Mixed-Domain vs. Multi-Domain Statistical Machine Translation

Matthias Huck mhuck@inf.ed.ac.uk
Alexandra Birch a.birch@ed.ac.uk
Barry Haddow bhaddow@inf.ed.ac.uk
School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Domain adaptation boosts translation quality on in-domain data, but translation quality for domain adapted systems on out-of-domain data tends to suffer. Users of web-based translation services expect high quality translation across a wide range of diverse domains, and what makes the task even more difficult is that no domain label is provided with the translation request.

In this paper we present an approach to domain adaptation which results in large-scale, general purpose machine translation systems. First, we tune our translation models to multiple individual domains. Then, by means of source-side domain classification, we are able to predict the domain of individual input sentences and thereby select the appropriate domain-specific model parameters. We call this approach multi-domain translation.

We develop state-of-the-art, domain-adapted translation engines for three broadly-defined domains: *TED talks*, *Europarl*, and *News*. Our results suggest that multi-domain translation performs better than a mixed-domain approach, which deploys a system that has been tuned on a development set composed of samples from many domains.

1 Introduction

Domain adaptation is a common approach to significantly improve machine translation quality on input documents from a given domain. Domain adaptation techniques for statistical machine translation (SMT) have been extensively studied and are well established (Federico and Bertoldi, 2012). In practice, machine translation systems are often engineered to perform well on the domain of one specific application. Most research on domain adaptation assumes that any prospective input data originates from a single domain, and the characteristics of this domain are known beforehand, e.g. by means of existing samples from the same domain which can be employed for training and tuning. The adaptation task is then defined as utilizing a small amount of in-domain training resources effectively in order to learn system parameters that are more appropriate for translating in-domain input. The in-domain training resources constitute a minor fraction of the overall training data only, the majority of which has a domain mismatch with the designated application.

The downside of systems that have been highly tweaked towards the characteristics of a single domain is a diminished translation quality on out-of-domain data (Haddow and Koehn, 2012). Online translation systems, on the other hand, are usually designed for open-domain scenarios where the domain of the input text is not predefined. Being able to take advantage of the benefits of domain adaptation while not having to compromise quality on out-of-domain data would be desirable for online systems.

A viable utilization of domain adaptation approaches in open-domain online translation systems comes in two components:

- A number of different parameter sets, each tuned to optimize translation quality on texts from a specific domain.
- A text classifier that predicts the domain of foreign-language input data prior to decoding.

As the input data is not labeled with a domain (and, in an open-domain setting, may even originate from a new domain), the text classifier first has to assign the most likely class from a set of multiple known domains. The decoder is then reconfigured with a domain-specific parameter set (e.g., a weight vector), which should be the most appropriate one for achieving high translation quality on the current input. We refer to this approach as *multi-domain SMT*.

In this work we investigate different methods for domain classification:

- Classification based on the scores of language models (LMs) which have been interpolated with interpolation weights tuned on in-domain development sets.
- Maximum entropy text classifiers trained on medium-sized training corpora.
- Maximum entropy text classifiers trained on the same smaller domain-specific development sets which are employed for tuning the machine translation systems.

An obvious alternative method for building an open-domain online translation system is tuning on a corpus containing samples of texts from all known domains, which collectively are considered to be representative for the application. We refer to this approach as *mixed-domain SMT*. A difficulty here is the choice of the corpus samples in a way that brings about good performance across all domains. However, a high-quality generic system with a single parameter set that does not depend on a domain label is appealing.

In the empirical part of this paper, we compare multi-domain and mixed-domain SMT on the English→German, English→Italian, English→Portuguese, and English→Greek language pairs using training corpora of diverse origin, totalling tens of millions of parallel sentences.

2 Related Work

A significant amount of research on domain adaptation for SMT has been conducted in recent years. Some methods which are commonly used are:

- Tuning of the decoder model weights (Och and Ney, 2002) on an in-domain development set (Pecina et al., 2012).
- Model combination (of language models, translation models, or reordering models) via interpolation or other schemes, e.g. phrase table fill-up (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Nakov, 2008; Bisazza et al., 2011; Niehues and Waibel, 2012; Chen et al., 2013).
- Data selection (Moore and Lewis, 2010; Axelrod et al., 2011).
- Instance weighting (Matsoukas et al., 2009; Foster et al., 2010; Shah et al., 2012; Mansour and Ney, 2012).
- Further exploitation of in-domain monolingual data (Ueffing et al., 2007; Bertoldi and Federico, 2009; Schwenk and Senellart, 2009; Lambert et al., 2011).
- Domain-specific features, e.g. binary features indicating the provenance of phrase pairs as implemented in the open-source Moses toolkit (Durrani et al., 2013b) or "domain augmentation" (Clark et al., 2012).

However, few authors have tackled the question of how to benefit from domain adaptation in scenarios where a domain label of the input is not present. An important aspect of our approach to multi-domain MT is the need for domain classification.

Xu et al. (2007) perform domain classification for a Chinese→English task. The domains are *newswire* and *newsgroup*. The classifiers operate on whole documents rather than on individual sentences. The authors propose two techniques for domain classification. Their first technique is based on interpolated LMs: a general-domain LM is interpolated with LMs which were trained on in-domain development sets, resulting in a number of domain-specific interpolated LMs. The interpolation weight is heuristically chosen. The classifier computes LM perplexities over input documents and assigns the domain with the lowest perplexity. Their second technique is based on a metric which measures similarity wrt. vocabulary.

