Experiments with Domain Adaptation Methods for Statistical MT: From European Parliament Proceedings to Finnish Newspaper Text

Marcus Dobrinkat and Jaakko J. Väyrynen
Aalto University School of Science and Technology
Department of Information and Computer Science
{marcus.dobrinkat,jaakko.j.vayrynen}@tkk.fi

Abstract

Statistical machine translation methodology is highly dependent of relevant parallel texts for training. However, available large parallel corpora are typically out-of-domain for many interesting translation tasks, such as news translation. We experiment with a very small Finnish–English news corpus using four different domain adaptation methods: language model adaptation, translation model adaptation, automatic post-editing and re-training with combined data. Translation quality is measured with the de-facto standard MT evaluation metric BLEU and we provide statistical significance testing for system comparison. Language model adaptation did not produce significant improvements. All other tested methods outperformed the baseline system.

1 Introduction

Machine translation (MT) adaptation aims to improve translation performance on text in a specific domain that is not present or pronounced in the bilingual training corpus. Domain adaptation is especially important in statistical machine translation (SMT) systems which are trained with empirical data and are closely tied to the training data domain. Text corpora can be very different in many aspects, such as vocabulary, style or grammar. Therefore, the performance of SMT systems is more susceptible to domain differences than traditional rule based systems which do not depend on example translations.

Our task is to improve the quality of a baseline SMT system with different domain adaptation methods. We use a very small in-domain corpus which makes training and evaluation challenging. We provide statistical significance testing to facilitate system comparison.

Our in-domain data consists of a small parallel Finnish–English news domain corpus that we have collected. The translations from Finnish to English were created by volunteers in an online system. The baseline system is trained on the out-of-domain Europarl parallel corpus.

The very small size of the in-domain corpus is highly problematic for parameter optimization and evaluation purposes as separate test and development sets would significantly reduce the size of the training data which would make the model learning very problematic. Our compromise to this was to use cross validation and bootstrap resampling for evaluation and to only try to get reasonable parameter values for the most interesting parameters related to domain adaptation. We believe this is a reasonable approach as we are interested in improving translation quality with different methods rather than optimizing the systems for the best possible performance.

We experiment with four different domain adaptation methods: language model (LM) adaptation, translation model (TM) adaptation, automatic post-editing (APE) and re-training of the model with combined in-domain and out-of-domain data. Language model adaptation does not modify the translation model, but influences the choices made by the translation model, whereas translation model adaptation methods include new translations and can be further combined with LM adaptation methods. Automatic post-editing does not modify the baseline translation model, but learns an additional translation model from the output of the baseline system to the reference translations.

In our goal to improve the baseline system translation quality, we evaluate the performance of four different domain adapation methods, namely language model adaptation, translation model adaptation, automatic post-editing and combination of in- and out-of-domain data for model retraining.

Each tested adaptation method significantly outperformed the baseline system in BLUE scores, except for the LM adaptation in which the improvement was not significant. There was no clear preference of the adaptation methods.

1.1 Statistical Machine Translation

In statistical machine translation, a statistical model governs the mapping from source (s) to target (t) sentence. Although the original ideas of SMT were already introduced in the work of Weaver (1949), the influential work of Brown et al. (1994) renewed the research in the SMT paradigm. The translation problem is represented in a probabilistic framework in which the Bayes formula gives the source-channel approach

$$P(t|s) \propto P(s|t)P(t) \quad (1)$$

for machine translation. It splits the conditional probability of t given s into two parts: a translation model ($P(s|t)$) and a language model ($P(t)$). The best translation is found by maximizing Equation 1.

