Adaptation of Reordering Models for Statistical Machine Translation

Boxing Chen, George Foster and Roland Kuhn
National Research Council Canada
first.last@nrc-cnrc.gc.ca

Abstract

Previous research on domain adaptation (DA) for statistical machine translation (SMT) has mainly focused on the translation model (TM) and the language model (LM). To the best of our knowledge, there is no previous work on reordering model (RM) adaptation for phrase-based SMT. In this paper, we demonstrate that mixture model adaptation of a lexicalized RM can significantly improve SMT performance, even when the system already contains a domain-adapted TM and LM. We find that, surprisingly, different training corpora can vary widely in their reordering characteristics for particular phrase pairs. Furthermore, particular training corpora may be highly suitable for training the TM or the LM, but unsuitable for training the RM, or vice versa, so mixture weights for these models should be estimated separately. An additional contribution of the paper is to propose two improvements to mixture model adaptation: smoothing the in-domain sample, and weighting instances by document frequency. Applied to mixture RMs in our experiments, these techniques (especially smoothing) yield significant performance improvements.

1 Introduction

A phrase-based statistical machine translation (SMT) system typically has three main components: a translation model (TM) that contains information about how to translate word sequences (phrases) from the source language to the target language, a language model (LM) that contains information about probable word sequences in the target language, and a reordering model (RM) that indicates how the order of words in the source sentence is likely to influence the order of words in the target sentence. The TM and the RM are trained on parallel data, and the LM is trained on target-language data. Usage of language and therefore the best translation practice differs widely across genres, topics, and dialects, and even depends on a particular author's or publication's style; the word "domain" is often used to indicate a particular combination of all these factors. Unless there is a perfect match between the training data domain and the (test) domain in which the SMT system will be used, one can often get better performance by adapting the system to the test domain.

In offline domain adaptation, the system is provided with a sample of translated sentences from the test domain prior to deployment. In a popular variant of offline adaptation, linear mixture model adaptation, each training corpus is used to generate a separate model component that forms part of a linear combination, and the sample is used to assign a weight to each component (Foster and Kuhn, 2007). If the sample resembles some of the corpora more than others, those corpora will receive higher weights in the combination.

Previous research on domain adaptation for SMT has focused on the TM and the LM. Such research is easily motivated: translations across domains are unreliable. For example, the Chinese translation of the English word "mouse" would most likely be "laoshu 老鼠" if the topic is the animal; if the topic is computer hardware, its translation would most

Proceedings of NAACL-HLT 2013, pages 938–946,
Atlanta, Georgia, 9–14 June 2013.

likely be "shubiao 鼠标". However, when the translation is for people in Taiwan, even when the topic is computer hardware, its translation would more likely be "huashu 滑鼠". It is intuitively obvious why TM and LM adaptation would be helpful here.

By contrast, it is not at all obvious that RM model adaptation will improve SMT performace. One would expect reordering behaviour to be characteristic of a particular language pair, but not of particular domains. At most, one might think that reordering is lexicalized—perhaps, (for instance) in translating from Chinese to English, or from Arabic to English, there are certain words whose English translations tend to undergo long-distance movement from their original positions, while others stay close to their original positions. However, one would not expect a particular Chinese adverb or a particular Arabic noun to undergo long-distance movement when being translated into English in one domain, but not in others. Nevertheless, that is what we observe: see section 5 below.

This paper shows that RM adaptation improves the performance of our phrase-based SMT system. In our implementation, the RM is adapted by means of a linear mixture model, but it is likely that other forms of RM adaptation would also work. We obtain even more effective RM adaptation by smoothing the in-domain sample and by weighting orientation counts by the document frequency of the phrase pair. Both improvements could be applied to the TM or the LM as well, though we have not done so.