Banerjee et al. (2010) conduct machine translation experiments with classification of two technical documentation domains, *availability* and *security* in the area of computing. Empirical results on Chinese→English and English→Chinese tasks are presented. The authors build a Support Vector Machine (SVM) classifier using Term Frequency Inverse Sentence Frequency features over bigrams of stemmed content words. Classification is carried out on the level of individual sentences. The SVM is trained on the SMT training corpora (∼226k sentences in total). Several setups with different domain-adapted and domain-agnostic systems are evaluated. The authors show that a pipeline with the SVM classifier is effective in multi-domain translation.

Wang et al. (2012) distinguish generic and patent domain data in experiments on 20 language pairs. For domain classification, the authors rely on averaged perceptron classifiers with various phrase-based features. The machine translation development sets serve as training data for the classifiers. An interesting aspect of their translation experiments is that they utilize a multi-domain optimization in order to jointly tune weights for all domains in a single run of lattice MERT (Macherey et al., 2008).

In a related strand of research, source-side text classifiers have recently been employed in order to detect Arabic dialects and select SMT systems accordingly (Salloum et al., 2014; Mansour et al., 2014).

3 Text Domains

Our application scenario is an online translation service with the requirement to provide high-quality translation not only of texts from a single domain, but of a wider range of text types. We therefore study a use case where the translation system is supposed to perform well on the following domains: *TED talks*, *Europarl*, and *News*. These three domains are fairly coarse-grained. Different documents from one of the domains are mostly not consistent regarding the covered topics. While all three domains comprise heterogeneous topics, the domains are set apart from each other by means of text style.

TED talks are transcripts of spoken language from short public presentations. The presentations often cover scientific subjects which are expressed in layman's terms and in an informal manner. TED talks are not spontaneous speech. They are however designed to be entertaining. *Europarl* texts are transcripts of speeches on political matters from parliamentary proceedings. *News* texts are written news articles.

TED talks, *Europarl*, and *News* could be described as "genres". We denote them as domains throughout this paper because the term "domain" is well established in related machine translation research literature and often used in a broad sense.

TED talks, *Europarl*, and *News* have been highly relevant domains in recent machine translation research. The International Workshop on Spoken Language Translation[1] (IWSLT) hosts a yearly open evaluation campaign which focuses on the translation of TED talks since 2011 (Federico et al., 2011). The European Parliament Proceedings Parallel Corpus (Koehn, 2005) has been an influential resource for machine translation research ever since its first release over

[1] http://www.iwslt.org

a decade ago. It is freely available and includes parallel text for 21 European languages. Test sets and training data that enables research on machine translation of texts from the *News* domain have regularly been released for the shared translation task of the Workshop on Statistical Machine Translation[2] (WMT). The WMT "newstest" corpora have become important test sets to measure progress in machine translation between English and several European languages.

Previous research has indicated that divergences of domains such as *TED talks*, *Europarl*, and *News* empirically matter for machine translation. For instance, Ruiz and Federico (2014) systematically analyze characteristics of TED talk transcripts and News Commentary texts and point out the differences in detail for English-German.

4 Adaptation Techniques

In order to adapt systems to a domain, we leverage previously proposed techniques. First, we tune the model weights on in-domain development sets. Secondly, we linearly interpolate language models: rather than training a single large LM on all the target-side data, we train separate models on each corpus and interpolate them based on weights that minimize perplexity over the development set, resulting in a new, domain-adapted large LM that can be used by the decoder. Finally, we add binary features indicating the provenance of phrase pairs: if a phrase pair has been seen in a particular training corpus, a binary indicator associated with the respective training corpus fires on application of that phrase pair during decoding. This increases the amount of features by a number equal to the number of parallel training corpora.

5 Domain Classification

A domain classifier is required for multi-domain SMT on unlabeled input data. The domain of the source-side text we receive for translation is unknown and we need to predict it in order to select appropriate decoder parameters.

We investigate classification based on source-side LMs as well as different variations of a maximum entropy classifier.

Unlike Xu et al. (2007), who classify documents, we predict the domain label on the level of single sentences. Sentence-level classification has the advantage that document boundaries do not need to be present, and we are able to decode an unstructured incoming stream of sentences.

5.1 Source LM Classifier

Classification based on source-side LMs predicts the domain label from LM scores. We train separate LMs on the source side of each parallel training corpus. Then we create adapted LMs for each domain by linearly interpolating those source language LMs, where the interpolation weights are tuned to minimize perplexity on the source side of the respective in-domain development set. The classifier computes LM scores with each of the domain-adapted source LMs and selects the domain label according to maximum score. Note that for the scores to be on a comparable level, all domain-adapted source LMs should be interpolations of the same set of individual LMs.

Besides classifying sentences rather than documents, our method differs from the one proposed in (Xu et al., 2007) with respect to another aspect: Xu et al. (2007) interpolate LMs trained on the respective in-domain development sets with a single huge generic LM. Disadvantages of their method are (1.) the tiny size of the development corpora in terms of LM training, and (2.) the necessity of setting the interpolation weights heuristically. We overcome these drawbacks by resorting to a more straightforward framework of reserving the in-domain devel-

[2] http://www.statmt.org/wmt15/translation-task.html

opment sets for tuning source LM interpolation weights. Furthermore, we argue that source-side scores of the interpolated LMs employed in our classifier may to some extent resemble target-side scores of the interpolated LMs which are applied in the respective domain-adapted SMT systems.