A more general approach using a log-linear model, employing a maximum entropy framework has been formulated in Och and Ney (2001). It provides M feature functions $h_m(t, s)$ with weights λ_m. The translation probability $P(t|s)$ is then defined as:

$$P(t|s) \propto \exp\left[\sum_{m=1}^{M} \lambda_m \, h_m(t, s)\right] \quad (2)$$

As before, the best translation is found when $P(t|s)$ is maximized. The source-channel model in Equation 1 can be modeled as a special case of the log-linear framework in Equation 2 by choosing equal weights, $\lambda_1 = \lambda_2$, and feature functions $h_1(t, s) = \log P(s|t)$ and $h_2(t, s) = \log P(t)$.

More details about how the model is discriminatively trained, can be found in Och and Ney (2001). One advantage of this more general model is that additional features can easily be included Koehn et al. (2003); Och and Ney (2001).

2 Related Work

Domain adaptation has been performed with a wide range of methods, which can be categorized by their use of in-domain resources (no additional in-domain data, monolingual in-domain data, dedicated in-domain dictionaries or dedicated in-domain parallel corpora) or the way these resources are used (interpolation of out-of-domain with in-domain data or models for language model or translation model adaptation or out-of-domain and in-domain system combinations).

Improving in-domain performance without a dedicated in-domain bilingual corpus is done by Ueffing et al. (2007), who call their approach transductive learning. Using the non-adapted system, they first translate a source language monolingual text corpus, select the good translations and paired them with their source sentences to build a new synthetic in-domain corpus. Re-training the system with this corpus strengthens valuable phrase table content and weakens less useful content, which makes the system gain knowledge from its own output.

Hildebrand et al. (2005) compile an in-domain corpus out of a large general domain bilingual corpus. Their basic assumption is that this large corpus contains different domain sub-corpora, which are obtained by selecting those sentence pairs only, which match the in-domain test set.

Xu et al. (2007) assume an existing bilingual in-domain corpus describing an approach towards a multi-domain machine translation system. The different domain LMs are combined as sentence level mixtures (Iyer and Ostendorf, 1996) using interpolation. Different domain translation models are trained and optimized separately and combined during decoding as different features in a log-linear model. Feature weights are chosen on-line, depending on the domain of the input text.

There are various approaches to LM adaptation, as enumerated by Béchet et al. (2004). He lists linear interpolation of out-of-domain and in-domain language models, and an information retrieval approach where documents matching the required domain are retrieved and trained on-line to create the in-domain language model.

The work of Zhao et al. (2004) combines language model adaptation and information retrieval in the context of machine translation. Using the nonadapted system, they generate a list of translation hypotheses, which are used to create a retrieval query run against large-scale monolingual text corpora. The best result sentences are then used to train a new in-domain language model which is linearly interpolated with the out-of-domain language model. Then translation hypotheses are re-created using the interpolated language model, before they build and run the queries and generate the in-domain language model. They achieve their best results using query models that incorporate additional structure in the queries.

Wu et al. (2008) use linear interpolation of language models as well as of translation models. However, instead of a given bilingual in-domain corpus, they employ an in-domain word dictionary for adaptation. They treat the dictionary as a small in-domain phrase table or as data for in-domain translation model training. In-domain and out-of-domain phrase tables are combined during decoding. Either each phrase table is used as factor in the log-linear

translation model, or both are linearly interpolated similar to the language models. Their intermediate results suggested that the log-linear approach works better.

Koehn and Schroeder (2007) have a similar arrangement. Their simplest phrase table adaptation setup is to combine in-domain and out-of-domain bilingual corpora before training. A more advanced way is to create two separate phrase tables, which are combined using factored translation models (Koehn and Hoang, 2007). They create an adapted language model in different ways, either using only the in-domain LM, linearly interpolating it with the out-of-domain LM, or using both as separate features in the log-linear translation model.

A quite different approach to domain adaptation is automatic post-editing (APE). In manual translation, a translator who corrects output from an MT system does post-editing. In automatic post-editing, manual corrections are used to train a system that automatically corrects the output of the original MT system. In such way, the post-edit system should learn to relieve the editors of repeatedly fixing the same mistakes made by the MT system. We experiment with this by training one SMT system to correct translations made by another SMT system.

