Translation Model Based Weighting for Phrase Extraction

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

Domain adaptation for statistical machine translation is the task of altering general models to improve performance on the test domain. In this work, we suggest several novel weighting schemes based on translation models for adapted phrase extraction. To calculate the weights, we first phrase align the general bilingual training data, then, using domain specific translation models, the aligned data is scored and weights are defined over these scores. Experiments are performed on two translation tasks, German-to-English and Arabic-to-English translation with lectures as the target domain. Different weighting schemes based on translation models are compared, and significant improvements over automatic translation quality are reported. In addition, we compare our work to previous methods for adaptation and show significant gains.

1 Introduction

In recent years, large amounts of monolingual and bilingual training corpora were collected for statistical machine translation (SMT). Early years focused on structured data translation such as newswire and parliamentary discussions. Nowadays, new domains of translation are being explored, such as talk translation in the IWSLT TED evaluation (Cettolo et al., 2012) and patents translation at the NTCIR PatentMT task (Goto et al., 2013).

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

The main component of an SMT system is the phrase table, providing the building blocks (i.e. phrase translation pairs) and corresponding translation model scores (e.g., phrase models, word lexical smoothing, etc.) to search for the best translation. In this work, we experiment with phrase model adaptation through training data weighting, where one assigns higher weights to relevant domain training instances, thus causing an increase of the corresponding probabilities. As a result, translation pairs which can be obtained from relevant training instances will have a higher chance of being utilized during search.

The main contribution of this work is designing several novel schemes for scoring sentences and assigning them appropriate weights to manifest adaptation. Our method consists of two steps: first, we find phrase alignments for the bilingual training data, then, the aligned data is scored using translation models and weights are generated.

Experiments using the suggested methods and a comparison to previous work are done on two tasks: Arabic-to-English and German-to-English TED lectures translation. The results show significant improvements over the baseline, and significant improvements over previous work are reported when combining our suggested methods with previous work.

The rest of the paper is organized as follows. Related work on adaptation and weighting is detailed in Section 2. The weighted phrase extraction training and the methods for assigning weights using translation models are described in

Section 3 and Section 4 correspondingly. Experimental setup including corpora statistics and the SMT system used in this work are described in Section 5. The results of the suggested methods are summarized in Section 6 and error analysis is given in Section 7. Last, we conclude with few suggestions for future work.

2 Related Work

A broad range of methods and techniques have been suggested in the past for domain adaptation for SMT. In recent work, language model and phrase model adaptation received most of the attention. In this work, we focus on phrase model adaptation. A prominent approach in recent work for phrase model adaptation is training samples weighting at different levels of granularity. Foster and Kuhn (2007) perform phrase model adaptation using mixture modeling at the corpus level. Each corpus in their setting gets a weight using various methods including language model (LM) perplexity and information retrieval methods. Interpolation is then done linearly or log-linearly. The weights are calculated using the development set therefore expressing adaptation to the domain being translated. A finer grained weighting is that of (Matsoukas et al., 2009), who assign each sentence in the bitexts a weight using features of meta-information and optimizing a mapping from feature vectors to weights using a translation quality measure over the development set. Foster et al. (2010) perform weighting at the phrase level, using a maximum likelihood term limited to the development set as an objective function to optimize. They compare the phrase level weighting to a "flat" model, where the weight directly models 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 phrase model is important for good performance.

Data filtering for adaptation (Moore and Lewis, 2010; Axelrod et al., 2011) can be seen as a special case of the sample weighting method where a weight of 0 is assigned to discard unwanted samples. These methods rely on an LM based score to perform the selection, though the filtered data will affect the training of other models such as the phrase model and other translation models. LM based scoring might be more appropriate for LM adaptation but not as much for phrase model adaptation as it does not capture bilingual dependencies. We score training data instances using translation models and thus model connections between source and target sentences.

In this work, we compare several scoring schemes at the sentence level for weighted phrase extraction. Additionally, we experiment with new scoring methods based on translation models used during the decoding process. In weighting, all the phrase pairs are retained, and only their probability is altered. This allows the decoder to make the decision whether to use a phrase pair or not, a more methodological way than removing phrase pairs completely when filtering.