5.2 Maximum Entropy Classifiers

Maximum entropy text classifiers can utilize a larger number of features in order to predict the label (Berger et al., 1996). We incorporate features from single words, pairs of adjacent words, the first word of the sentence, and the last word of the sentence. The model is trained with L-BFGS (Byrd et al., 1995) and regularized using a Gaussian prior.

We build maximum entropy (ME) classifiers under two different training conditions: using the MT development sets (which are rather small) as training data, and using selected other corpora as training data (which might not always exactly match what is defined as in-domain to the MT systems, as the development sets essentially constitute the domains).

In a further flavor of our ME classifiers, in addition to the previously described features, we include source LM indicator features in the ME model. To create these features, we score the sentence with the same domain-adapted source LMs as employed by the source LM classifier. For each of the domain-adapted LMs, an associated source LM indicator feature fires if the respective LM yields the maximum LM score.

Overall, we end up with four variations:

ME_{train} Classifier trained on medium-sized training corpora with the basic set of features.

$\text{ME}_{\text{train+lm}}$ Classifier trained on medium-sized training corpora with the basic set of features plus source LM indicator features.

ME_{dev} Classifier trained on the MT development sets with the basic set of features.

$\text{ME}_{\text{dev+lm}}$ Classifier trained on the MT development sets with the basic set of features plus source LM indicator features.

6 Experimental Setup

We use Moses (Koehn et al., 2007) for machine translation, MGIZA++ (Gao and Vogel, 2008) to train word alignments, KenLM (Heafield, 2011) for LM training and scoring, SRILM (Stolcke, 2002) for LM interpolation, and the Stanford Classifier[3] for maximum entropy text classification. We present experimental results on English→German, English→Italian, English→Portuguese, and English→Greek translation tasks.

6.1 Training Data

Our SMT systems are trained with the following bilingual corpora:

- TED from WIT3 (Cettolo et al., 2012)
- Europarl (Koehn, 2005)
- JRC-Acquis 3.0 (Steinberger et al., 2006)
- DGT's Translation Memory (Steinberger et al., 2012) as distributed in OPUS (Tiedemann, 2012)
- OPUS European Central Bank (ECB)
- OPUS European Medicines Agency (EMEA)
- OPUS EU Bookshop
- OPUS OpenSubtitles[4]

[3] http://nlp.stanford.edu/software/classifier.shtml

[4] http://www.opensubtitles.org

Parallel training corpus	En→De	En→It	En→Pt	En→El
sentences	23.97 M	29.55 M	32.15 M	30.79 M
source running words	464.82 M	445.97 M	427.53 M	417.45 M
source vocabulary	2.21 M	1.54 M	1.32 M	1.33 M
target running words	421.67 M	444.26 M	424.03 M	367.84 M
target vocabulary	3.81 M	1.62 M	1.39 M	2.19 M

Table 1: Statistics of the overall parallel training data after preprocessing.

Monolingual training corpus		De	It	Pt	El
Wikipedia	sentences	35.80 M	15.37 M	9.27 M	1.98 M
	running words	695.84 M	387.51 M	216.71 M	44.58 M
	vocabulary	6.50 M	2.42 M	1.82 M	0.93 M
News	sentences	159.66 M	—	—	—
	running words	2940.44 M	—	—	—
	vocabulary	8.83 M	—	—	—

Table 2: Statistics of additional monolingual training data after preprocessing.

- WMT News Commentary
- WMT CommonCrawl
- SETimes (Tyers and Alperen, 2010)

Statistics of a concatenation of all bilingual training corpora are presented in Table 1.

For language modeling on the target side, we furthermore add monolingual corpora from recent (April 2015) Wikipedia database dumps[5] and—for German—the News Crawl corpora provided for the WMT 2015 shared translation task. Plain text was obtained from the Wikipedia XML dumps with the Wikipedia Extractor[6] tool. Statistics of the additional monolingual training corpora are presented in Table 2.

6.2 Machine Translation Systems

Word alignments are created by aligning the data in both directions and symmetrizing the two trained alignments (Och and Ney, 2003; Koehn et al., 2003). We extract phrases up to a maximum length of five. The MT systems comprise these features:

- Phrase translation log-probabilities, smoothed with Good-Turing smoothing (Foster et al., 2006), and lexical translation log-probabilities in both directions.
- Phrase penalty and word penalty.
- Distance-based distortion cost.
- A hierarchical lexicalized reordering model (Galley and Manning, 2008).
- A 5-gram operation sequence model (Durrani et al., 2013a).
- Seven binary features indicating absolute occurrence count classes of phrase pairs.
- Sparse phrase length features.
- Sparse lexical features for the top 200 words.
- A 5-gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). We discard singleton *n*-grams of order three and higher.

Feature weights are optimized to maximize BLEU (Papineni et al., 2002) with batch MIRA (Cherry and Foster, 2012) on 1000-best lists. We prune the phrase table to a maximum of 100

[5] `http://dumps.wikimedia.org`

[6] `https://github.com/bwbaugh/wikipedia-extractor`

best translation options per distinct source side and apply a minimum score threshold of 0.0001 on the source-to-target phrase translation probability. We use cube pruning in decoding. Pop limit and stack limit are set to 1000 for tuning and to 5000 for testing. We disallow reordering over punctuation. Furthermore, Minimum Bayes Risk decoding is employed for testing. Translation quality is measured in truecase with BLEU.