Finally, the paper analyzes reordering to see why RM adaptation works. There seem to be two factors at work. First, the reordering behaviour of words and phrases often differs dramatically from one bilingual corpus to another. Second, there are corpora (for instance, comparable corpora and bilingual lexicons) which may contain very valuable information for the TM, but which are poor sources of RM information; RM adaptation downweights information from these corpora significantly, and thus improves the overall quality of the RM.

2 Reordering Model

In early SMT systems, such as (Koehn, 2004), changes in word order when a sentence is translated were modeled by means of a penalty that is incurred when the decoder chooses, as the next source phrase to be translated, a phrase that does not immediately follow the previously translated source sentence. Thus, the system penalizes deviations from monotone order, with the magnitude of the penalty being proportional to distance in the source sentence between the end of the previously translated source phrase and the start of the newly chosen source phrase.

Many SMT systems, including our own, still use this distance-based penalty as a feature. However, starting with (Tillmann and Zhang, 2005; Koehn et al., 2005), a more sophisticated type of reordering model has often been adopted as well, and has yielded consistent performance gains. This type of RM typically identifies three possible orientations for a newly chosen source phrase: monotone (M), swap (S), and discontinuous (D). The M orientation occurs when the newly chosen phrase is immediately to the right of the previously translated phrase in the source sentence, the S orientation occurs when the new phrase is immediately to the left of the previous phrase, and the D orientation covers all other cases.[1] This type of RM is lexicalized: the estimated probabilities of M, S and D depend on the source-language and target-language words in both the previous phrase pair and the newly chosen one.

Galley and Manning (2008) proposed a "hierarchical" lexicalized RM in which the orientation (M, S, or D) is determined not by individual phrase pairs, but by blocks. A block is the largest contiguous sequence of phrase pairs that satisfies the phrase pair consistency requirement of having no external links. Thus, classification of the orientation of a newly chosen phrase as M, S, or D is carried out as if the decoder always chose the longest possible source phrase in the past, and will choose the longest possible source phrase in the future.

The RM used in this paper is hierarchical and lexicalized. For a given phrase pair (f, e), we estimate the probabilities that it will be in an M, S, or D orientation o with respect to the previous phrase pair and the following phrase pair (two separate distributions). Orientation counts $c(o, f, e)$ are obtained from a word-aligned corpus using the method de-

[1] Some researchers have distinguished between left and right versions of the D orientation, but this 4-orientation scheme has not yielded significant gains over the 3-orientation one.

scribed in (Cherry et al., 2012), and corresponding probabilities $p(o|f,e)$ are estimated using recursive MAP smoothing:

$$p(o|f,e) = \frac{c(o,f,e) + \alpha_f \, p(o|f) + \alpha_e \, p(o|e)}{c(f,e) + \alpha_f + \alpha_e}$$

$$p(o|f) = \frac{c(o,f) + \alpha_g \, p(o)}{c(f) + \alpha_g}$$

$$p(o) = \frac{c(o) + \alpha_u/3}{c(\cdot) + \alpha_u}, \qquad (1)$$

where $p(o|e)$ is defined analogously to $p(o|f)$, and the four smoothing parameters α_e, α_f, α_g, and α_u are set to values that minimize the perplexity of the resulting model on held-out data.

During decoding, orientations with respect to the previous context are obtained from a shift-reduce parser, and orientations with respect to following context are approximated using the coverage vector (Cherry et al., 2012).

3 RM Adaptation

3.1 Linear mixture model

Following previous work (Foster and Kuhn, 2007; Foster et al., 2010), we adopt the linear mixture model technique for RM adaptation. This technique trains separate models for each training corpus, then learns weights for each of the models and combines the weighted component models into a single model.

If we have N sub-corpora, the global reordering model probabilities $p(o|f,e)$ are computed as in (2):

$$p(o|f,e) = \sum_{i=1}^{N} \alpha_i \, p_i(o|f,e) \qquad (2)$$

where $p_i(o|f,e)$ is the reordering model trained on sub-corpus i, and α_i is its weight.

