Combining Domain and Topic Adaptation for SMT

Eva Hasler[1] ehasler@ed.ac.uk
Barry Haddow[1] bhaddow@inf.ed.ac.uk
Philipp Koehn[1,2] pkoehn@inf.ed.ac.uk
[1]ILCC, School of Informatics, University of Edinburgh
[2]Center for Language and Speech Processing, Johns Hopkins University

Abstract

Recent years have seen increased interest in adapting translation models to test domains that are known in advance as well as using latent topic representations to adapt to unknown test domains. However, the relationship between domains and latent topics is still somewhat unclear and topic adaptation approaches typically do not make use of domain knowledge in the training data. We show empirically that combining domain and topic adaptation approaches can be beneficial and that topic representations can be used to predict the domain of a test document. Our best combined model yields gains of up to 0.82 BLEU over a domain-adapted translation system and up to 1.67 BLEU over an unadapted system, measured on the stronger of two training conditions.

1 Introduction

Domain adaptation is a very active area of research in statistical machine translation (SMT) and there is a large and growing body of work on different techniques to adapt translation and language models (Foster and Kuhn, 2007; Matsoukas et al., 2009; Foster et al., 2010; Sennrich, 2012) to specific target domains that are usually known in advance, for example the *news* domain. An extension of the standard domain adaptation task is *multi-domain adaptation* where a translation system is adapted to several known target domains (see for example Cui et al. (2013)). In cases where the target domains are not assumed to be known, dedicated domain classifiers can be trained and used to automatically predict the target domain and choose an appropriate model based on the prediction (Banerjee et al., 2010).

Recently, there has been increased work on the application of topic modelling to translation model adaptation (Gong et al., 2011; Su et al., 2012; Eidelman et al., 2012; Hewavitharana et al., 2013; Hasler et al., 2014a) in an attempt to move away from the notion of a *domain* as the source of a corpus. *Topic adaptation* techniques build on the assumption that the origin of sentences and documents is unknown at test time, and the unsupervised nature of topic models is useful for detecting structure across corpus boundaries in training sets to adapt to diverse test sets. While domain adaptation techniques rely on a given, hard clustering of the data, topic adaptation aims to induce a soft clustering that is more suited to the task.

While topic models are very useful for detecting and grouping the semantic differences in documents, making use of our knowledge about corpus boundaries in the training data could potentially help to adapt more specifically to style or genre on top of adapting to topics. By predicting the domain label of test documents, we can combine both approaches to translate unlabelled documents from different genres and topics. We show that domain and topic adaptation can be complementary and that finding the right balance between the two could lead to a more

efficient architecture that combines online and offline computation.

2 Related work

Domain classification for multi-domain adaptation has been the focus of several researchers in recent years. Xu et al. (2007) tune domain-specific features weights and build domain-specific language models. They use the perplexity of in-domain language models to classify test documents and select the appropriate weights and models per document. Banerjee et al. (2010) train domain-specific translation models and use SVMs to detect the domain of an input sentence to route it to a domain-specific model. Wang et al. (2012) follow a slightly different approach by re-using the same translation model for all domains and tuning domain-specific features weights with modified objectives. Sennrich et al. (2013) adapt the four standard translation model features to unsupervised clusters of the development data obtained by k-means clustering.

Another line of research aims to improve topic modelling by encoding domain information via a Dirichlet Forest Prior (Andrzejewski et al., 2009). By specifying Must-Link and Cannot-Link relations between words, topic modelling is guided to either separate words into different topics or merge them into the same topic. While the idea of combining domain and topic adaptation within the same model is appealing, the model requires manually constructed lists of words and seems more suited for fine-tuning specific topics, a process they call *interactive topic modeling*.

Different from previous work, we investigate the utility of combining domain adaptation with topic adaptation to capture potentially different levels of structure in test documents of unknown origin. We also show that topic modelling makes it straightfoward to predict the domain of a test document, circumventing the need for separately trained domain classifiers. This allows us to combine domain-adapted translation models with topic-adapted models dynamically at test time.

3 Topic modelling approach

We follow the approach described in Hasler et al. (2014b) to build a Phrase Pair Topic (PPT) model. We use a model with 50 topics for all translation experiments and evaluate different numbers of topics for the domain classifiers. In the PPT model, phrase pairs are represented as *distributional profiles* which are pseudo documents containing source words found in all the sentence contexts of a phrase pair in the training data. These pseudo documents are the input to the topic model which clusters context words into topics and infers a topic distribution for each phrase pair. At test time, context topic vectors are inferred by applying the model to each of the test documents. This contextual information can be used to measure the cosine similarity between a document context and each applicable phrase pair or for other topic-adapted features as described in the next section. Context adaptation works by adding topic-adapted features to each phrase pair in the filtered phrase table, depending on its topical similarity with the test document. Thus, the topic-adapted features are recomputed for each test document.

