A Simple and Effective Weighted Phrase Extraction for Machine Translation Adaptation

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

The task of domain-adaptation attempts to exploit data mainly drawn from one domain (e.g. news) to maximize the performance on the test domain (e.g. weblogs). In previous work, weighting the training instances was used for filtering dissimilar data. We extend this by incorporating the weights directly into the standard phrase training procedure of statistical machine translation (SMT). This allows the SMT system to make the decision whether to use a phrase translation pair or not, a more methodological way than discarding phrase pairs completely when using filtering. Furthermore, we suggest a combined filtering and weighting procedure to achieve better results while reducing the phrase table size. The proposed methods are evaluated in the context of Arabic-to-English translation on various conditions, where significant improvements are reported when using the suggested weighted phrase training. The weighting method also improves over filtering, and the combined filtering and weighting is better than a standalone filtering method. Finally, we experiment with mixture modeling, where additional improvements are reported when using weighted phrase extraction over a variety of baselines.

1. Introduction

Over the last 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, due to the success of SMT, new domains of translation are being explored, such as talk translation in the IWSLT TED evaluation [1] and dialects translation within the DARPA BOLT project [2]. The introduction of the BOLT project marks a shift in the Arabic NLP community, changing the focus from handling Modern Standard Arabic (MSA) structured data (e.g. news) to dialectal Arabic user generated noisy data (e.g. emails, weblogs). Dialectal Arabic is mainly spoken and scarcely written, even when it is written, the lack of common orthography causes significant variety and ambiguity in lexicon and morphology. The challenge is even greater due to the domain of informal communication, which is noisy by its nature. In this work, we perform experiments on both the BOLT and the IWSLT TED setups, allowing us to explore both lectures and weblogs domains, drawing more robust conclusions and enabling a larger group of researchers to reproduce our experiments and results.

The task of domain adaptation tackles the problem of utilizing existing resources in the most beneficial way for the new domain at hand. Given some general domain data and a new domain to tackle, adaptation is the task of modifying the SMT components in such a way that the new system will perform better on the new domain than the general domain system.

In this work, we focus on translation model (TM) adaptation. The TM (e.g. phrase model) is the core component of state-of-the-art SMT systems, providing the building blocks (e.g. phrase translation pairs) to perform the search for the best translation. Several methods were suggested already for TM adaptation. We experiment with training data weighting, where one assigns higher weights to relevant domain training instances, thus causing an increase of the corresponding probabilities. Therefore, translation pairs which can be obtained from relevant training instances will have a higher chance of being utilized during search.

Weighted phrase extraction can be done at several levels of granularity, including sub-corpus level, sentence level and phrase level. In this work, we focus on sentence level weighting for phrase extraction. Previous work also suggested filtering, which can be seen as a crude weighting were sentences are assigned $\{0, 1\}$ weights. We compare weighting to filtering and show superior results for weighting. In a scenario where efficiency constraints are imposed on the SMT system, reducing the TM size can serve as a solution. For such a scenario, we suggest filtering combined with weighting, and show that this method achieves better results than filtering alone.

Finally, we explore mixture modeling, where a purely in-domain TM is interpolated with various adapted TMs, and show further improvements. The resulting method described in this paper is simple and easy to reimplement, yet effective.

The rest of the paper is organized as follows. Related work on data filtering, weighting and mixture modeling is de-

tailed in Section 2. The weighted phrase extraction training and the method for assigning weights are described in Section 3. Section 4 recaps briefly mixture modeling methods that will be used in the paper. Experimental setup including corpora statistics and the SMT system are described in Section 5. The results of the described methods are summarized in Section 6. 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. The techniques include, among others: *(i)* semi-supervised training where one translates in-domain monolingual data and utilizes the automatic translations for retraining the LM and/or the TM ([3],[4]), *(ii)* different methods of interpolating in-domain and out-of-domain models ([5], [6], [7]) *(iii)* and sample weighting on the sentence or even the phrase level for LM training ([8],[9]) and TM training ([10],[11],[12]). Note that filtering is a special case of the sample weighting method where a threshold is assigned to discard unwanted samples.

