# Phrase Training Based 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

We present a novel approach for translation model (TM) adaptation using phrase training. The proposed adaptation procedure is initialized with a standard general-domain TM, which is then used to perform phrase training on a smaller in-domain set. This way, we bias the probabilities of the general TM towards the in-domain distribution. Experimental results on two different lectures translation tasks show significant improvements of the adapted systems over the general ones. Additionally, we compare our results to mixture modeling, where we report gains when using the suggested phrase training adaptation method.

## 1 Introduction

The task of domain-adaptation attempts to exploit data mainly drawn from one domain (e.g. news, parliamentary discussion) to maximize the performance on the test domain (e.g. lectures, web forums). In this work, we focus on translation model (TM) adaptation. A prominent approach in recent work is weighting at different levels of granularity. Foster and Kuhn (2007) perform weighting at the corpus level, where different corpora receive different weights and are then combined using mixture modeling. A finer grained weighting is that of Matsoukas et al. (2009), who weight each sentence in the bitexts using features of meta-information and optimize a mapping from the feature vectors to weights using a translation quality measure.

In this work, we propose to perform TM adaptation using phrase training. We start from a general-domain phrase table and adapt the probabilities by training on an in-domain data. Thus, we achieve direct phrase probabilities adaptation as opposed to weighting. Foster et al. (2010) perform weighting at the phrase level, assigning each phrase pair a weight according to its relevance to the test domain. They compare phrase weighting to a "flat" model, where the weight directly approximates the phrase probability. In their experiments, the weighting method performs better than the flat model, therefore, they conclude that retaining the original relative frequency probabilities of the TM is important for good performance. The "flat" model of Foster et al. (2010) is similar to our work. We differ in the following points: *(i)* we use the same procedure to perform the phrase training based adaptation and the search thus avoiding inconsistencies between the two; *(ii)* we do not directly interpolate the original statistics with the new ones, but use a training procedure to manipulate the original statistics. We perform experiments on the publicly available IWSLT TED task, on both Arabic-to-English and German-to-English lectures translation tracks. We compare our suggested phrase training adaptation method to a variety of baselines and show its effectiveness. Finally, we experiment with mixture modeling based adaptation. We compare mixture modeling to our adaptation method, and apply our method within a mixture modeling framework.

In Section 2, we present the phrase training method and explain how it is utilized for adaptation. Experimental setup including corpora statistics and the SMT system are described in Section 3. Section 4 summarizes the phrase training adaptation results ending with a comparison to mixture modeling.



*Proceedings of NAACL-HLT 2013*, pages 649–654,
Atlanta, Georgia, 9–14 June 2013. 

## 2 Phrase Training

The standard phrase extraction procedure in SMT consists of two phases: *(i)* word-alignment training (e.g., IBM alignment models), *(ii)* heuristic phrase extraction and relative frequency based phrase translation probability estimation. In this work, we utilize phrase training for the task of adaptation. We use the forced alignment (FA) method (Wuebker et al., 2010) to perform the phrase alignment training and probability estimation. We perform phrase training by running a normal SMT decoder on the training data and constrain the translation to the given target instance. Using n-best possible phrase segmentation for each training instance, the phrase probabilities are re-estimated over the output. Leaving-one-out is used during the forced alignment procedure phase to avoid over-fitting (Wuebker et al., 2010).

In the standard phrase training procedure, we are given a training set $y$, from which an initial heuristics-based phrase table $p_y^0$ is generated. FA training is then done over the training set $y$ using the phrases and probabilities in $p_y^0$ (possibly updated by the leaving-one-out method). Finally, re-estimation of the phrase probabilities is done over the decoder output, generating the FA phrase table $p^1$. We explain next how to utilize FA training for adaptation.