3 Weighted Phrase Extraction

The classical phrase model is estimated using relative frequency:

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r c_r(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} \sum_r c_r(\tilde{f}', \tilde{e})} \qquad (1)$$

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 in sentence pair (f_r, e_r). One method to introduce weights to eq. (1) is by weighting each sentence pair by a weight w_r. Eq. (1) will now have the extended form:

$$p(\tilde{f}|\tilde{e}) = \frac{\sum_r w_r \cdot c_r(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} \sum_r w_r \cdot c_r(\tilde{f}', \tilde{e})} \qquad (2)$$

It is easy to see that setting $\{w_r = 1\}$ will result in eq. (1) (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 perform weighting rather than filtering for adaptation as the former was shown to achieve better results (Mansour and Ney, 2012).

Next, we discuss several methods for setting the weights in a fashion which serves adaptation.

4 Weighting Schemes

Several weighting schemes can be devised to manifest adaptation. Previous work suggested perplexity based scoring to perform adaptation (e.g. (Moore and Lewis, 2010)). The basic idea is to generate a model using an in-domain training data and measure the perplexity of the in-domain model on new events to rank their relevance to the in-domain. We recall this method in Section 4.1.

In this work, we suggest to use several phrase-based translation models to perform scoring. The basic idea of adaptation using translation models is similar to the perplexity based method. We use an in-domain training data to estimate translation model scores over new events. Further details of the method are given in Section 4.2.

4.1 LM Perplexity Weighting

LM cross-entropy scoring can be used for both monolingual and bilingual data filtering (Moore and Lewis, 2010; Axelrod et al., 2011). Next, we recall the scoring methods introduced in the above previous work and utilize it for our proposed weighted phrase extraction method.

The scores for each sentence in the general-domain corpus are based on the cross-entropy difference of the in-domain (IN) and general-domain (GD) models. Denoting $H_{LM}(x)$ as the cross entropy of sentence x according to LM, then the cross entropy difference $DH_{LM}(x)$ can be written as:

$$DH_{LM}(x) = H_{LM_{IN}}(x) - H_{LM_{GD}}(x) \qquad (3)$$

The intuition behind eq. (3) 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. (3) for filtering performs better in terms of perplexity than using in-domain cross-entropy only ($H_{LM_{IN}}(x)$). For more details about the reasoning behind eq. (3) we refer the reader to (Moore and Lewis, 2010).

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

$$d_r = DH_{LM_{source}}(f_r) + DH_{LM_{target}}(e_r)$$

We utilize d_r 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 weights indicating sentences being closer to the in-domain, and use an exponent to ensure positive values. The final weight is of the form:

$$w_r = e^{-d_r} \qquad (4)$$

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

4.2 Translation Model Weighting

In state-of-the-art SMT several models are used during decoding to find the best scoring hypothesis. The models include, phrase translation probabilities, word lexical smoothing, reordering models, etc. We utilize these translation models to perform sentence weighting for adaptation. To estimate the models' scores, a phrase alignment is required. We use the forced alignment (FA) phrase training procedure (Wuebker et al., 2010) for this purpose. The general FA procedure will be presented next followed by an explanation how we estimate scores for adaptation using FA.

4.2.1 Forced Alignment 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 using the FA method for the task of adaptation. Unlike heuristic phrase extraction, the FA method performs actual phrase training. In the standard FA procedure, we are given a training set, from which an initial heuristics-based phrase table p^0 is generated. FA training is then done by running a normal SMT decoder (using p^0 phrases and models) on the training data and constrain the translation to the given target instance. Forced decoding generates n-best possible phrase alignments from which we are interested in the first-best (viterbi) one. Note that we do not use FA to generate a trained phrase table but only to get phrase alignments of the bilingual training data. We explain next how to utilize FA training for adaptation.

4.2.2 Scoring

The proposed method for calculating translation model scores using FA is depicted in Figure 1. We start by training the translation models using the standard heuristic method over the in-domain portion of the training data. We then use these in-domain translation models to perform the FA pro-