Isabelle et al. (2007) improve PORTAGE, a RBMT system, by the use of SMT as post-processing step. A bilingual corpus is constructed using the RBMT output translations as source text and the post-editor reference translations as target text. This corpus is used for SMT model training. In this setup, two translation steps are performed: the source text is translated by the RBMT to intermediate target language, which is translated by the APE layer to correct target language text. This process can be used to easily customize the RBMT system, or to adapt it to a specific domain. In their experiments, Isabelle et al. (2007) report results for a small APE-training corpus (< 500k words) of human corrections. The system yields almost the same results in BLEU score, as an RBMT system customized with 18000 manual entries. With an increase of APE training data, the overall quality improvement stagnates. The improvements seem to be limited by the output quality of the RBMT system.

Simard et al. (2007) report similar experiments using the PORTAGE system. Dugast et al. (2007) worked on improving the SYSTRAN RBMT system by statistical post-editing. They work with the English–French language pair and confirm good results by automatic evaluation as well as linguistic analysis. The SPE layer mostly improves local word choice, degrades morphological accuracy and does not affect long-distance reordering (which the RBMT does well).

In similar work, Díaz de Ilarraza et al. (2008) concentrated on the Spanish–Basque language pair, where little bilingual material is available. They use the open source RBMT system Matxin (Alegria et al., 2007). As Basque is a morphologically rich language, each word in the source corpus was replaced by its stem and additional morphological tags. Tests with this morpheme-based SMT system show significant improvements in NIST, WER and PER scores over the word-based SMT system (except for BLEU scores, which are worse). Their results are consistent with other research for a restricted domain corpus. However, for a general domain corpus the plain SMT system outperforms the combination of RMBT system with SPE module.

3 Data and Methods

3.1 Data

Our baseline translation system is trained on a reasonably large amount of out-of-domain parallel data to get a state-of-the-art SMT system. We have only little parallel in-domain data and have no specific monolingual in-domain target language corpus for language model training. The amount of parallel data could be alleviated by considering some other language pair than Finnish–English, but we are interested in considering Finnish, as it is the local language and the reported challenges in translating to and from Finnish with statistical methods (Koehn, 2005). Our experiments investigate the case of starting to adapt translation system with very little in-domain data.

The Europarl (denoted as 'ep') corpus Koehn (2005) was our out-of-domain corpus for training the baseline translation (TM), reordering (RM) and language (LM) models. We created a small in-domain news corpus (denoted as 'il') for domain adaptation and evaluation. For the adaptation methods we concatenated the two corpora (denoted as 'ep+il') or trained separate models that were combined in a log-linear framework (denoted as 'ep,il') or linearly interpolated (denoted as 'ep*il'). For post-edit models we paired the English output translations by the baseline system with the English reference translation for those sentences (corpus denoted as 'pec'). The corpora were preprocessed with the standard Moses scripts, included lowercasing and tokenization.

3.1.1 Europarl Parallel Corpus

The baseline models were generated from the Europarl corpus version 2 Koehn (2005), which is a widely used parallel corpus in SMT research. The corpus is freely available and based on the web versions of the European Parliament proceedings from April 1996 to September 2003 in eleven languages.

For the experiments in this paper, we selected the English–Finnish data. It contains the proceedings data from January 1997 to September 2003 with a total of 0.8 million sentence pairs after standard preprocessing.

3.1.2 Iltalehti Parallel Corpus

The monolingual Finnish in-domain corpus was extracted from the web version of Iltalehti, a Finnish daily tabloid newspaper. Sentence length was limited to minimum of 3 and maximum of 12 words. Sequences shorter than 3 words were not considered proper sentences and sequences longer than 11 words were considered too complex and laborious to manually evaluate and correct.