### 2.1 Adaptation

In this work, we utilize phrase training for the task of adaptation. The main idea is to generate the initial phrase table required for FA using a general-domain training data $y'$, thus resulting in $p_{y'}^0$, and perform the FA training over $y_{IN}$, the in-domain training data (instead of $y'$ in the standard procedure). This way, we bias the probabilities of $p_{y'}^0$ towards the in-domain distribution. We denote this new procedure by Y'-FA-IN. This differs from the standard IN-FA-IN by that we have more phrase pairs to use for FA. Thus, we obtain phrase pairs relevant to IN in addition to "general" phrase pairs which were not extracted from IN, perhaps due to faulty word alignments. The probabilities of the general phrase table will be tailored towards IN. In practice, we usually have in-domain IN and other-domain OD data. We denote by ALL the concatenation of IN and OD. To adapt the ALL phrase table, we perform the FA procedure ALL-FA-IN. We also utilize leaving-one-out to avoid over-fitting.

Another procedure we experimented with is adapting the OD phrase table using FA over IN, without leaving-one-out. We denote it by OD-FA$_0$-IN. In this FA scenario, we do not use leaving-one-out as IN is not contained in OD, therefore, over-fitting will not occur. By this procedure, we train phrases from OD that are relevant for both OD and IN, while the probabilities will be tailored to IN. In this case, we do not expect improvements over the IN based phrase table, but, improvements over OD and reduction in the phrase table size.

We compare our suggested FA based adaptation to the standard FA procedure.

## 3 Experimental Setup

### 3.1 Training Corpora

To evaluate the introduced methods experimentally, we use the IWSLT 2011 TED Arabic-to-English and German-to-English translation tasks. The IWSLT 2011 evaluation campaign focuses on the translation of TED talks, a collection of lectures on a variety of topics ranging from science to culture. For Arabic-to-English, the bilingual data consists of roughly 100K sentences of in-domain TED talks data and 8M sentences of "other"-domain United Nations (UN) data. For the German-to-English task, the data consists of 130K TED sentences and 2.1M sentences of "other"-domain data assembled from the news-commentary and the europarl corpora. For language model training purposes, we use an additional 1.4 billion words (supplied as part of the campaign monolingual training data).

The bilingual training and test data for the Arabic-to-English and German-to-English tasks are summarized in Table 1[1]. The English data was tokenized and lowercased while the Arabic data was tokenized and segmented using MADA v3.1 (Roth et al., 2008) with the ATB scheme. The German source is decompounded (Koehn and Knight, 2003) and part-of-speech-based long-range verb reordering rules (Popović and Ney, 2006) are applied.

From Table 1, we note that using the general data considerably reduces the number of out-of-

[1]For a list of the IWSLT TED 2011 training corpora, see http://www.iwslt2011.org/doku.php?id=06_evaluation

| Set | Sen | Tok | OOV/IN | OOV/ALL |
|---|---|---|---|---|
| German-to-English | | | | |
| IN | 130K | 2.5M | | |
| OD | 2.1M | 55M | | |
| dev | 883 | 20K | 398 (2.0%) | 215 (1.1%) |
| test | 1565 | 32K | 483 (1.5%) | 227 (0.7%) |
| eval | 1436 | 27K | 490 (1.8%) | 271 (1.0%) |
| Arabic-to-English | | | | |
| IN | 90K | 1.6M | | |
| OD | 7.9M | 228M | | |
| dev | 934 | 19K | 408 (2.2%) | 184 (1.0%) |
| test | 1664 | 31K | 495 (1.6%) | 228 (0.8%) |
| eval | 1450 | 27K | 513 (1.9%) | 163 (0.6%) |

Table 1: IWSLT 2011 TED bilingual corpora statistics: the number of tokens is given for the source side. OOV/X denotes the number of OOV words in relation to corpus X (the percentage is given in parentheses). *IN* is the TED in-domain data, *OD* denotes other-domain data, *ALL* denotes the concatenation of *IN* and *OD*.

vocabulary (OOV) words. This comes with the price of increasing the size of the training data by a factor of more than 20. A simple concatenation of the corpora might mask the phrase probabilities obtained from the in-domain corpus, causing a deterioration in performance. One way to avoid this contamination is by filtering the general corpus, but this discards phrase translations completely from the phrase model. A more principled way is by adapting the phrase probabilities of the full system to the domain being tackled. We perform this by phrase training the full phrase table over the in-domain training set.