Weighted phrase extraction can be done at several levels of granularity. [6] perform TM 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. [13] also performs weighting at the corpus level, but the weights are integrated into the phrase model estimation procedure. His method does not show an advantage over linear interpolation. A finer grained weighting is that of [10], 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. [11] 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 TM is important for good performance.

In this work, we propose a simple yet effective method for weighted phrase extraction expressing adaptation. Our method is comparable to [10] assigning each sentence pair in the training data a weight. We differ from them by using a weight based on the cross-entropy difference method proposed in [9] for LM filtering and later adapted in [12] for TM filtering. 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. We compare our weighting method to filtering and show superior results. In some cases, one might be interested in reducing the size of the TM for efficiency reasons. We combine filtering with weighting, and show that this leads to better performance than filtering alone.

Last, as done in some of the previous work mentioned above, we experiment with mixture modeling over the weighted phrase models. We use linear and log-linear interpolation similar to [6]. We differ from [13] by showing improved results over linear interpolation of baseline models. [14] analyze the effect of adding a general-domain corpus at different parts of the SMT training pipeline. A method denoted as "x+yE" performed best in their experiments. This method extracts all phrases from a concatenation of in-domain and general corpora, then, if a phrase pair exists in the in-domain phrase table it is assigned the in-domain probability, otherwise it is assigned the probability from the concatenation phrase table. We call this method an *ifelse* combination and test it in our experiments.

3. Weighted phrase extraction

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

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

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 (usually according to word alignment and heuristics) in sentence pair (s_r,t_r). One method to introduce weights to equation (1) is by weighting each sentence pair by a weight w_r. Equation (1) will now have the extended form:

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

It is easy to see that setting $\{w_r = 1\}$ will result in equation (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. Next, we discuss several methods for setting the weights in a fashion which serves adaptation.

3.1. Weight estimation

Several weighting schemes can be devised to manifest adaptation. One way is to manually assign suitable weights to corpora using information about genre, corpus provider, compilation method and other attributes of the corpora. For example, a higher weight (e.g. 10) can be assigned to in-domain

corpora sentences, while a lower weight (e.g. 1) is assigned to other corpora sentences.

LM cross-entropy scoring can be used for both monolingual data filtering for LM training as done in [9], or bilingual data filtering for TM training as done in [12]. Next, we recall the scoring methods introduced in the above previous work and utilize it for our proposed weighted phrase extraction method.

Given some corpus I which represents the domain we want to adapt to, and a general corpus O, [9] first generate a random subset $\hat{O} \subseteq O$ of approximately the same size as I (this is not required for the method to work, and is used to make the models generated by the corpora more comparable), and train the LMs LM_I and $LM_{\hat{O}}$ using the corresponding training data. Then, each sentence $o \in O$ is scored according to:

$$H_I(o) - H_{\hat{O}}(o) \tag{3}$$

where $H_M(o)$ ($M \in \{I, \hat{O}\}$) is the per-word cross-entropy according to a language model trained on M. Let $o = w_1 \dots w_n$, then we have

$$H_M(o) = -\frac{1}{n}\sum_{i=1}^{n} \log p_M(w_i|w_{i-1}) \tag{4}$$

for a 2-gram LM case.

The intuition behind equation (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. [9] show that using equation (3) performs better in terms of perplexity than using in-domain cross-entropy only ($H_I(o)$). For more details about the reasoning behind equation (3) we refer the reader to [9].

[12] adapted the LM scores for bilingual data filtering for the purpose of TM training. In this case, we have source and target in-domain corpora I_{src} and I_{trg}, and correspondingly, general corpora O_{src} and O_{trg}, with random subsets $\hat{O}_{src} \subseteq O_{src}$ and $\hat{O}_{trg} \subseteq O_{trg}$. Then, we score each sentence pair (s_r, t_r) by:

$$d_r = [H_{I_{src}}(s_r) - H_{\hat{O}_{src}}(s_r)] + [H_{I_{trg}}(t_r) - H_{\hat{O}_{trg}}(t_r)] \tag{5}$$

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:

$$w_r = e^{-d_r} \tag{6}$$

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

As done in [12], we compare using (5) to source only cross-entropy difference $[H_{I_{src}}(s) - H_{\hat{O}_{src}}(s)]$ and target only cross-entropy difference $[H_{I_{trg}}(t) - H_{\hat{O}_{trg}}(t)]$, in addition to source only in-domain cross-entropy $H_{I_{src}}(s)$.