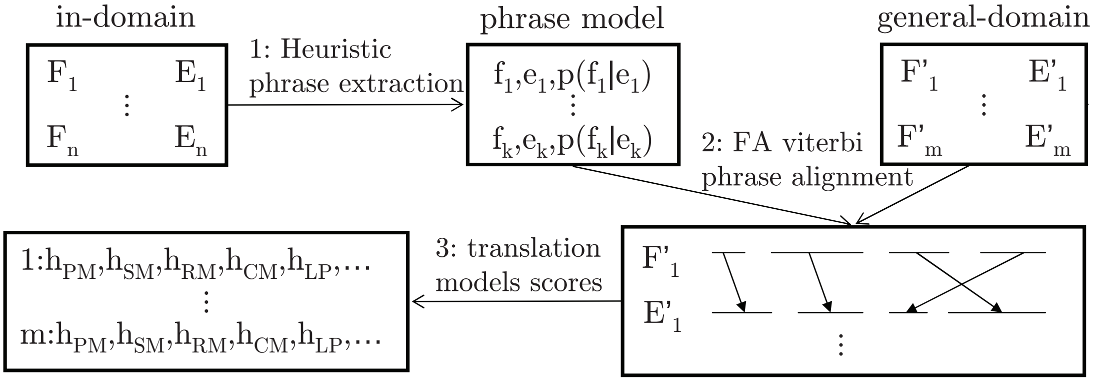

Figure 1: Translation model scores generation for general-domain sentence pairs using in-domain corpus and viterbi phrase alignments calculated by the FA procedure.

cedure over the general-domain (GD) data. The FA procedure provides n-best possible phrase alignments, but we are interested only in one alignment. Even though the IN data is small, we ensure that all GD sentences are phrase aligned using backoff phrases (Wuebker and Ney, 2013). Using the viterbi (first-best) phrase alignment and the in-domain models again, we generate the translation model scores for GD sentences. As the scores are calculated by IN models, they express the relatedness of the scored sentence to the in-domain. Note that the FA procedure for getting adaptation weights is different from the standard FA procedure. In the standard FA procedure, the same corpus is used to generate the initial heuristic phrase table as well as phrase training. The FA procedure to obtain adaptation weights uses an initial phrase table extracted from IN while the training is done over GD.

Next, we define the process for generating the scores with mathematical notation. Given a training sentence pair (f_1^J, e_1^I) from the GD corpus, we force decode $f_1^J = f_1...f_J$ into $e_1^I = e_1...e_I$ using the IN phrase table. The force decoding process generates a viterbi phrase alignment $s_1^K = s_1...s_K$, $s_k = (b_k, j_k; i_k)$ where (b_k, j_k) are the source phrase $\tilde{f}_k$ begin and end positions correspondingly, and i_k is the end position of translation target phrase $\tilde{e}_k$ (the start position of $\tilde{e}_k$ is $i_{k-1}+1$ by definition of phrase based translation). Using s_1^K we calculate the scores of 10 translation models which are grouped into 5 weighting schemes:

- PM: phrase translation models in both source-to-target (s2t) and target-to-source (t2s) directions

$$h_{PM_{s2t}}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \log p(\tilde{f}_k|\tilde{e}_k)$$

The t2s direction is defined analogously using the $p(\tilde{e}_k|\tilde{f}_k)$ probabilities.

- SM: word lexical smoothing models also in both translation directions

$$h_{SM_{s2t}}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K}\sum_{j=b_k}^{j_k} \log \sum_{i=i_{k-1}+1}^{i_k} p(f_j|e_i)$$

- RM: distance based reordering model

$$h_{RM}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} |b_k - j_{k-1} + 1|$$

- CM: phrase count models

$$h_{CM_i}(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \left[c(\tilde{f}_k, \tilde{e}_k) < i\right]$$

i is assigned the values 2,3,4 (3 count features). $c(\tilde{f}, \tilde{e})$ is the count of the bilingual phrase pair being aligned to each other (in the IN corpus).

- LP: length based word and phrase penalties

$$h_{LP_{wordPernalty}}(f_1^J, e_1^I, s_1^K) = I$$

$$h_{LP_{phrasePenalty}}(f_1^J, e_1^I, s_1^K) = K$$

We experiment with the PM scheme independently. In addition, we try using all models in a loglinear fashion for weighting (denoted by TM), and using TM and LM combined score (denoted by TM+LM). We use the decoder optimized lambdas to combine the models.