Following (Foster et al., 2010), we use the EM algorithm to learn the weights that maximize the probability of phrase-pair orientations in the development set (in-domain data):

$$\hat{\alpha} = \operatorname*{argmax}_{\alpha} \sum_{o,f,e} \tilde{p}(o,f,e) \log \sum_{i=1}^{N} \alpha_i \, p_i(o|f,e) \qquad (3)$$

where $\tilde{p}(o,f,e)$ is the empirical distribution of counts in the dev set (proportional to $c(o,f,e)$). Two separate sets of mixing weights are learned: one for the distribution with respect to the previous phrase pair, and one for the next phrase pair.

3.2 Development set smoothing

In Equation 3, $\tilde{p}(o,f,e)$ is extracted from the in-domain development set. Since dev sets for SMT systems are typically small (1,000-3,000 sentences), we apply smoothing to this RM. We first obtain a smoothed conditional distribution $p(o|f,e)$ using the MAP technique described above, then multiply by the empirical marginal $\tilde{p}(e,f)$ to obtain a final smoothed joint distribution $p(o,f,e)$.

There is nothing about this idea that limits it to the RM: smoothing could be applied to the statistics in the dev that are used to estimate a mixture TM or LM, in order to mitigate over-fitting. However, we note that, compared to the TM, the over-fitting problem is likely to be more acute for the RM, since it splits counts for each phrase pair into three categories.

3.3 Document-frequency weighting

Mixture models, like the RM in this paper, depend on the existence of multiple training corpora, with each sub-corpus nominally representing a domain. A recent paper suggests that some phrase pairs belong to general language, while others are domain-specific (Foster et al., 2010). If a phrase pair exists in all training corpora, it probably belongs to general language; on the other hand, if it appears in only one or two training corpora, it is more likely to be domain-specific.

We were interested in seeing whether information about domain-specificity could improve the estimation of mixture RM weights. The intuition is that phrase pairs that belong to general language should contribute more to determining sub-corpus weights, since they are the ones whose reordering behaviour is most likely to shift with domain. To capture this intuition, we multiplied the empirical distribution in (3) by the following factor, inspired by the standard document-frequency formula:

$$D(f,e) = \log(DF(f,e) + K), \qquad (4)$$

where $DF(f,e)$ is the number of sub-corpora that (f,e) appears in, and K is an empirically-determined smoothing term.

corpus	# segs	# en tok	%	genres
fbis	250K	10.5M	3.7	nw
financial	90K	2.5M	0.9	financial
gale_bc	79K	1.3M	0.5	bc
gale_bn	75K	1.8M	0.6	bn ng
gale_nw	25K	696K	0.2	nw
gale_wl	24K	596K	0.2	wl
hkh	1.3M	39.5M	14.0	Hansard
hkl	400K	9.3M	3.3	legal
hkn	702K	16.6M	5.9	nw
isi	558K	18.0M	6.4	nw
lex&ne	1.3M	2.0M	0.7	lexicon
others_nw	146K	5.2M	1.8	nw
sinorama	282K	10.0M	3.5	nw
un	5.0M	164M	58.2	un
TOTAL	10.1M	283M	100.0	(all)
devtest				
tune	1,506	161K		nw wl
NIST06	1,664	189K		nw bn ng
NIST08	1,357	164K		nw wl

Table 1: NIST Chinese-English data. In the *genres* column: nw=newswire, bc=broadcast conversation, bn=broadcast news, wl=weblog, ng=newsgroup, un=United Nations proceedings.