The extracted Finnish sentences were translated into English by volunteers using a web-based application. The created small in-domain parallel corpus consisted of 1076 sentences.

3.2 Baseline Translation System

All experiments were conducted with Moses Koehn et al. (2007), an open source, state of the art statistical machine translation system.[1] During decoding we used the default settings except for a translation table limit of 20, word penalty of -1 and a distortion limit of 6. We used the default reordering model (`msd-bidirectional-fe`).

The Finnish–English baseline models were trained with Europarl corpus (Koehn, 2005). The 4-gram language models used in all experiments were created using the SRILM toolkit Stolcke (2002) with Kneser-Ney smoothing.

3.3 Adaptation Methods

Here we provide brief descriptions of the adaptation methods in our experiments. A short identifier for each experiment is given in parenthesis.

Language model adaptation (Ln) only modifies the language model component of the system, whereas translation model adaptation (In/Cn) modifies models responsible for translation. The post-edit adaptation (Pn) does not change the baseline model (B) but builds a new translation system that takes as input the output of the baseline system. Table 1 describes the different data sets, models and methods.

3.3.1 Language Model Adaptation

We experiment with two different approaches: one new model based on all data (L2) and linear (L3) and log-linear (L1) model interpolation between baseline and in-domain models. We provide comparison with results with only in-domain (L4) and baseline language model (B).

We create linearly interpolated LMs with the SRILM toolkit and use Moses for log-linear combination (Equation 2) of LMs by using them as distinct features.

3.3.2 Translation Model Adaptation

Similarly to LM adaptation, we compare the baseline system to a new translation model that is trained on combined baseline and in-domain corpora (Cn) and a log-linear combination between separately trained baseline and in-domain models (In).

We show results with three different language models: a baseline language model (C1/I1), a LM trained on both corpora (C2/I2) and linear interpolation between baseline and in-domain LMs (C3/I3).

3.3.3 Post-edit Domain Adaptation

The automatic post-edit domain adaption method does not work in parallel with the baseline model as in model interpolation, but the baseline and the APE models are in a sequence. Therefore, the approach is feasible even if the baseline system details cannot be accessed or modified. This method does not translate between two different languages, but rather tries to correct the baseline system output to match the reference translations.

As in the translation model experiments, we show results with three different language models: a baseline language model (P1), a LM trained on both corpora (P2) and linear interpolation between baseline and in-domain LMs (P3).

3.4 Evaluation

We evaluate translation quality using the BLEU score measure (Papineni et al., 2002), which is commonly used in MT research, although it has received much

[1] Moses build from 11.12.2007 used in all experiments.

criticism (Callison-Burch et al., 2006; Lee and Przybocki, 2005). The basic idea of BLEU is to reward closeness to one of the human reference translations, using modified unigram precision. The precision is determined by the weighted overlap of n-grams between candidate and reference translations for $n = 1, .., 4$. The closeness between candidate and reference is given by the final score between 0 and 1.

Given our small bilingual in-domain corpus, it is hard to obtain a representative sample and the statistical significance of our results could be questioned. Therefore, a combination of 10-fold cross validation and bootstrap resampling Efron and Tibshirani (1986) was used. Cross-validation enlarges the variability of the training data and bootstrap resampling improves statistical accuracy for test set evaluation, while not assuming any specific distribution for the data. Bootstrap resampling has earlier been applied for significance tests in machine translation (Koehn, 2004; Zhang et al., 2004).

Due to to the very small in-domain adaptation corpus, we did not create a separate tuning data set as this would have reduced the available data for training too much. However, default parameters could be unfortunate in a way that the comparison between different systems is unfair. Therefore we decided to tradeoff proper division of validation and test set for better comparability between the different adaptation approaches, which was our main goal.

The training sets were used to train translation models, reordering models and language models. The testing sets were used during automatic evaluation and for trying to obtain reasonable interpolation weights between in-domain and out-of-domain models.