### 3.2 Translation System

The baseline system is built using the open-source SMT toolkit Jane 2.0, which provides a state-of-the-art phrase-based SMT system (Wuebker et al., 2012a). In addition to the phrase based decoder, Jane 2.0 implements the forced alignment procedure used in this work for the purpose of adaptation. 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 and an $n$-gram target language model. The SMT systems are tuned on the *dev* (dev2010) development set with minimum error rate training (Och, 2003) using BLEU (Papineni et al., 2002) accuracy measure as the optimization criterion. We test the performance of our system on the *test* (tst2010) and *eval* (tst2011) 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. The Arabic-English results are case sensitive while the German-English results are case insensitive.

## 4 Results

For TM training, we define three different sets: in-domain (IN) which is the TED corpus, other-domain (OD) which consists of the UN corpus for Arabic-English and a concatenation of news-commentary and europarl for German-English, and ALL which consists of the concatenation of IN and OD. We experiment with the following extraction methods:

- Heuristics: standard phrase extraction using word-alignment training and heuristic phrase extraction over the word alignment. The extraction is performed for the three different training data, IN, OD and ALL.
- FA standard: standard FA phrase training where the same training set is used for initial phrase table generation as well as the FA procedure. We perform the training on the three different training sets and denote the resulting systems by IN-FA, OD-FA and ALL-FA.
- FA adaptation: FA based adaptation phrase training, where the initial table is generated from some general data and the FA training is performed on the IN data to achieve adaptation. We perform two experiments, OD-FA$_0$-IN without leaving-one-out and ALL-FA-IN with leaving-one-out.

The results of the various experiments over both Arabic-English and German-English tasks are summarized in Table 2. The usefulness of the OD data differs between the Arabic-to-English and the German-to-English translation tasks. For Arabic-to-English, the *OD* system is 2.5%-4.3% BLEU worse than the *IN* system, whereas for the German-to-English task the differences between *IN* and *OD* are smaller and range from 0.9% to 1.6% BLEU. The

| Phrase training method | System | Rules number | dev | | test | | eval | |
|---|---|---|---|---|---|---|---|---|
| | | | BLEU | TER | BLEU | TER | BLEU | TER |
| **Arabic-to-English** | | | | | | | | |
| Heuristics | IN | 1.1M | 27.2 | 54.1 | 25.3 | 57.1 | 24.3 | 59.9 |
| | OD | 36.3M | 24.7 | 57.7 | 21.2 | 62.6 | 21.0 | 64.7 |
| | ALL | 36.9M | 27.1 | 54.8 | 24.4 | 58.6 | 23.8 | 61.1 |
| FA standard | IN-FA | 1.0M | 27.0 | 54.4 | 25.0 | 57.5 | 23.8 | 60.3 |
| | OD-FA | 1.8M | 24.5 | 57.7 | 21.0 | 62.4 | 21.2 | 64.3 |
| | ALL-FA | 2.0M | 27.2 | 54.2 | 24.5 | 58.1 | 23.8 | 60.6 |
| FA adaptation | OD-$FA_0$-IN | 0.3M | 25.8 | 55.8 | 23.6 | 59.4 | 22.7 | 61.7 |
| | ALL-FA-IN | 0.5M | 27.7 | 53.7 | 25.3 | 56.9 | 24.7 | 59.3 |
| **German-to-English** | | | | | | | | |
| Heuristics | IN | 1.3M | 31.0 | 48.9 | 29.3 | 51.0 | 32.7 | 46.8 |
| | OD | 7.3M | 29.8 | 49.2 | 27.7 | 51.5 | 31.8 | 47.5 |
| | ALL | 7.8M | 31.2 | 48.3 | 29.5 | 50.5 | 33.6 | 46.1 |
| FA standard | IN-FA | 0.5M | 31.6 | 48.2 | 29.7 | 50.5 | 33.3 | 46.4 |
| | OD-FA | 3.0M | 29.1 | 51.0 | 27.6 | 53.0 | 30.7 | 49.6 |
| | ALL-FA | 3.2M | 31.4 | 48.3 | 29.4 | 50.8 | 33.6 | 46.2 |
| FA adaptation | OD-$FA_0$-IN | 0.9M | 31.2 | 48.7 | 29.1 | 50.9 | 32.7 | 46.9 |
| | ALL-FA-IN | 0.9M | 31.8 | 47.4 | 29.7 | 49.7 | 33.6 | 45.5 |

