Medical Informatics Association.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard corpus of parallel English and French. Linguistic Data Consortium, Philadelphia, PA, USA.

Ruiz, M., Diekema, A., Sheridan, P., and Plaza, D. C. (1999). CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *The Eighth Text REtrieval Conference (TREC 8)*, pages 597–605, Gaithersburg, MD, USA.

Sadler, V. (1989). *Working with analogical semantics: disambiguation techniques in DLT.* Distributed Language Translation 5. Dordrecht: Foris.

Saleh, S., Bibyna, F., and Pecina, P. (2015). CUNI at the CLEF eHealth 2015 task 2. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, CEUR Workshop Proceedings*, volume 1391, Toulouse, France.

Saleh, S. and Pecina, P. (2014). CUNI at the ShARe/CLEF eHealth evaluation lab 2014. In *CLEF Online Working Notes, CEUR Workshop Proceedings*, volume 1180, pages 226–235, Sheffield, UK.

Saleh, S. and Pecina, P. (2016a). Adapting SMT query translation reranker to new languages in cross-lingual information retrieval. In *Proceedings of the Medical Information Retrieval (MedIR) Workshop. A SIGIR 2016 workshop*, Pisa, Italy.

Saleh, S. and Pecina, P. (2016b). CUNI at the CLEF eHealth evaluation lab 2016. In *CLEF Online Working Notes, CEUR Workshop Proceedings*, Évora, Portugal.

Saleh, S. and Pecina, P. (2016c). *Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval*, pages 54–66. Springer International Publishing, Évora, Portugal.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Sanchis-Trilles, G. and Casacuberta, F. (2010). Log-linear weight optimisation via bayesian adaptation in statistical machine translation. In *The 23rd International Conference on Computational Linguistics, Posters Volume*, page 1077–1085, Beijing, China.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, page 539–549, Avignon, France.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, Blackwell Publishing Ltd.

Sinaiko, H. and Klare, G. (1973). Further experiments in language translation: A second evaluation of the readability of computer translations. In *ITL, Review of Applied Linguistics*, volume 19, page 29–52.

Slocum, J. (1985). A survey of machine translation: its history, current status, and future prospects. *Computational Linguistics*, 11(1):1–17, MIT Press.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006a). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006b). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas*, page 223–231, Cambridge, MA, USA.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2008). Terp system description. In *MetricsMATR workshop at AMTA*.

Spink, A., Wolfram, D., Jansen, M. B. J., and Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, John Wiley & Sons, Inc.

Spousta, M., Marek, M., and Pecina, P. (2008). Victor: the Web-page cleaning tool. In *Proceedings of the 4th Web as Corpus Workshop - Can we beat Google?*, page 12–17, Marrakech, Morocco.

Spoustová, D., Spousta, M., and Pecina, P. (2010). Building a web corpus of Czech. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 998–1001, Valletta, Malta.

Srinivasan, P., Menczer, F., and Pant, G. (2005). A general evaluation framework for topical crawlers. *Information Retrieval*, 8:417–447, Kluwer Academic Publishers.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal.

Stanton, I., Ieong, S., and Mishra, N. (2014). Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142, Queensland, Australia.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 2141–2147, Genoa, Italy.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., and Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth evaluation lab 2013. In Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, CLEF 2013*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231. Springer Berlin Heidelberg, Valencia, Spain.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1).

Tamchyna, A., Dušek, O., Rosa, R., and Pecina, P. (2013). MTMonkey: A scalable infrastructure for a machine translation web service. *Prague Bulletin of Mathematical Linguistics*, 100:31–40, Charles University in Prague.

Tanaka, T. (2002). Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7.

Thouin, B. (1982). The METEO system. In 1982, L., editor, *Practical experience of machine translation*, page 39–44. North-Holland Publishing Company.

Thurmair, G. (1990). Complex lexical transfer in METAL. In *Proceedings of TMI'90*, page 91–107.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, page 2667–2670, Rhodes, Greece.

Toma, P. (1977). Systran as a multilingual machine translation system. In *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*, page 569–581.

Toral, A. (2013). Hybrid selection of language model training data using linguistic information and perplexity. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, page 8–12, Sofia, Bulgaria.

Toral, A., Pecina, P., Wang, L., and van Genabith, J. (2015). Linguistically-augmented perplexity-based data selection for language models. *Computer Speech & Language*, 32(1):11–26.

Tran, T. D., Garcelon, N., Burgun, A., and Beux, P. L. (2004). Experiments in cross-language medical information retrieval using a mixing translation module. *Studies in Health Technology and Informatic*, 107(Pt 2):946–949, IOS Press.

Tsujii, J. (1987). The current stage of the Mu-project. In *Machine Translation Summit*, Hakone Prince Hotel, Japan.

Urešová, Z., Hajič, J., Pecina, P., and Dušek, O. (2014). Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

U.S. National Library of Medicine (2009). UMLS reference manual. Metathesaurus. Bethesda, MD, USA.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing*, page 590–596, Borovets, Bulgaria.