4 Experiments

4.1 Data setting

We carried out experiments in two different settings, both involving data from NIST Open MT 2012.[2] The first setting uses data from the Chinese to English constrained track, comprising 283M English tokens. We manually identified 14 sub-corpora on the basis of genres and origins. Table 1 summarizes the statistics and genres of all the training corpora and the development and test sets; for the training corpora, we show their size in number of words as a percentage of all training data. Most training corpora consist of parallel sentence pairs. The *isi* and *lex&ne* corpora are exceptions: the former is extracted from comparable data, while the latter is a lexicon that includes many named entities. The development set (*tune*) was taken from the NIST 2005 evaluation set, augmented with some web-genre material reserved from other NIST corpora.

corpus	# segs	# en toks	%	genres
gale_bc	57K	1.6M	3.3	bc
gale_bn	45K	1.2M	2.5	bn
gale_ng	21K	491K	1.0	ng
gale_nw	17K	659K	1.4	nw
gale_wl	24K	590K	1.2	wl
isi	1,124K	34.7M	72.6	nw
other_nw	224K	8.7M	18.2	nw
TOTAL	1,512K	47.8M	100.0	(all)
devtest				
NIST06	1,664	202K		nw wl
NIST08	1,360	205K		nw wl
NIST09	1,313	187K		nw wl

Table 2: NIST Arabic-English data. In the *genres* column: nw=newswire, bc=broadcast conversation, bn=broadcase news, ng=newsgroup, wl=weblog.

The second setting uses NIST 2012 Arabic to English data, but excluding the UN data. There are about 47.8 million English running words in these training data. We manually grouped the training data into 7 groups according to genre and origin. Table 2 summarizes the statistics and genres of all the training corpora and the development and test sets. Note that for this language pair, the comparable isi data represent a large proportion of the training data: 72% of the English words. We use the evaluation sets from NIST 2006, 2008, and 2009 as our development set and two test sets, respectively.

4.2 System

Experiments were carried out with an in-house phrase-based system similar to Moses (Koehn et al., 2007). The corpus was word-aligned using IBM2, HMM, and IBM4 models, and the phrase table was the union of phrase pairs extracted from these separate alignments, with a length limit of 7. The translation model was smoothed in both directions with KN smoothing (Chen et al., 2011). The DF smoothing term K in equation 4 was set to 0.1 using held-out optimization. We use the hierarchical lexicalized RM described above, with a distortion limit of 7. Other features include lexical weighting in both directions, word count, a distance-based RM, a 4-gram LM trained on the target side of the parallel data, and a 6-gram English *Gigaword* LM. The sys-

[2] http://www.nist.gov/itl/iad/mig/openmt12.cfm

system	Chinese	Arabic
baseline	31.7	46.8
baseline+loglin	29.6	45.9
RMA	31.8	47.7**
RMA+DF	32.2*	47.9**
RMA+dev smoothing	32.3*	**48.3****
RMA+dev smoothing+DF	**32.8****	48.2**

Table 3: Results for variants of RM adaptation.

system	Chinese	Arabic
LM+TM adaptation	33.2	47.7
+RMA+dev-smoothing+DF	33.5	48.4**

Table 4: RM adaptation improves over a baseline containing adapted LMs and TMs.

system	Chinese	Arabic
baseline	31.7	46.8
LM adaptation	32.1*	47.0
TM adaptation	33.0**	47.5**
RM adaptation	32.8**	48.2**

Table 5: Comparison of LM, TM, and RM adaptation.

tem was tuned with batch lattice MIRA (Cherry and Foster, 2012).

4.3 Results

For our main baseline, we simply concatenate all training data. We also tried augmenting this with separate log-linear features corresponding to sub-corpus-specific RMs. Our metric is case-insensitvie IBM BLEU-4 (Papineni et al., 2002); we report BLEU scores averaged across both test sets. Following (Koehn, 2004), we use the bootstrap-resampling test to do significance testing. In tables 3 to 5, * and ** denote significant gains over the baseline at $p < 0.05$ and $p < 0.01$ levels, respectively.

Table 3 shows that reordering model adaptation helps in both data settings. Adding either document-frequency weighting (equation 4) or dev-set smoothing makes the improvement significant in both settings. Using both techniques together yields highly significant improvements.