3.1 Topic features

We consider several sets of topic features all of which are derived from distributions learned by the PPT model. The feature sets contain the individual features described below, where s and t denote a source and target phrase, c denotes the test document context, k denotes a latent topic and θ denotes a topic vector:

Conditional translation probability

$$p(t|s,c) = \sum_k p(t,k|s,c)$$
$$p(t,k|s,c) \propto p(t,s,k|c)$$
$$= p(t|s,k) \cdot p(s|k) \cdot p(k|c)$$

Joint-conditional probability

$$p(t,c|s) = p(c|t,s) \cdot p(t|s)$$
$$\approx p(\theta_c|\theta_{pp}) \cdot p(t|s)$$
$$\approx cos(\theta_c|\theta_{pp}) \cdot p(t|s)$$

Target-unigrams

$$trgUnigrams_t = \prod_{i=1}^{|t|} f(\frac{P_{doc}(w_i)}{P_{baseline}(w_i)}) \cdot f(\frac{P_{doc}(w_i)}{P_{topic0}(w_i)})$$
$$f(x) = \frac{2}{1+\frac{1}{x}}, \quad x > 0$$

Sim-phrasePair similarity = $cos(\theta_{pp}, \theta_c)$
Sim-targetPhrase similarity = $cos(\theta_{tp}, \theta_c)$
Sim-targetWord similarity = $cos(\theta_{tw}, \theta_c)$

The first two features are probabilistic features that take the topical context into account in computing the probability of a target phrase given a source phrase. The first feature, **Conditional**, factorises the joint probability of a target phrase t, source phrase s and topic k given a context c into the probabilities $p(t|s,k)$ and $p(s|k)$, which are estimated from relative counts of how often source and target phrases co-occur with each topic in the distributional profiles, and $p(k|c)$ which represents the inferred topic mixture for the test context. The second feature, **Joint-conditional**, estimates the joint probability of a target phrase and a test context given a source phrase. It is factorised as the (baseline) probability of a target phrase given a source phrase and the probability of the test context given the source and target phrase. The latter is approximated by the probility of the test context topic mixture given the phrase pair topic mixture, which is further approximated by the cosine similarity between the two topic mixtures.

The **Target-unigrams** feature is inspired by the lazy MDI adaptation of Ruiz and Federico (2012) and measures the probability ratio of a word under the document topic mixture versus under the baseline model[1]. We include an additional term to measure the topical relevance of a word by comparing against its probability under the asymmetric topic 0 of the PPT model[2]. **Sim-phrasePair** measures the cosine similarity of a phrase pair topic vector and the topic vector of a test context. **Sim-targetPhrase** is similar but uses an average topic vector over all phrase pairs with the same target phrase. **Sim-targetWord** instead replaces the phrase pair topic vector with the word topic vector of the word in the target phrase with the lowest topical entropy[3]. Target word topic vectors are derived from phrase pair topic vectors by averaging over all vectors of phrase pairs that include the target word. The features **Conditional**, **Target-unigrams** and

[1]The baseline model here corresponds to the relative frequency of target unigrams in the training data.

[2]Topic 0 has higher a priori probability and is supposed to capture common words that occur in the context of many translation units (Hasler et al., 2014b).

[3]The intuition behind this feature is that words with low topical entropy are expected to be more topically relevant.

Sim-phrasePair are similar or equivalent to features described in Hasler et al. (2014a,b), the remaining three features are new.

While the probabilistic features have some notion of the frequency of translations in the training corpus[4], similarity features are purely based on topic information and could be unreliable for rare translation units. On the other hand, similarity features are more efficient to compute at test time than the conditional translation probability which requires summation and normalisation for each phrase pair.

