# Unsupervised Adaptation for Statistical Machine Translation

**Saab Mansour** and **Hermann Ney**
Human Language Technology and Pattern Recognition
Computer Science Department
RWTH Aachen University
Aachen, Germany
{mansour,ney}@cs.rwth-aachen.de

## Abstract

In this work, we tackle the problem of language and translation models domain-adaptation without explicit bilingual in-domain training data. In such a scenario, the only information about the domain can be induced from the source-language test corpus. We explore unsupervised adaptation, where the source-language test corpus is combined with the corresponding hypotheses generated by the translation system to perform adaptation. We compare unsupervised adaptation to supervised and pseudo supervised adaptation. Our results show that the choice of the adaptation (target) set is crucial for successful application of adaptation methods. Evaluation is conducted over the German-to-English WMT newswire translation task. The experiments show that the unsupervised adaptation method generates the best translation quality as well as generalizes well to unseen test sets.

## 1 Introduction

Over the last few years, large amounts of statistical machine translation (SMT) monolingual and bilingual corpora were collected. Early years focused on structured data translation such as newswire. Nowadays, due to the relative success of SMT, new domains of translation are being explored, such as lecture and patent translation (Cettolo et al., 2012; Goto et al., 2013).

The task of domain adaptation tackles the problem of utilizing existing resources mainly drawn from one domain (e.g. parliamentary discussion) to maximize the performance on the target (test) domain (e.g. newswire).

To be able to perform adaptation, a *target set* representing the test domain is used to manipulate the general-domain models. Previous work on SMT adaptation focused on the scenario where (small) bilingual in-domain or pseudo in-domain training data are available. Furthermore, small attention was given to the choice of the target set for adaptation. In this work, we explore the problem of adaptation where no explicit bilingual data from the test domain is available for training, and the only resource encapsulating information about the domain is the source-language test corpus itself.

We explore how to utilize the source-language test corpus for adapting the language model (LM) and the translation model (TM). A combination of source and automatically translated target of the test set is compared to using the source side only for TM adaptation. Furthermore, we compare using the test set to using in-domain data and a pseudo in-domain data (e.g. news-commentary as opposed to newswire).

Experiments are done on the WMT 2013 German-to-English newswire translation task. Our best adaptation method shows competitive results to the best submissions of the evaluation.

This paper is structured as follows. We review related work in Section 2 and introduce the basic adaptation methods in Section 3. The experimental setup is described in Section 4, results are discussed in Section 5 and we conclude in Section 6.

## 2 Related Work

A broad range of methods and techniques have been suggested in the past for domain adaptation for both SMT and automatic speech recognition (ASR).

For ASR, (Bellegarda, 2004) gives an overview of LM adaptation methods. He differentiates between two cases regarding the availability of in-domain adaptation data: *(i)* the data is available and can be directly used to manipulate a background (general domain) corpus, and *(ii)* the data is not available or too small, and then it can be gathered or automatically generated during the

*Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 457–465,
Baltimore, Maryland USA, June 26–27, 2014. ©2014 Association for Computational Linguistics

recognition process. (Bacchiani and Roark, 2003) compare supervised against unsupervised (using automatic transcriptions) in-domain data for LM training for the task of ASR. They show that augmenting the supervised in-domain to the training of the LM performs better than the unsupervised in-domain. In addition, they perform "self-training", where the test set is automatically transcribed and added to the LM. When using a strong baseline, no improvements in recognition quality are achieved. We differ from their work by using the unsupervised test data to adapt a general-domain bilingual corpus. We also performed initial experiments of "self-training" for language modeling, where (artificial) perplexity improvement was achieved but without an impact on the machine translation (MT) quality.