4. Mixture modeling

Mixture modeling is a technique for combining several models using weights assigned to the different components. Domain adaptation could be achieved using mixture modeling when the weights are related to the proximity of the components to the domain being translated. As we generate several translation models differing by the training corpora domain and extraction method, interpolating these models could yield further improvements. In this work, we focus on two variants of mixture modeling, namely linear and loglinear interpolation.

4.1. Linear interpolation

Linear interpolation is a commonly used framework for combining different SMT models into one ([6]). As we experiment with interpolating two phrase models in this work (in-domain and other-domain), we obtain the following simplified interpolation formula:

$$p(\tilde{f}|\tilde{e}) = \lambda p_I(\tilde{f}|\tilde{e}) + (1-\lambda)p_O(\tilde{f}|\tilde{e}) = \tag{7}$$

λ is assigned a value in the range $[0, 1]$ to keep the resulting phrase model normalized. We set the value empirically on the development set testing different λ with steps of 0.1. Phrase pairs which appear in one model but not in the second are assigned small probabilities by the second model. The probabilities of the final mixture model are renormalized.

4.2. Loginear interpolation

Loglinear interpolation of phrase models fits directly into the loglinear framework of SMT ([7]). The weights of the different phrase models could be then tuned directly within the tuning procedure of the SMT system. This results in doubling the number of phrase model features, which could cause additional search errors, overfitting and finding an inferior local optima. Again, we assign a small probability to unknown phrase pairs. In this case, we do not perform renormalization to avoid overweighting of unknown phrase pairs.

5. Experimental setup

5.1. Training corpora

To evaluate the introduced methods experimentally, we use the BOLT Phase 1 Dialectal-Arabic-to-English task. The dialect chosen for Phase 1 is Egyptian Arabic (henceforth *Egyptian*). We confirm our findings by some final experiments on the IWSLT 2011 TED Arabic-to-English task.

The BOLT program goes beyond previous projects, shifting the focus from translating structured standardized text, such as Modern Standard Arabic (MSA) newswire, to a user generated noisy text such as Arabic dialect emails or weblogs. Translating Arabic dialects is a challenging task due to the scarcity of training data and the lack of common orthography causing a larger vocabulary size and higher ambigu-

Data style	Sentences	Tokens
United Nations	3557K	122M
Newswire	1918K	57M
Web	13K	280K
Newsgroup	25K	720K
Broadcast	91K	2M
Lexicons	213K	530K
Iraqi, Levantine	617K	4M
General (sum of above)	6434K	187M
Egyptian	240K	3M

Table 1: BOLT bilingual training corpora style and statistics. The number of tokens is given for the source side.

Set	Sen	Tok	OOV/IN	OOV/ALL
BOLT P1 Egyptian-to-English				
Egy (IN)	240K	3M		
General	6.4M	187M		
dev	1219	18K	387 (2.2%)	160 (0.9%)
test	1510	27K	559 (2.1%)	201 (0.7%)
IWSLT 2011 TED Arabic-to-English				
TED (IN)	90K	1.6M		
UN	7.9M	228M		
dev	934	19K	408 (2.2%)	184 (1.0%)
test	1664	31K	495 (1.6%)	228 (0.8%)

Table 2: 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). ALL denotes the concatenation of all training data for the specific task.

ity. Due to the scarcity of in-domain training data, MSA resources are being utilized for the project. In such a scenario, an important research question arises on how to use the MSA data in the most beneficial way to translate the given dialect. The training data for the BOLT Phase 1 program is summarized in Table 1. The table includes data style and size information. Most of the BOLT training data is available through the linguistic data consortium (LDC) and is regularly part of the NIST open MT evaluation [1].