To obtain the weights for a scheme which is composed of a set of models $\{h_1^n\}$, we normalize (the sum of absolute values equals 1) the corresponding lambdas obtaining $\{\lambda_1^n\}$, and calculate:

$$w(f,e,s) = e^{-\sum_{i=1}^{n} \lambda_i \cdot h_i(f,e,s)}$$

An alternative method to perform adaptation by force aligning GD using IN would be performing phrase probability re-estimation as done in the final step of standard FA training. In this case, n-best phrase alignments are generated for each sentence in GD using the IN models and the phrase model is then reestimated using relative frequencies on the n-bests. This way we directly use the FA procedure to generate the translation models. The problem with this approach is that due to the small size of IN, some sentences in GD can not be decoded with the initial phrase table and fallback runs using backoff phrases need to be used (Wuebker and Ney, 2013). Backoff phrases of a sentence pair contain all source and target sub-strings up-to a defined maximum length. Therefore, many of these backoff phrase pairs are not a translation of each other. Using such phrases to reestimate the phrase model might generate unwanted phrase translation candidates. In the case of weighting, the backoff probabilities are used indirectly to weight the initial counts, in addition, combining with other model scores remedies the problem further.

Another way to perform adaptation using FA is by starting with a GD heuristic phrase table and utilize it to force decode IN. This way, the probabilities of the general phrase model are biased towards the in-domain distribution. This method was presented by (Mansour and Ney, 2013) and will be compared to our work.

5 Experimental Setup

5.1 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 (OD) 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).

		de	**en**	**ar**	**en**
IN	sen	130K		90K	
	tok	2.5M	3.4M	1.6M	1.7M
	voc	71K	49K	56K	34K
OD	sen	2.1M		7.9M	
	tok	55M	56M	228M	226M
	voc	191K	129K	449K	411K
dev	sen	883		934	
	tok	20K	21K	19K	20K
	oov	215 (1.1%)		184 (1.0%)	
test10	sen	1565		1664	
	tok	31K	27K	31K	32K
	oov	227 (0.7%)		228 (0.8%)	
test11	sen	1436		1450	
	tok	27K	27K	27K	27K
	oov	271 (1.0%)		163 (0.6%)	

Table 1: IWSLT 2011 TED bilingual corpora statistics: the number of sentences (sen), running words (tok) and vocabulary (voc) are given for the training data. For the test data, the number of out-of-vocabulary (oov) words relatively to using all training data (concatenating IN and OD) is given (in parentheses is the percentage).

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 is tokenized and lowercased while the Arabic data was tokenized and segmented using MADA v3.1 (Roth et al., 2008) with the ATB scheme (this scheme splits all clitics except the definite article and normalizes the Arabic characters alef and yaa). The German source is decompounded and part-of-speech-based long-range verb reordering rules (Popović and Ney, 2006) are applied.

From Table 1, we note that the general data is more than 20 times bigger than the in-domain data. A simple concatenation of the corpora might mask the phrase probabilities obtained from the in-domain corpus, causing a deterioration in performance. This is especially true for the Arabic-to-

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

English setup, where the UN data is 100 times bigger than the TED data and the domains are distinct. 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 weighting the sentences of the corpora differently, such that sentences which are more related to the domain will have higher weights and therefore have a stronger impact on the phrase probabilities.

5.2 Translation 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). In addition to the phrase based decoder, Jane includes an implementation of the forced alignment procedure used in this work for the purpose of adaptation. We use the standard set of models with phrase translation probabilities and word lexical smoothing for source-to-target and target-to-source directions, a word and phrase penalty, distance-based reordering and an n-gram target language model. In addition, our baseline includes binary count features which fire if the count of the phrase pair in the training corpus is smaller than a threshold. We use three count features with thresholds $\{2, 3, 4\}$.

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 *test2010* and *test2011* 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. In addition to the raw automatic results, we perform significance testing over all evaluations sets. For both BLEU and TER, 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.

6 Results

In this section we compare the suggested weighting schemes experimentally using the final translation quality. We use two TED tasks, German-to-English and Arabic-to-English translation. In addition to evaluating our suggested translation models based weighting schemes, we evaluate methods suggested in previous work, including LM based weighting and FA based adaptation.

The results for both German-to-English and Arabic-to-English TED tasks are summarized in Table 2. Each language pair section is divided into three subsections which differ by the phrase table training method. The first subsection is using state-of-the-art heuristic phrase extraction, the second is using FA adaptation and the third is using weighted phrase extraction with different weighting schemes.

To perform weighted phrase extraction, we use all data (*ALL*, a concatenation of *IN* and *OD*) as the general-domain data (in eq. 3 and Figure 1). This way, we ensure weighting for all sentences in the training data, and, data from *IN* is still used for the generation of the weighted phrase table.