(Zhao et al., 2004) tackle LM adaptation for SMT. Similarly to our work, they use automatically generated hypotheses to perform adaptation. We extend their work by using the hypotheses also for TM adaptation. (Hildebrand et al., 2005) perform LM and TM adaptation based on information retrieval methods. They use the source-language test corpus to filter the bilingual data, and then use the target side of the filtered bilingual data to perform LM adaptation. We differ from their work by using both the in-domain source-language corpus and its corresponding automatic translation for adaptation, which is shown in our experiments to achieve superior results than when using the source-side information only. (Foster and Kuhn, 2007) perform LM and TM adaptation using mixture modeling. In their setting, the mixture weights are modified to express adaptation. They compare cross-domain (in-domain available) against dynamic adaptation. In the dynamic adaptation scenario, they utilize the source side of the development set to adapt the mixture weights (LM adaptation is possible as they only use parallel training data, which enables filtering based on the source side and then keeping the corresponding target side of the data). For an in-domain test set, the cross-domain setup performs better than the dynamic adaptation method. (Ueffing et al., 2007) use the test set translations as additional data to train the TM. One important aspect in their work is confidence measurement to remove noisy translation. In our approach, we use the automatic test set translations to adapt the SMT models rather than augmenting it as additional TM data. We also compare different adaptation sets. Furthermore, we do not use confidence measures to filter the automatic translations as they are only used to adapt the general-domain system and are not augmented to the TM.

In this work, we apply cross-entropy scoring for adaptation as done by (Moore and Lewis, 2010). Moore and Lewis (2010) apply adaptation by using an LM-based cross-entropy filtering for LM training. Axelrod et al. (2011) generalized the method for TM adaptation by interpolating the source and target LMs. These two works focused on a scenario where in-domain training data are available for adaptation. In this work, we focus on a scenario where in-domain training data is not labeled, and the main resource for adaptation is the source-language test data.

In recent WMT evaluations, the method of (Moore and Lewis, 2010) was utilized by several translation systems (Koehn and Haddow, 2012; Rubino et al., 2013). These systems use pseudo in-domain corpus, i.e., news-commentary, as the target domain (while the test domain is newswire). The contribution of this work is two fold: we show that the choice of the target set is crucial for adaptation, in addition, we show that an unsupervised target set performs best in terms of translation quality as well as generalization performance to unseen test sets (in comparison to using pseudo in-domain data or the references as target sets).

## 3 Cross-Entropy Adaptation

In this work, we use sample scoring for the purpose of adaptation. We start by introducing the scoring framework and then show how we utilize it to perform filtering based adaptation and weighted phrase extraction based adaptation.

LM cross-entropy scores can be used for both monolingual data weighting for LM training as done by (Moore and Lewis, 2010), or bilingual weighting for TM training as done by (Axelrod et al., 2011).

We differentiate between two types of data sets: the *adaptation set* (target) representative of the test-domain which we refer to also as in-domain (IN), and the general-domain (GD) set which we want to adapt.

The scores for each sentence in the general-domain corpus are based on the cross-entropy difference of the IN and GD models. Denoting $H_M(x)$ as the cross entropy of sentence $x$ accordinging to model $M$, then the cross entropy difference $DH_M(x)$ can be written as:

$$DH_M(x) = H_{M_{IN}}(x) - H_{M_{GD}}(x) \quad (1)$$

The intuition behind eq. (1) is that we are interested in sentences as close as possible to the in-domain, but also as far as possible from the general corpus. Moore and Lewis (2010) show that using eq. (1) for LM filtering performs better in terms of perplexity than using in-domain cross-entropy only ($H_{M_{IN}}(x)$). For more details about the reasoning behind eq. (1) we refer the reader to (Moore and Lewis, 2010).

Axelrod et al. (2011) adapted eq. (1) for bilingual data filtering for the purpose of TM training. The bilingual LM cross entropy difference for a sentence pair $(f_r, e_r)$ in the GD corpus is then defined by:

$$DH_{LM}(f_r, e_r) = DH_{LM_{src}}(f_r) + DH_{LM_{trg}}(e_r) \quad (2)$$

For IBM Model 1 (M1), the cross-entropy $H_{M1}(f_r|e_r)$ is defined similarly to the LM cross-entropy, and the resulting bilingual cross-entropy difference will be of the form:

$$DH_{M1}(f_r, e_r) = DH_{M1}(f_r|e_r) + DH_{M1}(e_r|f_r)$$

The combined LM+M1 score is obtained by summing the LM and M1 bilingual cross-entropy difference scores:

$$d_r = DH_{LM}(f_r, e_r) + DH_{M1}(f_r, e_r) \quad (3)$$

### 3.1 Filtering

A common framework to perform sample filtering is to score each sample according to a model, and then assigning a threshold on the score which filters out unwanted samples. If the score we generate is related to the probability that the sample was drawn from the same distribution as the in-domain data, we are selecting the samples most relevant to our domain. In this way we can achieve adaptation of the general-domain data.