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. It is important to stress that IWSLT 2011 is different from previous years' campaigns by the genre shifting from the traveling domain (BTEC task) to lectures (TED task). Further, the amount of training data provided for the TALK task is considerably larger than for the BTEC task. For Arabic-to-English, the bilingual data consists of roughly 100K sentences of in-domain TED talks data and 8M sentences of out-of-domain United Nations (UN) data. This makes the task more similar to real-life MT system conditions, and the discrepancy between the training and the test domain opens a window for a variety of adaptation methods.

The bilingual training and test data for the Egyptian-to-English and Arabic-to-English tasks are summarized in Table 2[2]. The English data was tokenized and lowercased while the Arabic data was tokenized and segmented with the ATB scheme (this scheme splits all clitics except the definite article and normalizes the Arabic characters alef and yaa).

From Table 2, we note that the general data considerably reduces the number of out-of-vocabulary (OOV) words. This comes with the price of increasing the size of the training data by a factor of more than 50. 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 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.

For language model training purposes, we use an additional 8 billion words for BOLT (4B words from the LDC gigaword corpus and 4B words collected from web resources) and 1.4 billion words for IWSLT (supplied as part of the campaign monolingual training data [3]).

5.2. Translation system

The baseline system is built using a state-of-the art phrase-based SMT system similar to Moses [15]. 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 lexical models are trained on the in-domain portion of the data and kept constant throughout the experiments. This way we achieve more control on the variability of the experiments. In the experiments, we update the phrase probability features in both directions of translation. The SMT systems are tuned on the *dev* development set with minimum error rate training [16] using BLEU [17] accuracy measure as the optimization criterion. We test the performance of our system on the *test* set using the BLEU and translation edit rate (TER) [18] measures. We use TER as an additional measure to verify the consistency of our improvements and avoid over-tuning. The BOLT results are case insensitive while the IWSLT results are case sensitive. In addition to the raw automatic results, we perform significance testing over the *test*

[1] For a list of the NIST MT12 corpora, see `http://www.nist.gov/itl/iad/mig/upload/OpenMT12_LDCAgreement.pdf`

[2] The test sets for BOLT are extracted from the LDC2012E30 corpus - BOLT Phase 1 DevTest Source and Translation V4.

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

Translation model	dev		test	
	BLEU	TER	BLEU	TER
Unfiltered				
EGY	24.6	61.2	22.2	62.6
EGY+GEN	25.3	60.6	22.5	61.9
Filtered				
EGY+GEN-1Mbest	25.4	60.5	22.9	61.6
EGY+GEN-1Mrand	25.3	60.6	22.6	61.7
Weighted phrase extr.				
10EGY+1GEN	25.6	60.2	22.8	61.5
ppl_1-src(EGY+GEN)	25.6	60.7	22.9	61.5
ppl-src(EGY+GEN)	25.6	60.6	23.3‡	61.0
ppl-trg(EGY+GEN)	25.6	60.6	22.8	61.8
ppl(EGY+GEN)	25.6	60.1	23.3‡	60.9‡
ppl(EGY+GEN1Mbest)	25.6	60.0	23.0	61.4
Mixture modeling				
-loglin-EGY+GEN	24.7	61.3	22.0	62.8
-loglin-ppl(EGY+GEN)	24.9	61.1	22.1	62.3
-linear-EGY+GEN	25.7	60.4	22.9	61.4
-linear-ppl(EGY+GEN)	26.0	59.9	23.3‡	60.6‡
-ifelse-EGY+GEN	25.6	60.2	23.0	61.1
-ifelse-ppl(EGY+GEN)	25.7	60.2	23.1	61.0

Table 3: BOLT 2012 Egyptian-English translation results. BLEU and TER results are in percentages. *EGY* denotes the Egyptian in-domain corpus, *GEN* denotes the general other corpora. Significance is marked with ‡ and measured over the *EGY+GEN* baseline.

set. For both BLEU and TER, we perform bootstrap resampling with bounds estimation as described in [19]. 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 proposed methods of weighted phrase extraction against unfiltered (in-domain and full) and filtered translation model systems. We start by testing our methods on the BOLT task, and finally verify the results on the IWSLT task.