Table 2: TED 2011 translation results. BLEU and TER are given in percentages. *IN* denotes the TED lectures in-domain corpus, *OD* denotes the other-domain corpus, *ALL* is the concatenation of *IN* and *OD*. $FA_0$ denotes forced alignment training without leaving-one-out (otherwise, leaving-one-out is used).

inferior performance of the *OD* system can be related to noisy data or bigger discrepancy between the OD data domain distribution and the IN distribution. The *ALL* system performs according to the usefulness of the OD training set, where for Arabic-to-English we observe deterioration in performance for all test sets and up-to -0.9% BLEU on the *test* set. On the other hand, for German-to-English, the *ALL* system is improving over *IN* where the biggest improvement is observed on the *eval* set with +0.9% BLEU improvement.

The standard FA procedure achieves mixed results, where *IN-FA* deteriorates the results over the *IN* counterpart for Arabic-English, while improving for German-English. *ALL-FA* performs comparably to the *ALL* system on both tasks, while reducing the phrase table size considerably. The *OD-FA* system deteriorates the results in comparison to the *OD* system in most cases, which is expected as training over the OD set fits the phrase model on the OD domain, making it perform worse on IN. (Wuebker et al., 2012b) also report mixed results with FA training.

The FA adaptation results are summarized in the last block of the experiments. The *OD-$FA_0$-IN* improves over the *OD* system, which means that the training procedure was able to modify the OD probabilities to perform well on the IN data. On the German-to-English task, the *OD-$FA_0$-IN* performs comparably to the IN system, whereas for Arabic-to-English *OD-$FA_0$-IN* was able to close around half of the gap between OD and IN.

The FA adapted ALL system (*ALL-FA-IN*) performs best in our experiments, improving on both BLEU and TER measures. In comparison to the best heuristics system (*IN* for Arabic-English and *ALL* for German-English), +0.4% BLEU and -0.6% TER improvements are observed on the *eval* set for Arabic-English. For German-English, the biggest improvements are observed on TER with -0.8% on *test* and -0.5% on *eval*. The results suggest that *ALL-FA-IN* is able to learn more useful phrases than the *IN* system and adjust the *ALL* phrase probabilities towards the in-domain distribution.

| System | dev | | test | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| **Arabic-to-English** | | | | |
| $\text{Heuristics}_{\text{best}}$ | 27.2 | 54.1 | 25.3 | 57.1 |
| IN,OD | 28.2 | 53.1 | 25.5 | 56.8 |
| IN,OD-$\text{FA}_0$-IN | 28.4 | 52.9 | 25.7 | 56.5 |
| **German-to-English** | | | | |
| $\text{Heuristics}_{\text{best}}$ | 31.2 | 48.3 | 29.5 | 50.5 |
| IN,OD | 31.6 | 48.2 | 29.9 | 50.5 |
| IN,OD-$\text{FA}_0$-IN | 31.8 | 47.8 | 30.0 | 50.0 |

Table 3: TED 2011 mixture modeling results. $\text{Heuristics}_{\text{best}}$ is the best heuristics based system, *IN* for Arabic-English and *ALL* for German-English. *X,Y* denotes linear interpolation between X and Y phrase tables.