We use the LM cross-entropy difference from eq. (1) for LM filtering and a combined LM+M1 score (eq. (3) for TM filtering. We sort the sentences in the general-domain according to the score and select the best 50%,25%,...,6.25% training instances. Our models are then trained on the selected portions of the training data, and the best performing portion (according to perplexity for LM training and BLEU for TM training) on the development set is chosen as the adapted corpus.

### 3.2 Weighted Phrase Extraction

The classical phrase model is trained using a "simple" maximum likelihood estimation, resulting in phrase translation probabilities being defined by relative frequency:

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r c_r(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} \sum_r c_r(\tilde{f}', \tilde{e})} \quad (4)$$

Here, $\tilde{f}, \tilde{e}$ are contiguous phrases, $c_r(\tilde{f}, \tilde{e})$ denotes the count of $(\tilde{f}, \tilde{e})$ being a translation of each other (usually according to word alignment and heuristics) in sentence pair $(f_r, e_r)$. One method to introduce weights to eq. (4) is by weighting each sentence pair by a weight $w_r$. Eq. (4) will now have the extended form:

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r w_r \cdot c_r(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} \sum_r w_r \cdot c_r(\tilde{f}', \tilde{e})} \quad (5)$$

It is easy to see that setting $\{w_r = 1\}$ will result in eq. (4) (or any non-zero equal weights). Increasing the weight $w_r$ of the corresponding sentence pair will result in an increase of the probabilities of the phrase pairs extracted. Thus, by increasing the weight of in-domain sentence pairs, the probability of in-domain phrase translations could also increase.

We utilize $d_r$ from eq. (3) using a combined LM+M1 scores for our suggested weighted phrase extraction. $d_r$ can be assigned negative values, and lower $d_r$ indicates sentence pairs which are more relevant to the in-domain. Therefore, we negate the term $d_r$ to get the notion of higher is closer to the in-domain, and use an exponent to ensure positive values. The final weight is of the form:

$$w_r = e^{-d_r} \quad (6)$$

This term is proportional to perplexities, as the exponent of entropy is perplexity by definition.

One could also use filtering for TM adaptation, but, as shown in (Mansour and Ney, 2012), filtering for TM could only reduce the size and weighting performs better than filtering.

## 4 Experimental Setup

### 4.1 Training Data

The experiments are done on the recent German-to-English WMT 2013 translation task [1]. For

[1] The translation task resources of WMT 2013 are available under: http://www.statmt.org/wmt13/

| Corpus | Sent | De | En |
|---|---|---|---|
| **Training data** | | | |
| news-commentary | 177K | 4.8M | 4.5M |
| europarl | 1 888K | 51.5M | 51.9M |
| common-crawl | 2 030K | 47.8M | 47.7M |
| total | 4 095K | 104.1M | 104M |
| **Test data** | | | |
| newstest08 | 2051 | 52446 | 49749 |
| newstest09 | 2525 | 68512 | 65648 |
| newstest10 | 2489 | 68232 | 62024 |
| newstest11 | 3003 | 80181 | 74856 |
| newstest12 | 3003 | 79912 | 73089 |
| newstest13 | 3000 | 69066 | 64900 |

Table 1: German-English bilingual training and test data statistics: the number of sentence pairs (Sent), German (De) and English (En) words are given.

German-English WMT 2013, the common-crawl bilingual corpus was introduced, enabling more impact for TM adaptation on the SMT system quality. Monolingual English data exists with more than 1 billion words, making LM adaptation and size reduction a wanted feature. We use *newstest08* throughout *newstest13* to evaluate the SMT systems. The baseline systems are built using all (unfiltered) available monolingual and bilingual training data. The bilingual corpora and the test data statistics are summarized in Table 1.