6.1. BOLT results

The results of the BOLT Phase 1 Egyptian-English task are summarized in Table 3. Adding the general-domain (*GEN*) corpora to the in-domain (*EGY*) corpora system (unfiltered) increases the translation quality slightly by +0.3% BLEU on the *test* set. This increase might be attributed to the fact that the number of OOVs is decreased by adding the GEN corpora three folds. But, in addition, the various corpora that assemble the general-domain corpus are collected from various resources, increasing the possibility that there exists relevant training data to the domain being tackled.

When adding to *EGY* a filtered *GEN* corpus, where the 1000K best sentences according to the bilingual cross-entropy difference (equation (5)) are kept (*EGY+GEN-1000K-best*), the results improve by another +0.4% BLEU on *test* in comparison to the full *EGY+GEN* system. Thus, the filtering is able to retain sentences which are more relevant to the domain being translated. As a control experiment, we selected 1000K sentences from the *GEN* corpus randomly and added them to the *EGY* corpus (*EGY+GEN-1000K-rand*). In the BOLT setup, the cross-entropy based filtering seems to have only slight edge over random selection, perhaps due to the generality and usefulness of *GEN*.

In the third block of experiments, we compare the suggested methods for weighted phrase extraction. In the first experiment, we give higher weights to bilingual sentences from in-domain (10) as opposed to smaller weights to the general corpus (1). The resulting system (*10EGY+1GEN*) is comparable to the filtered *EGY+GEN-1000K-best*. In comparison to the *EGY+GEN* baseline, small improvements are observed on *dev* (+0.3% BLEU) and on *test* (+0.3% BLEU). Next, we compare the suggested weighting schemes, including source only in-domain cross-entropy based (denoted by *ppl_1-src* in the table), source only cross-entropy difference (*ppl-src*), target only cross-entropy difference (*ppl-trg*) and bilingual cross-entropy difference (*ppl*). We weight the bilingual training sentences (both in-domain and general-domain *EGY+GEN*) by the corresponding perplexity weight. All the weighting schemes improve over the baseline, where *ppl_1-src* and *ppl-trg* perform worst among the methods, and bilingual cross-entropy difference *ppl* has a slight edge on TER over source side only *ppl-src*. The *ppl(EGY+GEN)* system achieves the best results where +0.8% BLEU and -1.0% TER are observed on *test* in comparison to the *EGY+GEN* baseline. The improvements on both BLEU and TER are statistically significant at the 95% level, the only system being able to achieve that among weighted and filtered systems. In the final experiment, we combine filtering with weighting, where the best 1000K sentences of *GEN* are concatenated to *EGY* and a weighted phrase extraction using perplexity is done over this concatenation (*ppl(EGY+GEN-1000K-best)*). This system improves slightly over the unweighted *EGY+GEN-1000K-best* system, with +0.2% BLEU and -0.5% TER on *dev*, and +0.1% BLEU and -0.2% TER on *test*. Thus, if one is interested in a smaller TM, filtering combined with weighting is the best method to use according to our experiments.

In the last block of experiments, model combination is tested. We compare mixing the in-domain TM *EGY* with standard *EGY+GEN* TM and weighted *ppl(EGY+GEN)* one, using log-linear and linear interpolation as done in [6], and ifelse combination as done in [14]. The first observation is that log-linear interpolation performs poorly and worse than linear interpolation, supporting the results of [6] and [13] and contradicting [12]. [12] describe a special case where the overlap between the combined phrase tables in their experiments is small, which could explain the difference. Linear