Target language translations of each cross-validation model were resampled with replacement to form 1 000 new sets of 100 sentence test corpora. The test set for each cross-validation fold was bootstrapped and the BLEU score for each of these $10 \cdot 1\,000$ test sets evaluated. These BLEU scores were then combined to determine the bootstrap confidence interval and mean estimates.

For system ranking we perform a pairwise comparison as described in Zhang et al. (2004), but use a one sided 95% interval. The method can be described as first calculating the difference between each paired sample and subsequently verifying if 95% of the differences are larger than zero for any one of the participating systems. If the condition is met, the score difference is significant at the 95% level. To our knowledge, this procedure has not been widely applied in MT research yet, therefore we also rank the systems using the Wilcoxon signed rank test Wilcoxon (1945).

Id	Description
ep	Europarl (Finnish,English) corpus
il	Iltalehti (Finnish,English) corpus
pec	Post-edit corrections (English,English)
ep+il	One model trained on combined corpora
ep,il	Log-linear combination of models
ep*il	Linear interpolation of models
B	Baseline translation system
Ln	Language model adaptation only
Cn	Adaptation with data combination
In	Adaptation with model interpolation
Pn	Adaptation with post-editing

Table 1: Data, model and system descriptions.

4 Results

Evaluation for all systems are given in Table 2 which shows the data and models used for training and BLEU scores for training and test sets.

Our results with language model adaptation (Ln) are not in line with existing research, as we were only able to get improved BLEU scores with a new model trained on all data (L2). Koehn and Schroeder (2007) try similar experiments as our Ln, although with a significantly larger in-domain corpus. Compared to the baseline, they report improvements in each of the LM adaptation methods, where the simple combination of corpora (comparable with L2) performs worse than the other methods. For our LM adaptation methods, L2 performs best. For linear LM interpolation (L3), we tried different weights using LM perplexity as a guide. However, the weight giving the lowest perplexity (0.5) did not result in the best translation score; similar scores were instead achieved for weights between 0.5 to 0.9.

Linear LM interpolation (L3) outperformed log-linear LM combination (L1), which agrees with the results in Wu et al. (2008). We tried different weights for the two LMs in the log-linear LM combination (L1), but additional in-domain LM weight only seemed to degrade translation performance.

All translation adaptation methods (Cn/In/Pn) outperformed the baseline system. Language model, however, did influence the results as interpolated LMs (ep*il) produced lower scores than the baseline model (ep), whereas the retrained model (ep+il) was always the best. The only exception was the post-edit adaptation (Pn), where also linear interpolation of LMs outperformed the baseline LM.

Id	Description	Data			Training	Testing		
		TM	RM	LM		cross-validation		bootstrap
						mean	interval	interval
B	baseline, no adaptaion	ep	ep	ep	16.49	16.43	[15.08, 17.79]	[12.64, 20.38]
L1	log-linear LM combination	ep	ep	ep, il	20.92	13.28	[11.89, 14.68]	[9.91, 16.88]
L2	combined corpus LM	ep	ep	ep+il	20.50	17.25	[15.78, 18.72]	[13.33, 21.40]
L3	linear LM interpolation	ep	ep	ep*il	19.79	14.86	[13.79, 15.93]	[11.42, 18.45]
L4	in-domain LM only	ep	ep	il	20.29	10.77	[9.45, 12.08]	[7.808, 13.87]
C1	combined corpus TM/RM	ep+il	ep+il	ep	48.92	21.41	[19.58, 23.23]	[16.79, 26.32]
C2	+combined corpus LM	ep+il	ep+il	ep+il	55.70	22.41	[20.55, 24.28]	[17.50, 27.57]
C3	+linear LM interpolation	ep+il	ep+il	ep*il	56.19	21.23	[19.73, 22.73]	[16.57, 26.20]
I1	log-linear TM combination	ep, il	ep	ep	62.92	23.75	[21.87, 25.64]	[18.77, 29.04]
I2	+combined corpus LM	ep, il	ep	ep+il	68.98	24.76	[22.49, 27.03]	[19.52, 30.39]
I3	+linear LM interpolation	ep, il	ep	ep*il	69.89	23.43	[21.41, 25.44]	[18.28, 29.08]
P1	post-edit TM/RM	pec	pec	ep	57.75	22.74	[21.24, 24.24]	[17.52, 28.48]
P2	+combined corpus LM	pec	pec	ep+il	61.02	24.05	[22.35, 25.75]	[18.47, 30.01]
P3	+linear LM interpolation	pec	pec	ep*il	61.23	23.49	[21.81, 25.16]	[17.99, 29.35]