### 4.1 Mixture Modeling

In this section, we compare our method to mixture modeling based adaptation, in addition to applying mixture modeling on top of our method. We focus on linear interpolation (Foster and Kuhn, 2007) of the in-domain (IN) and other-domain phrase tables, where we vary the latter between the heuristically extracted phrase table (*OD*) and the FA adapted one (*OD-*$\text{FA}_0$*-IN*). The interpolation weight is uniform for the interpolated phrase tables ($0.5$). The results of mixture modeling are summarized in Table 3. In this table, we include the best heuristics based system ($\text{Heuristics}_{\text{best}}$) from Table 2 as a reference system. The results on the *eval* set are omitted as they show similar tendencies to the *test* set results.

Linear interpolation of *IN* and *OD* (*IN,OD*) is performing well in our experiments, with big improvements over the *dev* set, +1.0% BLEU for Arabic-to-English and +0.4% BLEU for German-to-English. On the *test* set, we observe smaller improvements. Interpolating *IN* with the phrase training adapted system *OD-*$\text{FA}_0$*-IN* (*IN,OD-*$\text{FA}_0$*-IN*) achieves additional gains over the *IN,OD* system, the biggest are observed on TER for German-to-English, with -0.4% and -0.5% improvements on the *dev* and *test* sets correspondingly.

Comparing heuristics based interpolation (*IN,OD*) to our best phrase training adapted system (*ALL-FA-IN*) shows mixed results. For Arabic-to-English, the systems are comparable, while for the German-to-English *test* set, *IN,OD* is +0.2% BLEU better and +0.8% TER worse than *ALL-FA-IN*. We hypothesize that for Arabic-to-English interpolation is important due to the larger size of the OD data, where it could reduce the masking of the IN training data by the much larger OD data. Nevertheless, as mentioned previously, using phrase training adapted phrase table in a mixture setup consistently improves over using heuristically extracted tables.

## 5 Conclusions

In this work, we propose a phrase training procedure for adaptation. The phrase training is implemented using the FA method. First, we extract a standard phrase table using the whole available training data. Using this table, we initialize the FA procedure and perform training on the in-domain set.

Experiments are done on the Arabic-to-English and German-to-English TED lectures translation tasks. We show that the suggested procedure is improving over unadapted baselines. On the Arabic-to-English task, the FA adapted system is +0.9% BLEU better than the full unadapted counterpart on both test sets. Unlike the Arabic-to-English setup, the German-to-English OD data is helpful and produces a strong unadapted baseline in concatenation with IN. In this case, the FA adapted system achieves BLEU improvements mainly on the development set with +0.6% BLEU, on the *test* and *eval* sets, improvements of -0.8% and -0.6% TER are observed correspondingly. As a side effect of the FA training process, the size of the adapted phrase table is less than 10% of the size of the full table.

Finally, we experimented with mixture modeling where improvements are observed over the unadapted baselines. The results show that using our phrase training adapted OD table yields better performance than using the heuristically extracted OD in a mixture framework.

## 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


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.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA, October. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 347–354, Budapest, Hungary, April.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Singapore, August. 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.

M. Popović and H. Ney. 2006. POS-based Word Reorderings for Statistical Machine Translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283.

Ryan Roth, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin. 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In *Proceedings of ACL-08: HLT, Short Papers*, pages 117–120, Columbus, Ohio, June. 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.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden, July.

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

Joern Wuebker, Mei-Yuh Hwang, and Chris Quirk. 2012b. Leave-one-out phrase model training for large-scale deployment. In *NAACL 2012 Seventh Workshop on Statistical Machine Translation*, pages 460–467, Montreal, Canada, June. Association for Computational Linguistics.