Our second experiment measures the improvement from RM adaptation over a baseline that includes adapted LMs and TMs. We use the same technique—linear mixtures with EM-tuned weights—to adapt these models. Table 4 shows that adapting the RM gives gains over this strong baseline for both language pairs; improvements are significant in the case of Arabic to English.

The third experiment breaks down the gains in the last line of table 4 by individual adapted model. As shown in table 5, RM adaptation yielded the largest improvement on Arabic, while TM adaptation did best on Chinese. Surprisingly, both methods significantly outperformed LM adaptation, which only achieved significant gains over the baseline for Chinese.

5 Analysis

Why does RM adaptation work? Intuitively, one would think that reordering behaviour for a given phrase pair should not be much affected by domain, making RM adaptation pointless. That is probably why (as far as we know) no-one has tried it before. In this section, we describe three factors that account for at least part of the observed gains.

5.1 Weighting by corpus quality

One answer to the above question is that some corpora are better for training RMs than others. Furthermore, corpora that are good for training the LM or TM are not necessarily good for training the RM, and vice versa. Tables 6 and 7 illustrate this. These list the weights assigned to various sub-corpora for LM, TM, and RM mixture models.

The weights assigned to the *isi* sub-corpus in particular exhibit a striking pattern. These are high in the LM mixtures, moderate in the TM mixtures, and very low in the RM mixtures. When one considers that isi contains 72.6% of the English words in the Arabic training data (see table 2), its weight of 0.01 in the RM mixture is remarkable.

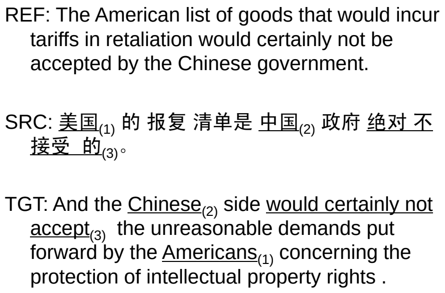
On reflection, it makes sense that EM would assign weights in the order it does. The *isi* corpus consists of comparable data: sentence pairs whose source- and target-language sides are similar, but often not mutual translations. These are a valuable source of in-domain n-grams for the LM; a somewhat noisy source of in-domain phrase pairs for the TM; and an unreliable source of re-ordering patterns for the RM. Figure 1 shows this. Although the two

LM	TM	RM
isi (0.23)	un (0.29)	un (0.21)
gale_nw (0.11)	fbis (0.15)	gale_nw (0.13)
un (0.11)	hkh (0.10)	lex&ne (0.12)
sino. (0.09)	gale_nw (0.09)	hkh (0.08)
fbis (0.08)	gale_bn (0.07)	fbis (0.08)
fin. (0.07)	oth_nw (0.06)	gale_bn (0.08)
oth_nw (0.07)	sino. (0.06)	gale_wl (0.06)
gale_bn (0.07)	isi (0.05)	gale_bc (0.06)
gale_wl (0.06)	hkn (0.04)	hkn (0.04)
hkh (0.06)	fin. (0.04)	fin. (0.04)
hkn (0.03)	gale_bc (0.03)	oth_nw (0.03)
gale_bc (0.02)	gale_wl (0.02)	hkl (0.03)
lex&ne (0.00)	lex&ne (0.00)	isi (0.01)
hkl (0.00)	hkl (0.00)	sino. (0.01)

Table 6: Chinese-English sub-corpora for LM, TM, and RM mixture models, ordered by mixture weight.

LM	TM	RM
isi (0.41)	isi (0.35)	gale_bc (0.21)
oth_nw (0.19)	oth_nw (0.29)	gale_ng (0.20)
gale_ng (0.15)	gale_bc (0.10)	gale_nw (0.20)
gale_wl (0.09)	gale_ng (0.08)	oth_nw (0.13)
gale_nw (0.07)	gale_bn (0.07)	gale_ng (0.12)
gale_bc (0.05)	gale_nw (0.07)	gale_wb (0.11)
gale_bn (0.02)	gale_wl (0.05)	isi (0.01)

Table 7: Arabic-English sub-corpora for LM, TM, and RM mixture models, ordered by mixture weight.