Table 2: BLEU score evaluation of the in-domain news corpus test set for all adaptation systems using 10-fold cross-validation and bootstrap resampling. 95% confidence intervals are reported for bootstrap resampling and cross-validation data, for the latter assuming Student's t-distributed data. Mean values for the bootstrap data are not shown as they were equal to the reported cross-validation means when using four significant figures.

A comparison of the best systems (B/L2/C2/I2/P2) in each method is shown in Figure 1 as a histogram of the BLEU scores from the bootstrap resampling sets. This result was used to rank the adaptation methods, which gave the result given in Table 3. [2]

comparison	result	p-value
B	$\not>$ L2	0.93
B	$<$ C2	0.0011
B	$<$ I2	< 0.001
B	$<$ P2	0.015
L2	$<$ C2	0.0029
L2	$<$ I2	< 0.001
L2	$<$ P2	0.028
C2	$\not>$ I2	0.92
C2	$\not>$ P2	0.66
I2	$\not>$ P2	0.43

(a) Bootstrap method

comparison	result	p-value
B	$<$ L2	0.0020
B	$<$ C2	< 0.001
B	$<$ I2	< 0.001
B	$<$ P2	< 0.001
L2	$<$ C2	< 0.001
L2	$<$ I2	< 0.001
L2	$<$ P2	< 0.001
C2	$<$ I2	0.0029
C2	$\not>$ P2	0.90
I2	$\not>$ P2	0.35

(b) Wilcoxon signed-rank test

Table 3: A system ranking of the best systems of each method. Using the more pessimistic bootstrap method, we get the ranking $(C2, I2, P2) > (L2, B)$. Applying the more sensitive Wilcoxon signed-rank test on the cross-validation data results in the ranking $I2 > C2 > L2 > B$ and $P2 > L2 > B$.

The ranking obtained by the bootstrap method is more pessimistic than the Wilcoxon signed-rank test over the cross-validation data. All adaptation methods that include translation model adaptation (C2, I2, P2) perform significantly better than the baseline and the LM only adaptation (L2) using any of the testing methods. Wilcoxon signed-rank test shows the significant difference that the interpolation model (I2) is better than the model from the combined corpora (C2) and both are better than the LM adaptation (L2) alone. The post-editing model (P2) is also better than the baseline and L2, but not significantly better than combined corpora model C2 or interpolation model I2.

The detailed model evaluation results in Table 2 show that there is a considerable difference between the estimated cross-validation confidence interval [3] and the confidence intervals created from the bootstrapped data, which have a much larger interval.

5 Conclusions and Discussion

This paper experimented with a statistical machine translation framework and four different domain adaptation methods from the baseline system trained on the Finnish–English part of the Europarl corpus to a news domain. The in-domain news corpus was a very small parallel corpus collected by the authors. The results show that the adaptation methods can sig-

[2] All rankings use a significance level of 95%.