6.1 German-to-English

Focusing on the German-to-English translation results, we note that using all data (ALL system) for the heuristic phrase extraction improves over the in-domain system (IN), with gains up-to +0.9% BLEU and -0.7% TER on the test2011 set. We perform significance testing in comparison to the ALL system as this is the best baseline system (among IN and ALL).

Mansour and Ney (2013) method of adaptation using the FA procedure (ALL-FA-IN) consistently outperforms the baseline system, with significant improvements on test10 TER.

Comparing the weighting schemes, weighting based on the phrase model (PM) and language model (LM) perform similarly, without a clear advantage to one method. The standalone weighting schemes do not achieve improvements over the baseline. Combining all the translation models (PM,SM,RM,CM,LP) into the TM scheme generates improvements over the standalone weighting schemes. TM also improves over the LM scheme suggested in previous work. We hypothesize that TM scoring is better for phrase model adaptation as it captures bilingual dependencies, unlike the LM scheme. In an experiment we do not report

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

System	dev		test2010		test2011	
	BLEU	TER	BLEU	TER	BLEU	TER
German-to-English						
IN	31.0	48.9	29.3	51.0	32.7	46.8
ALL	31.2	48.3	29.5	50.5	33.6	46.1
Forced alignment based adaptation						
ALL-FA-IN	31.8	47.4†	29.7	49.7†	33.6	45.5
Weighted phrase extraction						
LM	31.1	48.7	29.2	51.1	33.6	46.2
PM	31.5	48.8	29.2	50.9	33.1	46.4
TM	31.7	48.4	29.8	50.2	33.8	45.8
TM+LM	32.2†	47.5†	30.1	49.5‡	34.4†	44.8‡
Arabic-to-English						
IN	27.2	54.1	25.3	57.1	24.3	59.9
ALL	27.1	54.8	24.4	58.6	23.8	61.1
ALL-FA-IN	27.7	53.7	25.3	56.9	24.7	59.3
LM	28.1†	52.9‡	26.0	56.2†	24.6	59.3
PM	27.2	54.4	25.1	57.5	24.1	60.3
TM	27.4	53.9	25.4	57.0	24.4	59.5
TM+LM	28.3‡	52.8‡	26.2†	55.9‡	25.1†	58.7‡

Table 2: TED 2011 translation results. BLEU and TER are given in percentages. *IN* denotes the TED lectures in-domain corpus and *ALL* is using all available bilingual data (including *IN*). Significance is marked with † for 90% confidence and ‡ for 95% confidence, and is measured over the best heuristic system.

here, we tried to remove one translation model at a time from the TM scheme, the results always got worse. Therefore, we conclude that using all translation models is important to achieve robust weighting and generate the best results.

Combining TM with LM weighting (TM+LM) generates the best system overall. Significant improvements at the 95% level are observed for TER, BLEU is significantly improved for test11. TM+LM is significantly better than LM weighting on both test sets. In comparison to ALL-FA-IN, TM+LM is significantly better on test11 BLEU. TM+LM combines the advantages of both scoring methods, where TM ensures in-domain lexical choice while LM achieves better sentence fluency.

6.2 Arabic-to-English

To verify our results, we repeat the experiments on the Arabic-to-English TED task. The scenario is different here as using the OD data (UN) deteriorates the results of the IN system by 0.9% and 0.5% BLEU on test2010 and test2011 correspondingly. We attribute this deterioration to the large size of the UN data (a factor of 100 bigger than IN) which causes bias to OD. In addition, UN is more distinct from the TED lecture domain. We use the IN system as baseline and perform significance testing in comparison to this system.

FA adaptation (ALL-FA-IN) results are similar to the German-to-English section, with consistent improvements over the baseline but no significance is observed in this case.

For the weighting experiments, combining the translation models into the TM scheme improves over the standalone schemes. The LM scheme is performing better than TM in this case. We hypothesize that this is due to the big gap between the in-domain TED corpus and the other-domain UN corpus. The LM scheme is combining a term which overweights sentences further from the other-domain. This factor proves to be crucial in the case of a big gap between IN and OD. Such a term is not present in the translation model weighting schemes, we leave its incorporation for future work.