REF: The American list of goods that would incur tariffs in retaliation would certainly not be accepted by the Chinese government.

SRC: 美国$_{(1)}$ 的 报复 清单是 中国$_{(2)}$ 政府 绝对 不 接受 的$_{(3)}$。

TGT: And the Chinese$_{(2)}$ side would certainly not accept$_{(3)}$ the unreasonable demands put forward by the Americans$_{(1)}$ concerning the protection of intellectual property rights .

Figure 1: Example of sentence pair from comparable data; underlined words with the same number are translations of each other

Corpus	M	S	D	Count
fbis	0.50	0.07	0.43	685
financial	0.32	0.28	0.41	65
gale_bc	0.60	0.10	0.31	50
gale_bn	0.47	0.15	0.37	109
gale_nw	0.51	0.05	0.44	326
gale_wl	0.42	0.26	0.32	52
hkh	0.29	0.23	0.48	130
hkl	0.28	0.16	0.56	263
hkn	0.30	0.27	0.43	241
isi	0.24	0.16	0.60	240
lex&ne	0.94	0.03	0.02	1
others_nw	0.29	0.16	0.55	23
sinorama	0.44	0.07	0.49	110
un	0.37	0.10	0.53	15
dev	0.46	0.24	0.31	11

Table 8: Orientation frequencies for the phrase pair "立即 immediately", with respect to the previous phrase.

sides of the comparable data are similar, they give the misleading impression that the phrases labeled 1, 2, 3 in the Chinese source should be reordered as 2, 3, 1 in English. We show a reference translation of the Chinese source (not found in the comparable data) that reorders the phrases as 1, 3, 2.

Thus, RM adaptation allows the RM to learn that certain corpora whose reordering information is of lower quality corpora should have lower weights. The optimal weights for corpora inside an RM may be different from the optimal weights inside a TM or LM.

5.2 Weighting by domain match

So is this all that RM adaptation does: downweight poor-quality data? We believe there is more to RM adaptation than that. Specifically, even if one considers only high-quality data for training RMs (ignoring comparable data, etc.) one sees differences in reordering behaviour between different domains. This isn't just because of differences in word frequencies between domains, because we observe domain-dependent differences in reordering for the same phrase pair. Two examples are given below: one Chinese-English, one Arabic-English.

Table 8 shows reordering data for the phrase pair "立即 immediately" in various corpora. Notice the strong difference in behaviour between the three Hong Kong corpora—*hkh, hkl* and *hkn*—and some of the other corpora, for instance *fbis*. In the

Corpus	M	S	D	Count
gale_bc	0.50	0.27	0.22	233
gale_bn	0.56	0.21	0.23	226
gale_ng	0.51	0.13	0.37	295
gale_nw	0.47	0.20	0.33	167
gale_wl	0.56	0.18	0.26	127
isi	0.50	0.06	0.44	5502
other_nw	0.50	0.16	0.34	1450
dev	0.75	0.12	0.13	52

Table 9: Orientation frequencies for the phrase pair “work AlEml” with respect to the previous phrase.

Hong Kong corpora, *immediately* is much less likely (probability of around 0.3) to be associated with a monotone (M) orientation than it is in *fbis* (probability of 0.5). This phrase pair is relatively frequent in both corpora, so this disparity seems too great to be due to chance.

Table 9 shows reordering behaviour for the phrase pair “work AlEml”[3] across different sub-corpora. As in the Chinese example, there appear to be significant differences in reordering patterns for certain corpora. For instance, *gale_bc* swaps this well-attested phrase pair twice as often (probability of 0.27) as *gale_ng* (probability of 0.13).