6.2.1 Development and Test Sets

Wherever possible, we tune and test on common sets as distributed on `http://matrix.statmt.org/test_sets/list/` and `https://wit3.fbk.eu/`. The domain-adapted Europarl systems are tuned on the `test2006` set. The domain-adapted TED systems are tuned on a concatenation of `TED-dev2010` and `TEDX-dev2012` (En→De), on a concatenation of `TED-dev2010` and `TEDX-dev2014` (En→It), and on `TED-dev2010` (En→Pt). An English→Greek translation task was so far never organized as part of any of the IWSLT evaluation campaigns and for that reason no common TED sets exist for that language pair. However, the 2012-02 release of the WIT3 corpus contains a parallel corpus of 84 831 English-Greek sentence pairs. We reserved every fifth sentence of this data for tuning and testing. Of the tuning and testing part of the corpus, we assign every fourth sentence to the test set and the rest to the tuning set.[7] The domain-adapted News systems are tuned on a concatenation of the `newstest2008-2012` sets (En→De) and on `newstest2009` (En→It). We use `news-syscomb2009` as an English→Italian News domain test set for lack of other English-Italian News test sets. Note that `newssyscomb2009` is a small set of only 502 sentences. No News test data was available to us for the English→Portuguese and English→Greek language pairs, so we experiment with only two domains (TED and Europarl) on these tasks.

The Portuguese *TED* development and test sets are Brazilian Portuguese whereas the *Europarl* sets are European Portuguese. The two Portuguese dialects have a number of differences in written language. Marujo et al. (2011) give a brief overview.

6.2.2 Domain-Adapted SMT

For our domain adaptation experiments, we first tune the systems with the features described above on the respective in-domain development set (*TED-tuned*, *Europarl-tuned*, *News-tuned*). We next replace the large baseline LM with a domain-specific interpolated LM (+ *LM interp.*). We then add binary features indicating the provenance of phrase pairs (+ *LM interp.* + *indicator feat.*).

6.2.3 Mixed-Domain SMT

We build mixed-domain SMT systems by tuning on a development corpus containing samples of texts from all domains. We include a balanced amount of development data from the different domains in the mixed-domain development set in order to avoid a bias towards any specific domain.

The mixed-domain systems (*Mixed-domain-tuned*) are tuned on a concatenation of `TED-dev2010` and `TEDX-dev2012` and Europarl `test2006` and `newstest2009` (En→De), on a concatenation of `TED-dev2010` and `TEDX-dev2014` and Europarl `test2006` and `newstest2009` (En→It), on a concatenation of `TED-dev2010` and every second sentence from Europarl `test2006` (En→Pt), and on a concatenation of every sixth sentence from our Greek TED development set and the full Europarl `test2006` (En→El).

The LMs for the mixed-domain systems are trained on the full target language monolingual training data, not interpolated from individual LMs. Binary features indicating the provenance of phrase pairs are not used.

[7] We end up with English-Greek TED corpus sizes of 67 865 sentences for training, 12 725 sentences for tuning, and 4 241 for testing.

Classifier	Accuracy [%]			
	En→De	En→It	En→Pt	En→El
LM	78.1	90.2	96.9	96.5
ME_{train}	83.1	92.4	96.6	96.7
$\text{ME}_{\text{train+lm}}$	84.8	92.5	96.9	96.6
ME_{dev}	78.1	83.1	91.0	93.0
$\text{ME}_{\text{dev+lm}}$	81.8	88.4	96.9	96.6

Table 3: Domain classifier accuracies on the English side of a concatenation of all test sets for the respective language pairs.

6.2.4 Multi-Domain SMT

Multi-domain systems classify the input sentence with a domain classifier. They parameterize the decoder according to the predicted domain label. We use the parameters of the *Domain-tuned + LM interp. + indicator feat.* MT setups. We evaluate five multi-domain system per language pair, one for each of our domain classifiers.

6.2.5 Oracle-Domain SMT

In an oracle domain setup, we assume that the correct domain label of each input sentence is given. We can parameterize the decoder according to the gold-standard domain label.

6.3 Domain Classifiers

The ME_{train} classifiers are trained on the source language side of the TED portion of the training data and fractions of both the Europarl portion of the training data and the English News Crawl 2014 corpus as provided for the WMT 2015 shared translation task. Again, we include a balanced amount of data from the different domains (e.g. 10% of the Europarl data and 1% of the News Crawl 2014 data for English→German) in order to not give preference to any of the domains. The ME_{dev} classifiers are trained on the MT development sets as used for mixed-domain MT tuning.

While building common English domain classifiers would be possible, we decided to train separate ones for each task and utilize the data resources from the respective language pair.

7 Experimental Results

Domain classifiers. The accuracies of the domain classifiers are presented in Table 3. We report accuracies (micro-averaged F1) on a concatenation of all test sets for each of the language pairs with the source LM classifier and four variations of the ME classifier (cf. Section 5). Naturally, accuracies are higher for the tasks where only two domains have to be distinguished (En→Pt, En→El) than on the tasks with three domain classes (En→De, En→It). Accuracies are generally of a high level, even for the simple source LM classifier. We are going to evaluate in MT experiments whether any of the differences in classification accuracy carry over to translation quality of multi-domain systems.

Translation quality. Tables 4-7 contain BLEU scores obtained with all MT systems on the test sets from the various domains for the four language pairs. The *TED* test sets are the common IWSLT tests sets, the *Europarl* test sets have been downloaded from `matrix.statmt.org`, and the *News* test sets are the standard ones from the WMT shared tasks. We test all systems on the sets from all domains, in particular, domain-adapted systems are tested on out-of-domain sets as well. We also report BLEU scores on concatenations of all test sets (*all*) in order to measure overall performance in open-domain scenarios.