Translation model	dev		test	
	BLEU	TER	BLEU	TER
Unfiltered				
TED	27.2	54.1	25.3	57.1
TED+UN	27.1	54.8	24.4	58.6
Filtered				
TED+UN-1Mbest	27.7	53.7	25.5	56.9
TED+UN-1Mrand	27.4	54.0	25.1	57.1
Weighted phrase extr.				
10TED+1UN	28.2	53.4	25.4	56.8
ppl_I-src(TED+UN)	27.9	53.3	25.5	55.8
ppl-src(TED+UN)	28.1	53.2	26.0	56.5
ppl-trg(TED+UN)	28.0	53.0	25.8	56.2
ppl(TED+UN)	28.1	52.9	26.0	56.2†
ppl(TED+UN-1Mbest)	28.1	53.1	25.8	56.3
Mixture modeling				
-loglin-TED+UN	26.8	53.9	24.0	58.3
-loglin-ppl(TED+UN)	27.2	53.9	24.7	57.6
-linear-TED+UN	28.0	53.1	25.9	56.2†
-linear-ppl(TED+UN)	28.1	53.3	25.9	56.1‡
-ifelse-TED+UN	28.4	52.6	25.9	56.0
-ifelse-ppl(TED+UN)	28.2	52.8	25.7	56.4

Table 4: IWSLT TED 2011 Arabic-English translation results. BLEU and TER results are in percentages. *TED* denotes the TED lectures in-domain corpus, *UN* denotes the united nations corpus. Significance is marked with ‡ and measured over the *TED* baseline.

combination on the other hand performs well, always improving over the respective combined standalone TMs. The mixture weight value for linear interpolation is set empirically by ranging the weight of the in-domain corpus *EGY* from $[0, 1]$ with steps of 0.1. The best result on the development set was achieved for a weight of 0.9. The linear mixture of *EGY* and *EGY+GEN* already achieves large improvements over the baseline. Still, interpolation with the weighted phrase table system (*EGY-linear-ppl(EGY+GEN)*) achieves the best results, improving over the mixture counterpart *EGY-linear-EGY+GEN* by +0.4% BLEU and up-to -0.8% TER on *test*. For both linear interpolation settings, $\lambda = 0.9$ for equation (7) performed best on the development set. Even though the ifelse combination is rather simple, the results are surprisingly good, still, the best linear combination performs better than the ifelse method. Similar to the other combination methods, using the weighted phrase table has a slight edge over the unweighted counterpart.

6.2. IWSLT TED results

The results of the IWSLT TED 2011 Arabic-English task are summarized in Table 4. Unlike the BOLT task, adding the out-of-domain *UN* corpus to the in-domain *TED* corpus system decreases the translation quality by -0.9% BLEU on the *test* set. This suggests a big discrepancy between the in-domain and the out-of-domain bilingual training corpora. Even though the *UN* corpus decreases the OOV ratio by a factor of 2 according to Table 2, the 100 times larger *UN* corpus masking the in-domain phrase probabilities seems to be more important and decisive for the degradation in performance. This claim is supported by the result of the *TED+UN-1000K-rand* system, which improves over *TED+UN*, due to the smaller *UN* selection that is being used and reducing the contamination of the in-domain phrase probabilities. When adding to *TED* a filtered *UN* corpus, where the 1000K best sentences according to the bilingual cross-entropy difference are kept (*TED+UN-1000K-best*), the results improve by 0.8% BLEU on *dev*, but smaller improvement of 0.2% BLEU is observed on *test*. In the context of filtering, cross-entropy based filtering is again performing better than random selection.

In the third block of experiments, we compare the suggested methods for weighted phrase extraction. The trends are similar to the BOLT results, where the perplexity based weighting achieves the best results and big improvements over the in-domain baseline, where the improvements on TER are statistically significant at the 90% level. A combined filtering and weighting (*ppl(TED+UN-1000K-best)*) performs better than unweighted filtering (*TED+UN-1000K-best*) by +0.3% BLEU and bigger -0.6% TER improvements on *test*.

For the mixture modeling results, loglinear interpolation decreases the performance dramatically, while linear interpolation achieves comparable results to the best weighted extraction, and no further improvements were observed. We hypothesize that mixture modeling did not yield improvements for IWSLT due to the big discrepancy between *TED* and *UN*, limiting the margin of improvements that is possible to achieve.