For Chinese, it is possible that dialect plays a role in reordering behaviour. In theory, Mandarin Chinese is a single language which is quite different, especially in spoken form, from other languages of China such as Cantonese, Hokkien, Shanghainese, and so on. In practice, many speakers of Mandarin may be unconsciously influenced by other languages that they speak, or by other languages that they don’t speak but that have an influence over people they interact with frequently. Word order can be affected by this: the Mandarin of Mainland China, Hong Kong and Taiwan sometimes has slightly different word order. Hong Kong Mandarin can be somewhat influenced by Cantonese, and Taiwan Mandarin by Hokkien. For instance, if a verb is modified by an adverb in Mandarin, the standard word order is “adverb verb”. However, since in Cantonese, “verb adverb” is a more common word order, speakers and writers of Mandarin in Hong Kong may adopt the “verb adverb” order in that language as well. Figure 2 shows how a different word order in the Mandarin source affects reordering when translating into English. Perhaps in situations where different training corpora represent different dialects, RM adaptation involves an element of dialect adaptation. We are eager to test this hypothesis for Arabic—different dialects of Arabic are much more different from each other than dialects of Mandarin, and reordering is often one of the differences—but we do not have access to Arabic training, dev, and test data in which the dialects are clearly separated.

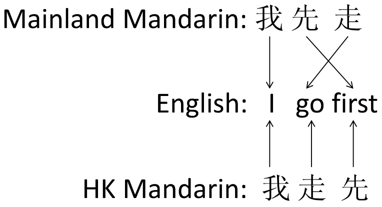


Figure 2: An example of different word ordering in Mandarin from different area.

It is possible that RM adaptation also has an element of genre adaptation. We have not yet been able to confirm or refute this hypothesis. However, whatever is causing the corpus-dependent reordering patterns for particular phrase pairs shown in the two tables above, it is clear that they may explain the performance improvements we observe for RM adaptation in our experiments.

5.3 Penalizing highly-specific phrase pairs

In section 3.3 we described our strategy for giving general (high document-frequency) phrase pairs that occur in the dev set more influence in determining mixing weights. An artifact of our implementation applies a similar strategy to the probability estimates for *all* phrase pairs in the model. This is that 0 probabilities are assigned to all orientations whenever a phrase pair is absent from a particular sub-corpus.

Thus, for example, a pair (f, e) that occurs only in sub-corpus i will receive a probability $p(o|f, e) = \alpha_i\, p_i(o|f, e)$ in the mixture model (equation 2). Since $\alpha_i \leq 1$, this amounts to a penalty on pairs that occur in few sub-corpora, especially ones with low mixture weights.

The resulting mixture model is deficient (non-

[3] We represent the Arabic word *AlEml* in its Buckwalter transliteration.

normalized), but easy to fix by backing off to a global distribution such as $p(o)$ in equation 1. However, we found that this "fix" caused large drops in performance, for instance from the Arabic BLEU score of 48.3 reported in table 3 to 46.0. We therefore retained the original strategy, which can be seen as a form of instance weighting. Moreover, it is one that is particularly effective in the RM, since, compared to a similar strategy in the TM (which we also employ), it applies to whole phrase pairs and results in much larger penalties.

6 Related work

Domain adaptation is an active topic in the NLP research community. Its application to SMT systems has recently received considerable attention. Previous work on SMT adaptation has mainly focused on translation model (TM) and language model (LM) adaptation. Approaches that have been tried for SMT model adaptation include mixture models, transductive learning, data selection, data weighting, and phrase sense disambiguation.

Research on mixture models has considered both linear and log-linear mixtures. Both were studied in (Foster and Kuhn, 2007), which concluded that the best approach was to combine sub-models of the same type (for instance, several different TMs or several different LMs) linearly, while combining models of different types (for instance, a mixture TM with a mixture LM) log-linearly. (Koehn and Schroeder, 2007), instead, opted for combining the sub-models directly in the SMT log-linear framework.