System (En→De)	**TED**				**Europarl**		**News**			**all**
	2010	**2011**	**2012**	**2013**	**2007**	**2008**	**2013**	**2014**	**2015**	
TED-tuned	24.3	26.7	22.9	25.3	21.9	21.7	20.2	20.8	22.9	22.4
+ LM interp.	24.2	26.8	22.7	24.6	21.2	20.9	19.5	20.1	22.2	21.9
+ LM interp. + indicator feat.	24.4	26.7	23.1	25.0	20.9	20.7	19.5	20.2	22.3	21.9
Europarl-tuned	24.4	26.3	22.8	25.2	22.4	22.1	20.4	21.2	23.0	22.6
+ LM interp.	23.1	24.8	22.0	24.0	22.5	22.2	19.4	19.7	21.8	21.8
+ LM interp. + indicator feat.	22.8	24.7	21.7	24.0	22.6	22.1	19.3	19.5	21.6	21.7
News-tuned	23.7	26.2	22.2	24.5	21.5	21.4	20.6	21.1	22.9	22.2
+ LM interp.	23.3	25.8	22.1	24.2	21.3	21.2	20.4	20.9	22.6	21.9
+ LM interp. + indicator feat.	23.4	25.7	22.2	24.4	21.4	21.0	20.5	20.9	22.6	21.9
Mixed-domain-tuned	24.6	26.9	23.1	25.3	22.2	22.0	20.6	21.2	23.2	22.7
Multi-domain, LM classifier	24.4	26.8	23.1	24.9	22.5	22.1	20.1	20.6	22.5	22.4
Multi-domain, ME_{train} classifier	24.3	26.7	22.8	24.7	22.5	22.1	20.3	20.7	22.5	22.4
Multi-domain, $\text{ME}_{\text{train+lm}}$ classifier	24.3	26.7	23.0	24.9	22.5	22.2	20.4	20.8	22.5	22.5
Multi-domain, ME_{dev} classifier	24.2	26.8	22.8	24.7	22.5	22.1	20.3	20.8	22.5	22.4
Multi-domain, $\text{ME}_{\text{dev+lm}}$ classifier	24.3	26.9	22.9	24.9	22.5	22.1	20.3	20.7	22.5	22.5
Oracle-domain	24.4	26.7	23.1	25.0	22.6	22.1	20.5	20.9	22.6	22.5

Table 4: English→German experimental results (truecase BLEU scores).

System (En→It)	**TED**				**Europarl**		**News**	**all**
	2010	**2011**	**2012**	**2013**	**2007**	**2008**	**syscomb2009**	
TED-tuned	26.8	26.6	27.3	32.9	24.6	25.3	27.5	26.9
+ LM interp.	27.0	27.3	27.7	33.5	24.3	24.7	27.3	27.0
+ LM interp. + indicator feat.	27.3	27.7	28.0	33.4	24.1	24.6	27.5	27.0
Europarl-tuned	24.6	24.3	25.5	30.5	24.9	25.6	26.5	25.8
+ LM interp.	23.9	24.4	25.6	30.0	24.9	25.4	25.2	25.6
+ LM interp. + indicator feat.	24.2	24.3	25.8	29.9	25.0	25.5	25.2	25.7
News-tuned	26.8	26.5	27.2	32.5	24.6	25.3	27.7	26.9
+ LM interp.	26.8	27.4	28.0	33.3	24.7	25.1	27.8	27.2
+ LM interp. + indicator feat.	26.9	27.5	27.9	33.1	24.6	25.1	27.8	27.1
Mixed-domain-tuned	26.9	26.7	27.0	32.5	25.0	25.6	27.9	27.0
Multi-domain, LM classifier	27.3	27.7	27.9	33.4	24.9	25.5	27.3	27.2
Multi-domain, ME_{train} classifier	27.3	27.6	27.8	33.3	25.0	25.5	26.1	27.1
Multi-domain, $\text{ME}_{\text{train+lm}}$ classifier	27.3	27.6	27.9	33.3	25.0	25.5	25.8	27.1
Multi-domain, ME_{dev} classifier	27.3	27.7	27.9	33.1	25.0	25.4	27.4	27.1
Multi-domain, $\text{ME}_{\text{dev+lm}}$ classifier	27.3	27.8	27.9	33.3	24.9	25.5	27.6	27.2
Oracle-domain	27.3	27.7	28.0	33.4	25.0	25.5	27.8	27.2

Table 5: English→Italian experimental results (truecase BLEU scores).

Domain adaptation. As expected, the domain-adapted systems perform better on in-domain data than when evaluated in a cross-domain experiment. The in-domain systems outperform systems tuned on out-of-domain development sets. More aggressive domain adaptation via LM interpolation and binary provenance indicator features gives mixed results. For English→German, we basically do not observe gains over in-domain tuning with any of the two adaptation methods. They just further reduce quality on out-of-domain data. For English→Italian, we observe gains with both methods on in-domain test sets on top of the *TED-tuned* system, but not on top of the *Europarl-tuned* and *News-tuned* systems. For English→Portuguese and English→Greek, we see larger gains mostly due to LM interpolation and on TED-domain adaptation.