Vogel, S., Ney, H., and Tillmann, C. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Copenhagen, Denmark.

Volk, M., Ripplinger, B., Vintar, Š., Buitelaar, P., Raileanu, D., and Sacaleanu, B. (2002). Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1):97–112, Elsevier.

Voorhees, E. M. and Harman, D. K., editors (2005). *TREC: Experiment and evaluation in information retrieval*, volume 63 of *Digital libraries and electronic publishing series.* MIT press Cambridge, Cambridge, MA, USA.

Voorhees, E. M. and Tong, R. M. (2011). Overview of the TREC 2011 Medical Records Track. In *The Eleventh Text Retrieval Conference (TREC 2002)*, pages 1–11, Gaithersburg, MD, USA.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Columbus, Ohio.

Weaver, W. (1955). Translation. In *Machine Translation of Languages*, page 15–23. MIT Press, Cambridge, Massachusetts.

White, J. S., O'Connell, T., and O'Mara, F. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and further approaches. In *Proceedings of the 1994 Conference of the Association for Machine Translation in the Americas*, pages 193–205, Columbia, MD.

Wu, C., Xia, F., Deleger, L., and Solti, I. (2011). Statistical machine translation for biomedical text: are we there yet? *AMIA Annual Symposium proceedings*, page 1290–1299.

Wu, H. and Wang, H. (2004). Improving domain-specific word alignment with a general bilingual corpus. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, page 262–271, Washington, DC, USA.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd Internation-*

al Conference on Computational Linguistics, volume 1, page 993–1000, Manchester, United Kingdom.

Yu, H., Han, J., and Chang, K. C.-C. (2004). PEBL: Web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):70–81, IEEE Educational Activities Department.

Zhang, Y., Wu, K., Gao, J., and Vines, P. (2006). Automatic acquisition of Chinese-English parallel corpus from the Web. In *Proceedings of the 28th European Conference on Information Retrieval*, page 420–431, London, UK.

Zhou, D., Truran, M., Brailsford, T., Wade, V., and Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, 45(1):1:1–1:44, Association for Computing Machinery.

Part II

Selected Publications

1 Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation

Pecina et al. (2011): Pavel Pecina, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou. **Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation**. In Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 297–304, Leuven, Belgium. European Association for Machine Translation, 2011.

2 Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A Case Study

Pecina et al. (2012a): Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. **Domain Adaptation of Statistical Machine Translation using Web-Crawled Resources: A Case Study**. In Mauro Cettolo and Marcello Federico and Lucia Specia and Andy Way, editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152, Trento, Italy. European Association for Machine Translation, 2012.

3 Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation

Pecina et al. (2012b): Pavel Pecina, Antonio Toral, and Josef van Genabith. **Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation**. In Martin Kay and Christian Boitet, editors, *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2209–2224, Mumbai, India. Coling 2012 Organizing Committee, 2012.

4 Domain Adaptation of Statistical Machine Translation with Domain-focused Web Crawling

Pecina et al. (2015): Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way, and Josef van Genabith. **Domain Adaptation of Statistical Machine Translation with Domain-focused Web Crawling**. In *Language Resources and Evaluation*, 49 (1), pp. 147–193., Springer Netherlands, 2015. `https://doi.org/10.1007/s10579-014-9282-3`.

5 Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain

Urešová et al. (2014): Zdeňka Urešová, Ondřej Dušek, Jan Hajič, and Pavel Pecina. **Multilingual Test Sets for Machine Translation of Search Queries for Cross-Lingual Information Retrieval in the Medical Domain**. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (editors:) *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 3244–3247, Reykjavik, Iceland. European Language Resources Association, 2014.

6 Machine Translation of Medical Texts in the Khresmoi Project

Dušek et al. (2014): Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, and Daniel Zeman. **Machine Translation of Medical Texts in the Khresmoi Project**. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 221–228, Baltimore, USA. Association for Computational Linguistics, 2014. `https://doi.org/10.3115/v1/w14-3326`.

7 Adaptation of Machine Translation for Multilingual Information Retrieval in the Medical Domain

Pecina et al. (2014): Pavel Pecina, Ondřej Dušek, Lorraine Goeuriot, Jan Hajič, Jaroslava Hlaváčová, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling, David Mareček, Michal Novák, Martin Popel, Rudolf Rosa, Aleš Tamchyna, and Zdeňka Urešová. **Adaptation of Machine Translation for Multilingual Information Retrieval in the Medical Domain**. In Hanna Suominen (editor): *Artificial Intelligence in Medicine, 61 (3), Text Mining and Information Analysis of Health Documents.*, pages 165–185, Elsevier, 2014. `https://doi.org/10.1016/j.artmed.2014.01.004`.

8 Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval

Saleh and Pecina (2016c): Shadi Saleh and Pavel Pecina. **ing Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval**. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. The 7th International Conference of the CLEF Association, Évora, Portugal. Volume 9822 of the series Lecture Notes in Computer Science*, pages 54–66, Springer International Publishing, 2016. `https://doi.org/10.1007/978-3-319-44564-9_5`.