In transductive learning, an MT system trained on general domain data is used to translate in-domain monolingual data. The resulting bilingual sentence pairs are then used as additional training data (Ueffing et al., 2007; Chen et al., 2008; Schwenk, 2008; Bertoldi and Federico, 2009).

Data selection approaches (Zhao et al., 2004; Lü et al., 2007; Moore and Lewis, 2010; Axelrod et al., 2011) search for bilingual sentence pairs that are similar to the in-domain "dev" data, then add them to the training data. The selection criteria are typically related to the TM, though the newly found data will be used for training not only the TM but also the LM and RM.

Data weighting approaches (Matsoukas et al., 2009; Foster et al., 2010; Huang and Xiang, 2010; Phillips and Brown, 2011; Sennrich, 2012) use a rich feature set to decide on weights for the training data, at the sentence or phrase pair level. For instance, a sentence from a corpus whose domain is far from that of the dev set would typically receive a low weight, but sentences in this corpus that appear to be of a general nature might receive higher weights.

The 2012 JHU workshop on Domain Adaptation for MT [4] proposed phrase sense disambiguation (PSD) for translation model adaptation. In this approach, the context of a phrase helps the system to find the appropriate translation.

All of the above work focuses on either TM or LM domain adaptation.

7 Conclusions

In this paper, we adapt the lexicalized reordering model (RM) of an SMT system to the domain in which the system will operate using a mixture model approach. Domain adaptation of translation models (TMs) and language models (LMs) has become common for SMT systems, but to our knowledge this is the first attempt in the literature to adapt the RM. Our experiments demonstrate that RM adaptation can significantly improve translation quality, even when the system already has TM and LM adaptation. We also experimented with two modifications to linear mixture model adaptation: dev set smoothing and weighting orientation counts with document frequency of phrase pairs. Both ideas are potentially applicable to TM and LM adaptation. Dev set smoothing, in particular, seems to improve the performance of RM adaptation significantly. Finally, we investigate why RM adaptation helps SMT performance. Three factors seem to be important: downweighting information from corpora that are less suitable for modeling reordering (such as comparable corpora), dialect/genre effects, and implicit instance weighting.

[4] http://www.clsp.jhu.edu/workshops/archive/ws-12/groups/dasmt

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP 2011*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, Athens, March. WMT.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting n-best hypotheses for smt self-enhancement. In *ACL 2008*.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and transforming feature functions: New ways to smooth phrase tables. In *MT Summit 2011*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *NAACL 2012*.

Colin Cherry, Robert C. Moore, and Chris Quirk. 2012. On hierarchical re-ordering and permutation parsing for phrase-based decoding. In *WMT 2012*.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, June. WMT.

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 (EMNLP)*, Boston.

Michel Galley and C. D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP 2008*, pages 848–856, Hawaii, October.

Fei Huang and Bing Xiang. 2010. Feature-rich discriminative phrase rescoring for SMT. In *COLING 2010*.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June. Association for Computational Linguistics.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, D. Talbot, and M. White. 2005. Edinburgh system description for the 2005 NIST MT evaluation. In *Proceedings of Machine Translation Evaluation Workshop*.

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. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Demonstration Session*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, Georgetown University, Washington D.C., October. Springer-Verlag.

Yajuan Lü, Jin Huang, and Qun Liu. 2007. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague, Czech Republic.

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 (EMNLP)*, Singapore.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *ACL 2010*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July. ACL.

Aaron B. Phillips and Ralf D. Brown. 2011. Training machine translation with a second-order taylor approximation of weighted translation instances. In *MT Summit 2011*.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *IWSLT 2008*.

Rico Sennrich. 2012. Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. In *EACL 2012*.

Christoph Tillmann and Tong Zhang. 2005. A localized prediction model for statistical machine translation. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, July. ACL.

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

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the International Conference on Computational Linguistics (COLING) 2004*, Geneva, August.