System (En→Pt)	TED					Europarl		all
	2010	2011	2012	2013	2014	2007	2008	
TED-tuned	32.1	34.0	34.8	33.9	32.8	27.6	27.7	30.6
+ LM interp.	33.7	35.2	36.1	34.9	34.6	25.5	25.2	30.2
+ LM interp. + indicator feat.	33.7	35.6	36.2	35.4	35.2	24.9	24.5	30.0
Europarl-tuned	29.9	32.0	31.0	31.6	31.2	30.1	30.1	30.7
+ LM interp.	26.7	28.4	27.5	28.6	28.2	30.3	30.1	29.1
+ LM interp. + indicator feat.	26.5	27.7	27.1	28.0	27.6	30.4	30.1	28.8
Mixed-domain-tuned	32.3	34.1	33.4	33.2	33.5	29.4	29.5	31.6
Multi-domain, LM classifier	33.7	35.4	35.9	35.1	35.0	30.4	30.1	32.6
Multi-domain, ME$_{\text{train}}$ classifier	33.6	35.4	35.9	35.0	34.8	30.4	30.1	32.6
Multi-domain, ME$_{\text{train+lm}}$ classifier	33.7	35.4	35.9	35.1	34.9	30.4	30.1	32.6
Multi-domain, ME$_{\text{dev}}$ classifier	33.3	34.9	35.4	34.6	34.3	30.2	29.9	32.3
Multi-domain, ME$_{\text{dev+lm}}$ classifier	33.7	35.4	35.9	35.1	35.0	30.4	30.1	32.6
Oracle-domain	33.7	35.6	36.2	35.4	35.2	30.4	30.1	32.8

Table 6: English→Portuguese experimental results (truecase BLEU scores).

System (En→El)	TED	Europarl		all
		2007	2008	
TED-tuned	28.2	25.3	24.6	26.3
+ LM interp.	29.0	24.7	24.3	26.4
+ LM interp. + indicator feat.	28.9	24.6	24.2	26.3
Europarl-tuned	27.0	25.6	25.0	26.0
+ LM interp.	25.7	25.6	25.3	25.7
+ LM interp. + indicator feat.	25.7	25.6	25.1	25.6
Mixed-domain-tuned	27.5	25.6	25.0	26.2
Multi-domain, LM classifier	28.9	25.6	25.1	26.8
Multi-domain, ME$_{\text{train}}$ classifier	28.9	25.6	25.1	26.8
Multi-domain, ME$_{\text{train+lm}}$ classifier	28.9	25.6	25.1	26.8
Multi-domain, ME$_{\text{dev}}$ classifier	28.8	25.6	25.1	26.7
Multi-domain, ME$_{\text{dev+lm}}$ classifier	28.9	25.6	25.1	26.8
Oracle-domain	28.9	25.6	25.1	26.8

Table 7: English→Greek experimental results (truecase BLEU scores).

On *TED*, our domain-adapted English→Italian and English→Portuguese systems outperform the best submissions from recent IWSLT evaluation campaigns by several BLEU points.[8] On *News*, our domain-adapted English→German systems are on par with the best phrase-based system submissions at the WMT shared translation task.[9]

Mixed-domain vs. multi-domain SMT. Looking at the performance on the concatentation of all test sets, mixed-domain SMT yields a higher BLEU score than any of the domain-adapted systems on two out of four language pairs (En→De: +0.3; En→It: -0.2; En→Pt: +0.9; En→El: -0.2). Apart from English→Portuguese, the differences are small.

Multi-domain SMT clearly outperforms mixed-domain SMT for English→Portuguese (up to +1.0 on *all*) and English→Greek (up to +0.6 on *all*). The choice of the domain classifier barely matters wrt. translation quality. Due to its compact model, the ME$_{\text{dev}}$ classifier would for instance be a reasonable choice despite not providing the highest classification accuracy.

[8] MT English→Italian: +4.3 points BLEU on `tst2013` (Cettolo et al., 2013). MT English→Portuguese: +2.8 points BLEU on `tst2014` (Cettolo et al., 2014).

[9] `http://matrix.statmt.org`

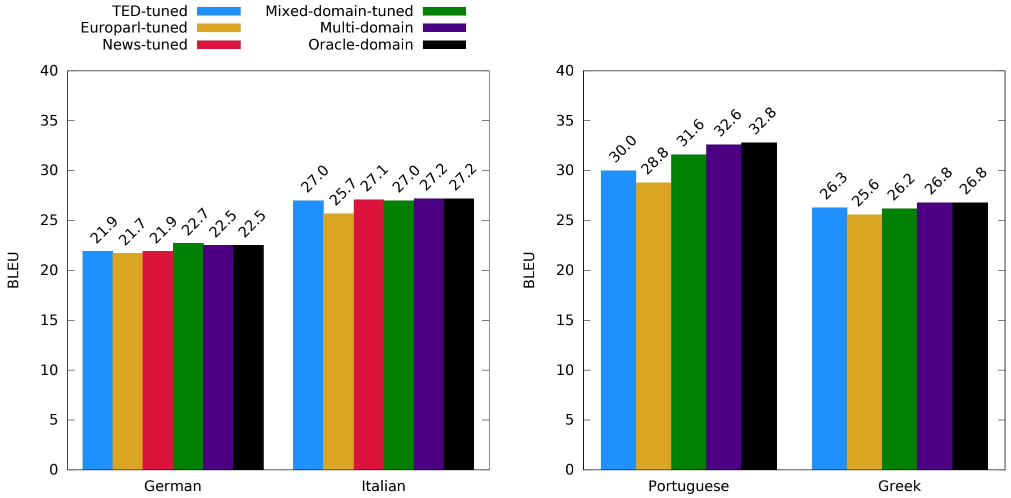

Figure 1: BLEU scores on a concatenation of all test sets.

Compared to oracle-domain SMT, which is equivalent to choosing the respective in-domain translation from the *Domain-tuned + LM interp. + indicator feat.* system, the best multi-domain results are on the same level of quality across the board, with a maximum drop of 0.2 points BLEU (En→Pt).