In Table 2, we summarize the size and LM perplexity of the different monolingual corpora for the German-English task over the LM development set *newstest09* and test set *newstest13*. The corpora are split into three parts, the English side of the bilingual side (*bi.en*), the giga-fren joined with undoc (*giun*) and the news-shuffle (*ns*) corpus. To keep the perplexity results comparable, we use the intersection vocabulary of the different corpora as a reference vocabulary. From the table, we notice that as expected, the in-domain corpus *news-shuffle* generate the best perplexity values.

### 4.2 SMT System

The baseline system is built using the open-source SMT toolkit Jane[2], which provides state-of-the-art phrase-based SMT system (Wuebker et al., 2012). We use the standard set of models with phrase translation probabilities for source-to-target and target-to-source directions, smoothing with lexical weights, a word and phrase penalty, distance-based reordering, hierarchical reordering model (Galley and Manning, 2008) and a 4-gram target language model. The baseline system is competitive and using adaptation we will show comparable results to the best systems of WMT 2013. The SMT system was tuned on the development set *newstest10* with minimum error rate training (MERT) (Och, 2003) using the BLEU (Papineni et al., 2002) error rate measure as the optimization criterion. We test the performance of our system on the *newstest08...newstest13* sets using the BLEU and translation edit rate (TER) (Snover et al., 2006) measures. We use TER as an additional measure to verify the consistency of our improvements and avoid over-tuning. All results are based on true-case evaluation. We perform bootstrap resampling with bounds estimation as described by (Koehn, 2004). We use the 90% and 95% (denoted by † and ‡ correspondingly in the tables) confidence thresholds to draw significance conclusions.

| Corpus | Tokens [M] | ppl dev | ppl test |
|---|---|---|---|
| bi.en | 88 | 216.5 | 192.7 |
| giun | 775 | 229.0 | 198.9 |
| ns | 1 479 | 144.1 | 122.7 |

Table 2: German-English monolingual corpora statistics: the number of tokens is given in millions [M], *ppl* is the perplexity of the corresponding corpus.

## 5 Results

To perform adaptation, an adaptation set representing the in-domain needs to be specified to be plugged in eq. (1) as IN. The choice of the adaptation corpus is crucial for the successful application of the cross-entropy based scoring, as the closer the corpus is to our test domain, the better adaptation we get. For the WMT task, the choice of the adaptation corpus is not an easy task. The genre of the test sets is newswire, while the bilingual training data is composed of news-commentary, parliamentary records (europarl) and common-crawl noisy data. On the other hand, the monolingual data includes large amounts of in-domain newswire data (news-shuffle).

For LM training, the task of adaptation might be unprofitable in terms of performance, as the

[2] www.hltpr.rwth-aachen.de/jane

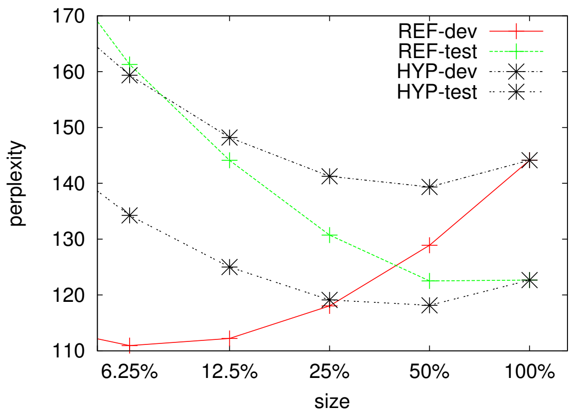

Figure 1: Size (fraction of *news-shuffle* data) against the resulting LM perplexity on *dev* and *test*, using different filtering sets.

majority of the training is in-domain. Still, one might hope that by using adaptation, a more compact and comparable LM can be generated. Another point is that LM training is less demanding than TM training, and a comparison of the results of LM and TM adaptation might prove fruitful and convey additional information.