7. Conclusions

In this work, we utilize cross-entropy based weights for domain adaptation. We extend on previous work, where the weights are used for filtering purposes, by incorporating the weights directly into the standard maximum likelihood estimation of the phrase model. The weighted phrase extraction influences the phrase translation probabilities, while keeping the set of phrase pairs intact. We find this a more methodological way for adaptation than a hard decision where filtering is done. In some scenarios where efficiency constraints are imposed on the SMT system, filtering might be necessary. We propose a combined filtering and weighting method.

The proposed methods are evaluated in the context of Arabic-to-English translation on two conditions, IWSLT TED MSA lectures and BOLT Egyptian weblogs. The weighted phrase extraction method shows consistent improvements on both tasks, with up-to +1.1% BLEU and -1.7% TER improvements over the purely in-domain BOLT baseline, and +0.7% BLEU and -0.9% TER over the TED

baseline. The new method is also improving over filtering, and the combined filtering and weighting is better than a standalone filtering method. Thus, if one is interested in a smaller TM, filtering combined with weighting is the best method to use according to our experiments.

Finally, we tried mixture modeling of the in-domain and the various adapted TMs. Log-linear interpolation performed poorly in our experiments, which is consistent with previous work. On the other hand, linear interpolation performed well, achieving comparable results to the best system on the TED task, and further improvements on the BOLT task. We hypothesize that interpolation could not help for the TED task due to the big distance between the (scientific, cultural) lectures and the parliamentary discussions domains, limiting the improvement range of adaptation at the sentence level. On the BOLT task, interpolation with weighted phrase extraction performed better than interpolation with a standard phrase model, supporting the good performance of our suggested new method.

In future work, it will be interesting to compare different weighting methods in the weighted maximum likelihood estimation framework. Additionally, the effect of the granularity of weighting could be evaluated, comparing sentence versus corpus versus documents (any set of sentences) weighting.

8. Acknowledgements

This work was partially realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and also partially funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 4911028154.0. 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.

9. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stker, "Overview of the IWSLT 2011 evaluation campaign," in *International Workshop on Spoken Language Translation*, 2011, pp. 11–27.

[2] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. M. Schwartz, J. Makhoul, O. Zaidan, and C. Callison-Burch, "Machine translation of arabic dialects," in *HLT-NAACL*, 2012, pp. 49–59.

[3] N. Ueffing, G. Haffari, and A. Sarkar, "Transductive learning for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 25–32. [Online]. Available: http://www.aclweb.org/anthology/P07-1004

[4] H. Schwenk, "Investigations on large-scale lightly-supervised training for statistical machine translation," in *International Workshop on Spoken Language Translation*, 2008, pp. 182–189.

[5] Y. Lu, J. Huang, and Q. Liu, "Improving statistical machine translation performance by training data selection and optimization," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 343–350. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1036

[6] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0717

[7] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 224–227. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0733

[8] J. Gao, J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for chinese," *ACM Transactions on Asian Language Information Processing*, vol. 1, pp. 3–33, March 2002. [Online]. Available: http://doi.acm.org/10.1145/595576.595578

[9] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 220–224. [Online]. Available: http://www.aclweb.org/anthology/P10-2041

[10] S. Matsoukas, A.-V. I. Rosti, and B. Zhang, "Discriminative corpus weight estimation for machine translation," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, August 2009, pp. 708–717. [Online]. Available: http://www.aclweb.org/anthology/D/D09/D09-1074

[11] G. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, October 2010, pp. 451–459. [Online]. Available: http://www.aclweb.org/anthology/D10-1044

[12] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 355–362. [Online]. Available: http://www.aclweb.org/anthology/D11-1033

[13] R. Sennrich, "Perplexity minimization for translation model domain adaptation in statistical machine translation," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 539–549. [Online]. Available: http://www.aclweb.org/anthology/E12-1055

[14] B. Haddow and P. Koehn, "Analysing the effect of out-of-domain data on smt systems," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 422–432. [Online]. Available: http://www.aclweb.org/anthology/W12-3154

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantine, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, June 2007, pp. 177–180.

[16] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003, pp. 160–167.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[18] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "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*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.

[19] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Barcelona, Spain, July 2004, pp. 388–395.