[3] based on the assumption that the data is t-distributed

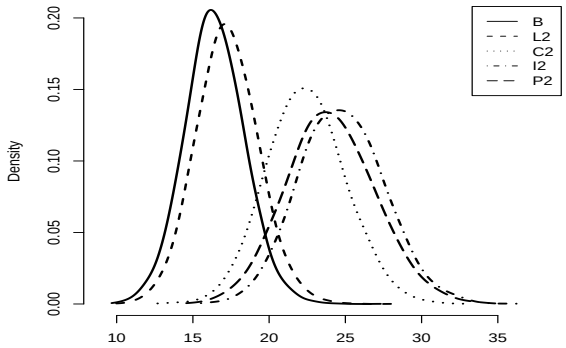

Figure 1: Best models of each family compared by smoothed BLEU score histograms created from the bootstrap resampling test sets.

nificantly improve translation quality measured with the BLEU score, even when the in-domain training corpus is very small. Our results suggest that language model adaptation combined with translation model adaptation methods or post-editing methods produces the best results. Language model adaptation methods by itself may not always improve the results.

We were not able to show a clear ranking of the adaptation methods. However, the choice of the appropriate method might depend not only on the improvements in translation quality but also on other performance measures, such as training time, model size and translation time.

Due to the small size of the in-domain parallel corpus, parameters of each system were not fully optimized. Some of our results deviate from existing research, especially regarding the language model interpolation, which we suspect to be a result of the very small in-domain corpus size. In further work, we believe that combining the methodologically different adaptation methods would produce even greater improvements in translation quality.

References

Iñaki Alegria, Arantza Díaz de Ilarraza, Gorga Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. In *LNCS 4394*, 2007.

Frédéric Béchet, Renato de Mori, and David Janiszek. Data augmentation and language model adaptation using singular value decomposition. *Pattern Recognition Letters*, 25(1):15–19, 2004.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1994.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL-2006*, pages 249–256, 2006.

Arantza Díaz de Ilarraza, Gorga Labaka, and Kepa Sarasola. Statistical post-editing: a valuable method in domain adaptation of RBMT systems for less-resourced languages. In *Proceedings of MATMT2008: Mixing Approaches to Machine Translation*, pages 35–40, 2008.

Loïc Dugast, Jean Senellart, and Philipp Koehn. Statistical post-editing on SYSTRAN's rule-based translation system. In *Proceedings of WMT07*, pages 220–223, 2007.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Almut S. Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the EAMT 2005*, 2005.

Pierre Isabelle, Cyril Goutte, and Michel Simard. Domain adaptation of MT systems through automatic post-editing. In *MT Summit XI*, pages 255–261, 2007.

Rukmini Iyer and Mari Ostendorf. Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models. In *IEEE Transactions on Speech and Audio Processing*, pages 236–239, 1996.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, 2004.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, 2005.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of EMNLP-CoNLL 2007*, pages 868–876, 2007.

Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of WMT07*, pages 224–227, 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of NAACL'03*, pages 48–54, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL'2007*, 2007.

Audrey Lee and Mark Przybocki. NIST 2005 machine translation evaluation official results. official release of automatic evaluation scores for all submissions, 2005.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL'02*, pages 295–302, 2001.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002.

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. Rule-based translation with statistical phrase-based post-editing. In *Proceedings of WMT07*, pages 203–206, 2007.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP-2002*, pages 901–904, 2002.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2):77–94, 2007.

Warren Weaver. *Translation (1949)*. The Technology Press of the Massachussetts Institute of Technology/John Wiley, New York/Clapham & Hall (London), 1949.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, (1):80–83, 1945.

Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of COLING'08*, pages 993–1000, 2008.

Jia Xu, Yonggang Deng, Yuqing Gao, and Hermann Ney. Domain dependent statistical machine translation. In *Proceedings of MT Summit XI*, pages 515–520, 2007.

Ying Zhang, Stephan Vogel, and Alex Waibel. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of LREC*, pages 2051–2054, 2004.

Bing Zhao, Matthias Eck, and Stephan Vogel. Language model adaptation for statistical machine translation with structured query models. In *COLING'04*, pages 411–417, 2004.