Next, we start with LM adaptation experiments where we mainly compare different adaptation sets for filtering over the final translation quality. A comparison to the full (unfiltered LM) is also produced. For TM adaptation, we repeat the adaptation sets choice experiment and analyze the difference between the sets.

## 5.1 LM Adaptation

To evaluate our methods experimentally, we use the German-English translation task to compare different adaptation sets for filtering and then analyze the full versus the filtered LM SMT system results. We recall that *newstest09* is used as a development set and *newstest13* as a test set in the LM experiments.

The different adaptation sets for filtering that we explore are: *(i)* unsupervised: an automatic translation of the test sets (*newstest08...newstest13*), where the baseline system (without adaptation) is used to generate the hypotheses which then define the adaptation corpus for filtering (*HYP*), *(ii)* supervised: the references of the test sets *newstest08...newstest12* concatenated, *newstest13* is kept as a blind set, which will also help us determine if overfitting occurs (*REF*), and *(iii)* pseudo supervised: a pseudo in-domain corpus, *news-commentary*, where the domain is similar to the test set domain, but the style might differ (*NC*). Next, we filter the *news-shuffle* (*ns*) and *giga-fren+undoc* (*giun*) according to the three suggested adaptations sets, where we plug each adaptation set in eq. (1) as IN and compare their performance.

| **Corpus** | **Adapt set** | **Optimal size** | **ppl** | |
|---|---|---|---|---|
| | | | **dev** | **test** |
| ns | none | 100% | 144 | 123 |
| | NC | 100% | 144 | 123 |
| | REF | 6.25% | 111 | 161 |
| | HYP | 50% | 139 | 118 |
| giun | none | 100% | 229 | 199 |
| | NC | 50% | 215 | 185 |
| | REF | 6.25% | 161 | 171 |
| | HYP | 12.5% | 187 | 159 |

Table 3: Optimal size portion and resulting perplexities, across adaptation sets (**NC**, **REF** and **HYP**) and monolingual LM training corpora.

### 5.1.1 Perplexity Results

In Figure 1, we draw the size portion versus the dev and test perplexities for the *REF* and *HYP* adaptation sets over the *news-shuffle* corpus. *REF* performs best for filtering the dev set, where an optimum is achieved when using only 6.25% of the news-shuffle data, with a perplexity of 111 in comparison to 144 perplexity of the full LM. Measuring perplexities over newstest08-12, *REF* based filtering achieves 109 while the full LM achieves 140. The good performance on the seen sets comes with the cost of severe overfitting, where the test set perplexity using 6.25% of the data is 161, much higher than 123 generated by the full LM. On the other hand, *HYP* achieves an optimum for both sets when using 50% of the data. A summary of the best results across monolingual corpora and adaptation sets is given in Table 3. Filtering the *giun* monolingual corpus shows similar results to *ns* filtering, where overfitting occurs on the blind test set when using *REF* as the target domain. *HYP*-based adaptation achieves the best LM perplexity on the blind test set. *NC*-based adaptation retains the biggest amount of data, 50% for the *giun* corpus and 100% (no filtering) for the *ns* corpus. *REF*-based adaptation shows overfitting on the seen dev set, and the worst results on the blind test set when filtering the *ns* corpus.

| LM data | Adapt. set | ppl | newstest10 | | newstest11 | | newstest12 | | newstest13 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER |
| bi.en+giun | none | 162 | 23.2 | 59.6 | 21.2 | 61.0 | 21.8 | 60.9 | 24.6 | 57.2 |
| | NC | 160 | 23.2 | 59.3 | 21.5 | 61.0 | 21.9 | 60.7 | 24.6 | 57.0 |
| | REF | 158 | 23.7 | 59.2 | 21.9 | 60.5 | 22.2 | 60.5 | 24.5 | 57.3 |
| | HYP | 151 | 23.6 | 59.2 | 21.5 | 60.9 | 22.2 | 60.4 | 25.1 | 56.7 |
| +ns | none | 111 | 24.5 | 59.1 | 22.1 | 61.3 | 23.3 | 60.1 | 25.9 | 56.7 |
| | NC | 111 | 24.4 | 58.7 | 22.1 | 60.5 | 23.4 | 59.7 | 25.5 | 56.6 |
| | REF | 143 | 25.7 | 57.8 | 23.0 | 59.9 | 24.2 | 59.4 | 24.1 | 57.8 |
| | HYP | 109 | 25.0 | 58.2 | 22.1 | 60.6 | 23.5 | 59.6 | 25.9 | 56.3 |

Table 4: German-English LM filtering results using different adaptation sets. The LM perplexity over the blind test set *nestest13*, as well as BLEU and TER percentages are presented.