Finally, similar to the German-to-English results, the combined TM+LM achieves the best results, with significant improvements at the 90% level for all sets and error measures, and at the

Type	DE-EN		AR-EN	
	base	TM+LM	base	TM+LM
lexical	23695	23451	26679	25813
reorder	1193	1106	935	904

Table 3: Error analysis. A comparison of the error types along with the error counts are given. The systems include the baseline system and the TM+LM weighted system.

95% level for most. TM+LM improves over the baseline with +1.1% BLEU and -1.3% TER on dev, +0.9% BLEU and -1.2% TER on test2010 and +0.8% BLEU and -1.2% TER on test2011.

7 Error Analysis

In this section, we perform automatic and manual error analysis. For the automatic part, we use addicter[3] (Berka et al., 2012), which performs HMM word alignment between the reference and the hypothesis and measures lexical (word insertions, deletions and substitutions) and reordering errors. Addicter is a good tool to measure tendencies in the errors, but the number of errors might be misleading due to alignment errors. The summary of the errors is given in Table 3. From the table we clearly see that the majority of the improvement comes from lexical errors reduction. This is an indication of an improved lexical choice, due to the improved phrase model probabilities.

Translation examples are given in Table 4. The examples show that the lexical choice is being improved when using the weighted TM+LM phrase extraction. For the first example in German, “grossartig” means “great”, but translated by the baseline as “a lot”, which causes the meaning to be distorted. For the second Arabic example, the word معدل is ambiguous and could mean both “rate” and “modified”. The TM+LM system does the correct lexical choice in this case.

8 Conclusion

In this work, we investigate several weighting schemes for phrase extraction adaptation. Unlike previous work where language model scoring is used for adaptation, we utilize several translation models to perform the weighing.

Sample sentences	
src	es fuehlt sich grossartig an .
ref	it feels great .
base	it feels like a lot .
TM+LM	it feels great .
src	es haelt dich frisch .
ref	it keeps you fresh .
base	it’s got you fresh .
TM+LM	it keeps you fresh .
src	كيف ستقوم بإطعام العالم
ref	How are you going to feed the world
base	How will feed the world
TM+LM	How are you going to feed the world
src	ولماذا ؟ غذاء معدل جينيا
ref	And why? Genetically engineered food
base	And why ? Food rate genetically
TM+LM	And why ? Genetically modified food

Table 4: Sample sentences. The source, reference, baseline hypothesis and TM+LM weighted system hypothesis are given.

The translation models used for weighting are calculated over phrase aligned general-domain sentences using an in-domain phrase table.

Experiments on two language pairs show significant improvements over the baseline, with gains up-to +1.0% BLEU and -1.3% TER when using a combined TM and LM (TM+LM) weighting scheme. The TM+LM scheme also shows improvements over previous work, namely scoring using LM and using FA training to adapt a general-domain phrase table to the in-domain (ALL-FA-IN method).

In future work, we plan to investigate using translation model scoring in a fashion similar to the cross entropy difference framework. In this case, the general-domain data will be phrase aligned and scored using a general-domain phrase table, and the difference between the in-domain based scores and the general-domain ones can be calculated. Another interesting scenario we are planning to tackle is when only monolingual in-domain data exists, and whether our methods could be still applied and gain improvements, for example using automatic translations.

Acknowledgments

This material is based upon work supported by the DARPA BOLT project under Contract No. HR0011-12-C-0015.

[3] https://wiki.ufal.ms.mff.cuni.cz/user:zeman:addicter

References

Axelrod, Amittai, 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.

Berka, Jan, Ondrej Bojar, Mark Fishel, Maja Popovic, and Daniel Zeman. 2012. Automatic MT Error Analysis: Hjerson Helping Addicter. In *LREC*, pages 2158–2163, Istanbul, Turkey.

Cettolo, M Federico M, 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.

Foster, George 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.

Foster, George, 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.

Goto, Isao, 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.

Koehn, Philipp. 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.

Mansour, Saab 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.

Mansour, Saab and Hermann Ney. 2013. Phrase training based adaptation for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 649–654, Atlanta, Georgia, June. Association for Computational Linguistics.

Matsoukas, Spyros, 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.

Moore, Robert C. 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.

Och, Franz J. 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.

Papineni, Kishore, 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.

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

Roth, Ryan, 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.

Snover, Matthew, 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.

Wuebker, Joern and Hermann Ney. 2013. Length-incremental phrase training for smt. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pages 309–319, Sofia, Bulgaria, August.

Wuebker, Joern, 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.

Wuebker, Joern, 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.