We visualized the results in a couple of plots (Figures 1-4). We use *Domain-tuned* as a shortcut for *Domain-tuned + LM interp. + indicator feat.* in all plots, i.e. the in-domain system results in the plots include LM interpolation and the provenance indicator. *Multi-domain* in the plots is the variant based on the $\mathrm{ME}_{\mathrm{dev+lm}}$ classifier.

BLEU histograms on the concatenation of all test sets are shown in Figure 1. The figure illustrates the results we just discussed, on the concatentation of all test sets. In Figures 2–4 we plotted average BLEU differences wrt. the in-domain system on the *TED*, *Europarl*, and *News* test sets. In terms of averaged BLEU scores over the in-domain test sets, in-domain systems are up to 7.8 points BLEU better than out-of-domain systems. Multi-domain SMT is up to 1.7 points BLEU better than mixed-domain SMT but can also perform minimally worse in some cases, for instance English→German *TED* and *News*, where mixed-domain SMT performs better than in-domain SMT. However, multi-domain SMT is typically on par with in-domain SMT.

8 Conclusion

While mixed-domain tuning worked for half of the language pairs, our results indicate that multi-domain SMT is the more reliable choice. Multi-domain SMT is always on par with in-domain SMT translation quality and under some circumstances mixed-domain SMT can perform much worse.

Multi-domain SMT can be easily implemented with a domain classifier and by allowing for run-time reconfiguration of the decoder with domain-specific weight vectors. Satisfactory domain classification accuracy can be achieved with a simple and compact maximum entropy text classifier trained on the small MT development sets and applied at the sentence level.

For scenarios where the model is expected to translate a wide variety of input text, the approach presented in this paper balances ease of implementation with high performance.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements n^{o} 644333 (TraMOOC) and 644402 (HimL).

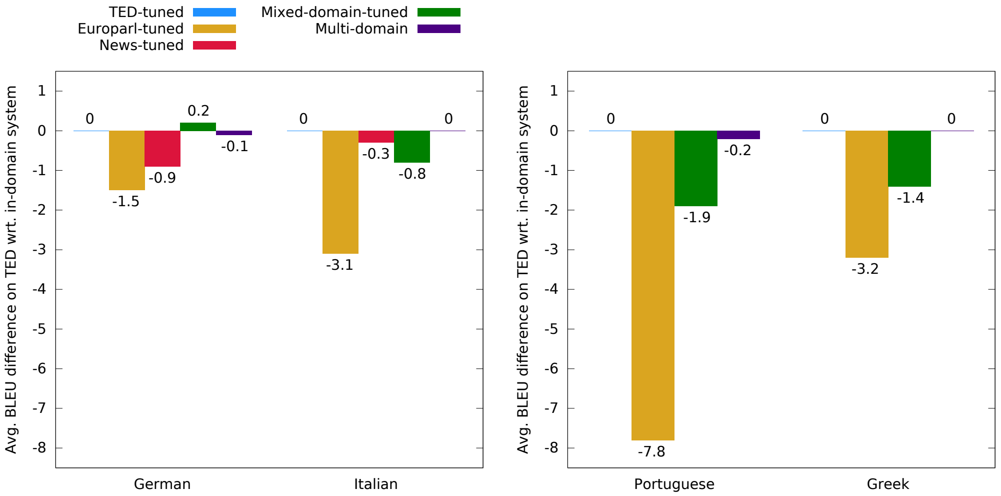

Figure 2: Average BLEU differences on TED test sets.

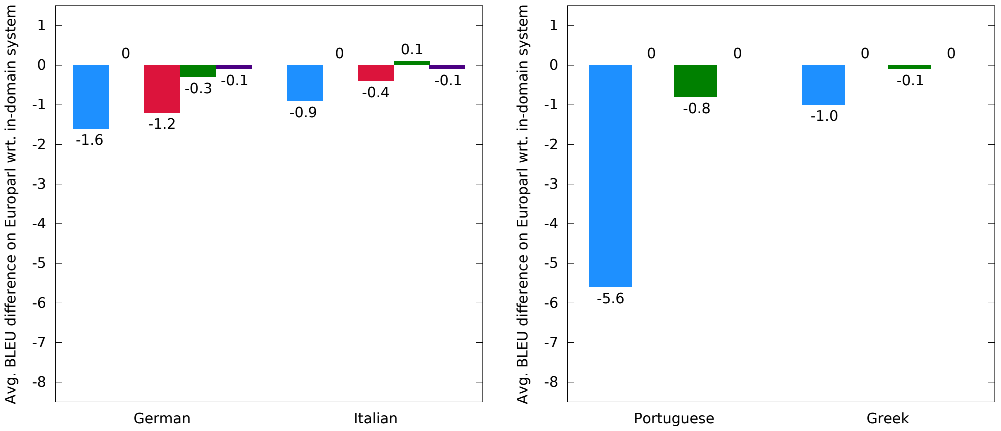

Figure 3: Average BLEU differences on Europarl test sets.

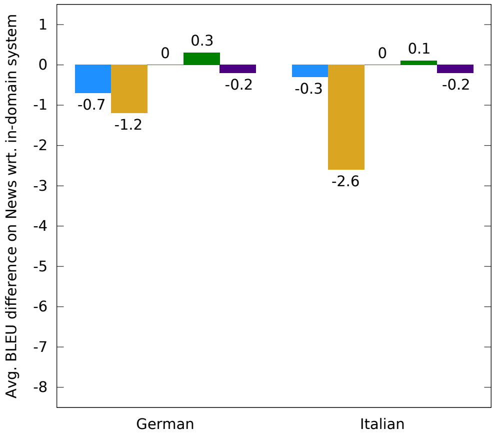

Figure 4: Average BLEU differences on News test sets.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-domain Data Selection. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 355–362, Edinburgh, Scotland, UK.