### 5.1.2 Translation Results

Next, we measure whether the improvements of the single adapted corpora carry over to the mixture LM both in perplexity and translation quality. The mixture LM is created by linear interpolation (of *bi.en*, *giun* and *ns*) with perplexity minimization on the dev set using the SRILM toolkit[3]. We carry out two experiments, in the first we interpolate the English side of the bilingual data with a *giun* LM, then we add the *ns* LM. This way we measure whether the effects of adaptation carry over to a stronger baseline.

The SMT systems built using the full and filtered LMs are compared in Table 4. The table includes the data used for LM training, the adaptation set used to filter the data, the perplexity of the resulting LM on the test set (*newstest13*) and the resulting SMT system quality over *newstest10...newstest13*.

Starting with the first block of experiments using LM data composed from the English side of the bilingual corpora and the *giun* corpus (bi.en+giun), the unfiltered LM performs worse, both in terms of perplexity and translation quality. The *NC* based adaptation improves the results slightly, with gains upto +0.3% BLEU on *newstest11* and -0.3% TER on *newstest10*. The overfitting behavior of *REF* adapted LMs carries over to the mixture LM, mainly on the translation quality. The *REF* adapted LM system translation results are better on the test sets used to perform the adaptation, but worse on the blind test set (*newstest13*). The *HYP* system performs best in terms of perplexity. *REF* is better than *HYP* over the non-blind test sets, but *HYP* outperforms *REF* on *newstest13* with an improvement of +0.6% BLEU and -0.6% TER.

The second block of experiments where news-shuffle (*ns*) is added to the mixture shows even stronger overfitting for *REF*. The *REF* based adaptation is performing worse in terms of perplexity, 143 in comparison to 111 for the full LM. On the blind set *newstest13*, *REF* is hindering the results with a loss of -1.8% BLEU in comparison to the full system, and a loss of -0.4% BLEU in comparison to the corresponding system without *ns*. On the non-blind sets, *REF* is performing best, showing typical overfitting. Comparing the full LM system to the *HYP* adapted LM, big improvements are mainly observed on TER, with significance at the 95% level for *newstest10*.

We conclude that using the references as adaptation set causes overfitting, using a pseudo in-domain set as the news-commentary does not improve the results, and the best choice is using the automatic translations (*HYP*).

As already mentioned in Section 2, we experimented with adding the automatic translations of the test sets (*HYP*) to the LM. Doing so resulted in 8 points perplexity reduction, but no impact on the MT quality was observed. Therefore, we deem these perplexity improvements by adding *HYP* as artificial.

## 5.2 TM Adaptation

In the LM adaptation experiments, we found that using the test sets automatic translation as the adaptation set (*HYP* system) for filtering performed best, in terms of LM quality (perplexity) and translation quality, when compared to the other suggested adaptation sets, especially on the blind test set.

[3] http://www.speech.sri.com/projects/srilm/

| LM | TM | newstest10 | | newstest11 | | newstest12 | | newstest13 | |
|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER |
| full | full | 24.5 | 59.1 | 22.1 | 61.3 | 23.3 | 60.1 | 25.9 | 56.7 |
| HYP | full | 25.0 | 58.2‡ | 22.1 | 60.6 | 23.5 | 59.6 | 25.9 | 56.3 |
| | TM Filtering | | | | | | | | |
| | REF-25% | 25.1 | 57.9‡ | 22.4 | 60.2‡ | 24.0‡ | 59.1‡ | 25.5 | 56.7 |
| | HYP-50% | 25.2 | 58.0‡ | 22.2 | 60.5† | 23.8† | 59.4‡ | 26.0 | 56.4 |
| | TM Weighting | | | | | | | | |
| | ppl.NC | 25.0 | 58.1‡ | 22.5 | 60.2‡ | 23.6 | 59.5† | 26.1 | 56.2 |
| | ppl.TST | 24.8 | 58.8 | 22.3 | 60.7 | 23.6 | 59.7 | 26.0 | 56.3 |
| | ppl.REF | 24.8 | 58.2‡ | 22.2 | 60.3† | 23.7 | 59.5† | 25.5 | 56.4 |
| | ppl.HYP | 25.4‡ | 57.8‡ | 22.5 | 60.1‡ | 23.9‡ | 59.3‡ | 26.4† | 55.9‡ |

Table 5: German-English TM filtering and weighting results using different adaptation sets. The results are given in BLEU and TER percentages. Significance is measured over the full system (first row).

For TM adaptation, we experiment with filtering and weighting based adaptation. By using weighting, we expect further improvements over the baseline and better differentiation between the competing adaptation sets.

To perform filtering, we concatenate all the bilingual corpora in Table 1 and sort them according to the combined LM+M1 cross-entropy score. We then extract the top 50%,25%,... bilingual sentence from the sorted corpus, generate the phrase table for each setup and reoptimize the system using MERT on the development set.

Weighted phrase extraction is based on the same LM+M1 combined cross entropy score as filtering, but instead of discarding whole sentences we weight them according to their relevance to the adaptation set being used.

In this section, we compare the three adaptation sets suggested for LM filtering for the TM component. In addition, one might argue that for the bilingual case, the source side of the test set might be sufficient to perform adaptation, or even it might perform better for TM adaptation as the automatically generated translation might not be as reliable. We perform an experiment using the source side of the test sets as an adaptation set to score the source side of the bilingual corpora (denoted *TST* in the experiments). To summarize, we collect 4 corpora as adaptation sets to be used for adapting the TM: *(i) NC*, *HYP*, and *REF* as defined for LM but using both source and target (automatically generated for *HYP*) sides, and *(ii) TST* using only the source side of the test sets.

The results comparing the 4 suggested adaptation sets for filtering and weighting are given in Table 5. In this table, we use *newstest10* as before for MERT optimization and display results for *newstest10...newstest13*. Note that for TM filtering and weighting we use the *HYP* adapted LM as it achieves the best results in the previous section.

For filtering, the *NC* and *TST* adaptation sets could not improve the dev results over the full system therefore they are omitted. *REF* based adaptation achieves the best dev results when using 25% of the bilingual data while *HYP* based adaptation uses 50% of the data. For TM filtering, only slight overfitting is observed, where the *REF* system is slightly better than *HYP* on the non blind sets and is worse on the blind test set. We hypothesize that no severe overfitting is observed for TM filtering as we use a strong LM adapted with the *HYP* set, therefore degradation is lessened.

Next, we focus on weighted phrase extraction for adaptation using the various adaptation sets. Comparing filtering to weighting, weighting improves for the *ppl.HYP* based adaptation but a slight loss is observed for the *ppl.REF* system except on the blind test set. We conclude that due to the usage of more data in the weighting scenario, overfitting is lessened. Using the source side of the test sets for weighting (*ppl.TST*) achieves good results, with improvements over the *ppl.REF* system on *newstest13*.

The *ppl.HYP* system achieves the best results among the weighted systems. Comparing the full unadapted system with the LM+TM adapted *ppl.HYP* system, we achieve significant BLEU improvements on most sets, TER improvements are significant in all cases with 95% significance level. The highest gains are on the development set with

+0.9% BLEU and -1.3% TER improvements, on the test sets, *newstest12* improves with +0.6% BLEU and -0.8% TER and *newstest13* improves with +0.5% BLEU and -0.8% TER. The *ppl.HYP* system is comparable to the best single system of WMT 2013 [4] (26.4% BLEU vs 26.8% BLEU for Edinburgh submission, RWTH submission is a system combination). Note that we are not using the LDC GigaWord corpus.