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 *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, USA.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–72.

Bertoldi, N. and Federico, M. (2009). Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 182–189, Athens, Greece.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA, USA.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT Evaluation Campaign. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 15–32, Heidelberg, Germany.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 2–17, Lake Tahoe, CA, USA.

Chen, B., Foster, G., and Kuhn, R. (2013). Adaptation of Reordering Models for Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 938–946, Atlanta, GA, USA.

Chen, S. F. and Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, USA.

Cherry, C. and Foster, G. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 427–436, Montréal, Canada.

Clark, J. H., Lavie, A., and Dyer, C. (2012). One System, Many Domains: Open-Domain Statistical Machine Translation via Feature Augmentation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA.

Durrani, N., Fraser, A., and Schmid, H. (2013a). Model With Minimal Translation Units, But Decode With Phrases. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 1–11, Atlanta, GA, USA.

Durrani, N., Haddow, B., Heafield, K., and Koehn, P. (2013b). Edinburgh's Machine Translation Systems for European Language Pairs. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 114–121, Sofia, Bulgaria.

Federico, M., Bentivogli, L., Paul, M., and Stueker, S. (2011). Overview of the IWSLT 2011 Evaluation Campaign. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 11–27, San Francisco, CA, USA.

Federico, M. and Bertoldi, N. (2012). Practical Domain Adaptation in SMT. Tutorial at the Conf. of the Assoc. for Machine Translation in the Americas (AMTA). `http://www.mt-archive.info/10/AMTA-2012-Bertoldi-ppt.pdf`.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 451–459, Cambridge, MA, USA.

Foster, G. and Kuhn, R. (2007). Mixture-Model Adaptation for SMT. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 128–135, Prague, Czech Republic.

Foster, G., Kuhn, R., and Johnson, H. (2006). Phrasetable Smoothing for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 53–61, Sydney, Australia.

Galley, M. and Manning, C. D. (2008). A Simple and Effective Hierarchical Phrase Reordering Model. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 847–855, Honolulu, HI, USA.

Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Columbus, OH, USA.

Haddow, B. and Koehn, P. (2012). Analysing the Effect of Out-of-Domain Data on SMT Systems. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 422–432, Montréal, Canada.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 187–197, Edinburgh, Scotland, UK.

Kneser, R. and Ney, H. (1995). Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proc. of the MT Summit X*, Phuket, Thailand.

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 *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 127–133, Edmonton, Canada.

Koehn, P. and Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 224–227, Prague, Czech Republic.

Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011). Investigations on Translation Model Adaptation Using Monolingual Data. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 284–293, Edinburgh, Scotland, UK.

Macherey, W., Och, F., Thayer, I., and Uszkoreit, J. (2008). Lattice-based Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 725–734, Honolulu, HI, USA.

Mansour, S., Al-Onaizan, Y., Blackwood, G., and Tillmann, C. (2014). Automatic Dialect Classification for Statistical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Vancouver, BC, Canada.

Mansour, S. and Ney, H. (2012). A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 193–200, Hong Kong.

Marujo, L., Grazina, N., Luis, T., Ling, W., Coheur, L., and Trancoso, I. (2011). BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, pages 129–136, Leuven, Belgium.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative Corpus Weight Estimation for Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 708–717, Singapore.

Moore, R. C. and Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 220–224, Uppsala, Sweden.

Nakov, P. (2008). Improving English-Spanish Statistical Machine Translation: Experiments in Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasing. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 147–150, Columbus, OH, USA.

Niehues, J. and Waibel, A. (2012). Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA.

Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, 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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA.

Pecina, P., Toral, A., and van Genabith, J. (2012). Simple and Effective Parameter Tuning for Domain Adaptation of Statistical Machine Translation. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, pages 2209–2224, Mumbai, India.

Ruiz, N. and Federico, M. (2014). Complexity of Spoken Versus Written Language for Machine Translation. In *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, pages 173–180, Dubrovnik, Croatia.

Salloum, W., Elfardy, H., Alamir-Salloum, L., Habash, N., and Diab, M. (2014). Sentence Level Dialect Identification for Machine Translation System Selection. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 772–778, Baltimore, MD, USA.

Schwenk, H. and Senellart, J. (2009). Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In *Proc. of the MT Summit XII*, Ottawa, Canada.

Shah, K., Barrault, L., and Schwenk, H. (2012). A General Framework to Weight Heterogeneous Parallel Data for Model Adaptation in Statistical MT. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. In *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, pages 454–459, Istanbul, Turkey.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, pages 2142–2147, Genoa, Italy.

Stolcke, A. (2002). SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, volume 3, Denver, CO, USA.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, pages 2214–2218, Istanbul, Turkey.

Tyers, F. M. and Alperen, M. S. (2010). South-East European Times: A parallel corpus of Balkan languages. In *Procȯf the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Malta.

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

Wang, W., Macherey, K., Macherey, W., Och, F., and Xu, P. (2012). Improved Domain Adaptation for Statistical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA.

Xu, J., Deng, Y., Gao, Y., and Ney, H. (2007). Domain Dependent Statistical Machine Translation. In *Proc. of the MT Summit XI*, pages 515–520, Copenhagen, Denmark.