We conclude that using in-domain automatic translations (*HYP*) for TM weighting performs best, better than using source side only in-domain (*TST*) and better than using the references (*REF*) especially on the blind test set. TM adaptation shows further improvements on top of LM adaptation and achieves significant gains.

## 6 Conclusion

In this work, we tackle the problem of adaptation without labeled bilingual in-domain training data. The only information about the test domain is encapsulated in the test sets themselves. We experiment with unsupervised adaptation for SMT, using automatic translations of the test sets, focusing on adaptation for the LM and the TM components. We use cross-entropy based scoring for the task of adaptation, as this method proved successful in previous work. We utilize filtering for LM adaptation, while we compare filtering and weighting for TM adaptation.

For LM adaptation, the setup we devise already contains a majority of in-domain data, still we could report improvements over the unadapted baseline. We compose three different adaptation sets for filtering using automatic translation of the test data (*HYP*), a pseudo in-domain set (*NC*) and the references (*REF*) of the test sets (keeping one blind test set). The *NC* based filtering is not able to perform good selection, for *news-shuffle* the whole corpus is retained and for *giun* 50% of the data is retained. The perplexity results and the translation quality are virtually unchanged in comparison to the full system. Using *REF* as the target set causes overfitting, where the results are better on the seen test sets but worse on the blind test set. The best performing target set in our experiments is the unsupervised *HYP* adaptation set, achieving the best perplexity as well as the best translation quality on the blind test set. Therefore, we conclude that for developing a successful SMT system that can generalize to new data the *HYP* based adaptation is preferred.

Next, we perform TM adaptation, where we repeat the comparison between the different adaptation sets for filtering as well as weighting. We also compare to adaptation based only on the source side of the test sets (*TST*). The LM adaptation results hold for TM adaptation, where using the automatic translations method shows the best results for the blind test set. Our experiments show that using the source side only of the test set for adaptation performs worse than the unsupervised method, reminiscent to results reported in previous work comparing supervised source side against bilingual filtering (Axelrod et al., 2011). For filtering, the *REF* system suffers from overfitting, while when using weighting for adaptation, overfitting is lessened. Comparing the unadapted baseline to the adapted LM and TM system using the *HYP* set, improvements of +1.0% BLEU and -1.3% TER are reported on the development set while +0.5% BLEU and -0.8% TER improvements are reported on the blind test set.

[4] http://matrix.statmt.org/matrix/systems_list/1712

## Acknowledgments

This material is based upon work supported by the DARPA BOLT project under Contract No. HR0011-12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

## References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

M. Bacchiani and B. Roark. 2003. Unsupervised language model adaptation. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–224 – I–227 vol.1, april.

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93 – 108. Adaptation Methods for Speech Recognition.

M Federico M Cettolo, L Bentivogli, M Paul, and S Stüker. 2012. Overview of the iwslt 2012 evaluation campaign. In *International Workshop on*

*Spoken Language Translation*, pages 12–33, Hong Kong, December.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 847–855, Honolulu, Hawaii, USA, October.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. 2013. Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR Conference*, volume 10, pages 260–286, Tokyo, Japan, June.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th EAMT conference on "Practical applications of machine translation"*, pages 133–1142, May.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 317–321, Montréal, Canada, June. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain, July.

Saab Mansour and Hermann Ney. 2012. A simple and effective weighted phrase extraction for machine translation adaptation. In *International Workshop on Spoken Language Translation*, pages 193–200, Hong Kong, December.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July. Association for Computational Linguistics.

Franz J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Raphael Rubino, Antonio Toral, Santiago Cortés Vaíllo, Jun Xie, Xiaofeng Wu, Stephen Doherty, and Qun Liu. 2013. The CNGL-DCU-Prompsit translation systems for WMT13. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 213–218, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Joern Wuebker, Matthias Huck, Stephan Peitz, Malte Nuhn, Markus Freitag, Jan-Thorsten Peter, Saab Mansour, and Hermann Ney. 2012. Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, Mumbai, India, December.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.