For the adaptation experiments, we evaluate a topic feature set that contains all the features described above, as well as smaller subsets thereof. The large feature set overlaps with the unadapted and domain-adapted features sets in that each contains probabilistic translation features. The smaller sets only contain features that have no direct correspondence in the baseline models. The topic feature sets are defined as:

Overlap Conditional, Joint-conditional, Target-unigrams, Sim-phrasePair, Sim-targetPhrase, Sim-targetWord
Sim-combine similarity = $\frac{1}{3}$ (Sim-phrasePair + Sim-targetPhrase + Sim-targetWord)
Sim-combine-loglin Sim-phrasePair, Sim-targetPhrase, Sim-targetWord
Sim-combine+trgUnigrams Sim-combine, Target-unigrams

4 Data

Our experiments were carried out on a French-English data set consisting of the TED corpus (Cettolo M. and Federico, 2012), parts of the News Commentary corpus (NC) and parts of the Commoncrawl corpus (CC) from the WMT13 shared task (Bojar et al., 2013) as described in Table 1, condition 1. To ensure that the baseline model does not have an implicit preference for any particular domain, we selected subsets of the NC and CC corpora such that the training data contains 2.7M English words per domain. The data set simulates an environment where very diverse documents have to be translated, which is a typical scenario for web translation engines, for example.

We evaluate our models on a second training set (condition 2) where we add the Europarl corpus to the translation and language model training data. This increases the number of parallel training sentences to 2.3M. For condition 2, the phrase pair contexts for topic modelling are extracted from a much larger number of sentence pairs, therefore we sample up to 50 contexts per phrase pair to keep the training size manageable. We also do not learn topic vectors for singleton phrase pairs or phrase pairs that occur more than 20000 times in the training data, as we expect that such pairs are less dependent on the context.

Data	Mixed		CC	NC	TED	Europarl
Train (condition 1)	354K	(6450)	110K	103K	140K	-
Train (condition 2)	2.3M		110K	103K	140K	1.9M
Dev	2453	(39)	818	817	818	-
Test	5664	(112)	1892	1878	1894	-

Table 1: Number of sentence pairs and documents (in brackets) in the data sets.

5 Predicting domain labels

While previous approaches to automatic domain classification have built dedicated classifiers such as SVMs and perceptrons or used in-domain language model perplexity, we use our

[4]For **Conditional**, this is implicit in the number of context words in a distributional profile.

trained topic model to assign domain labels to documents. We apply the PPT model to all documents from the training domains of condition 1 (CC, NC, TED) to get one topic vector per training document. We then experiment with three types of classifiers using the induced topic vectors:

Single-prototype Compute the average of all document vectors of the same training domain ($\rightarrow$ domain vectors), then compute the cosine similarity of a test document with the three domain vectors and predict the domain with the highest similarity.

Multi-prototype Compute the cosine similarity of a test document with the topic vectors of all training documents and predict the domain according to the label of the most similar training document.

Single-prototype-threshold Like single-prototype but with a threshold of 0.35 for prediction[5]. If a test document is not similar to any of the domain vectors according to the threshold, predict *unknown* and use the baseline model in place of a domain-adapted model.

The results of the single- and multi-prototype classifiers on the development and test documents are shown in Table 2. While for NC and TED documents, we can get perfect domain predictions with the single-protoype classifier, the accuracy on CC is at most 0.82, depending on the number of latent topics in the topic vectors. However, the multi-prototype classifier does better for CC in all cases. This suggests that there are subclusters of documents in the CC corpus to which some of the CC test documents are similar while not being as similar to a global average of all CC documents. Table 3 shows the accuracy of the single-protoype classifier when using a fixed threshold, with the results split into *correct*, *other* and *unknown*. While NC and TED documents are still labelled accurately, the proportion of correct predictions drops for CC. This confirms that NC and TED can be considered domains in the sense that the documents all have certain properties in common, while this is not the case for CC. This is also supported by Figure 1 which shows the average domain vectors for each of the three corpora, with some of the topical peaks labelled according to their most likely words. While CC documents can belong to rather diverse clusters such as IT, arts, hotel reviews or speech, NC documents belong to more related topics along the themes of politics and economy. These topics are more likely to be active within the same document and thus a document with political or economical content would likely overlap with the NC domain vector on several dimensions. TED documents share two topical components that capture words that are typical in speech like 1st and 2nd person verb forms (*speech*) as well as a rather broad *science* topic. Thus, a document with a high proportion of these verb forms would be likely to be classified as TED.

Model	**CC**		**NC**		**TED**	
# dev+test docs	88		39		24	
	sgl	mlt	sgl	mlt	sgl	mlt
k=10	0.70	**0.88**	**1.0**	0.95	**1.0**	0.96
k=20	0.82	**0.94**	**1.0**	0.97	**1.0**	1.0
k=50	0.73	**0.93**	**1.0**	0.95	**1.0**	1.0
k=100	0.76	**0.93**	**1.0**	1.0	**1.0**	0.92

Table 2: Accuracy of domain prediction using single-prototype (sgl) or multi-prototype (mlt) domain vectors with different numbers of topics (k).

[5]Cosine similarity ranges from 0 to 1. The threshold was set on the development set.

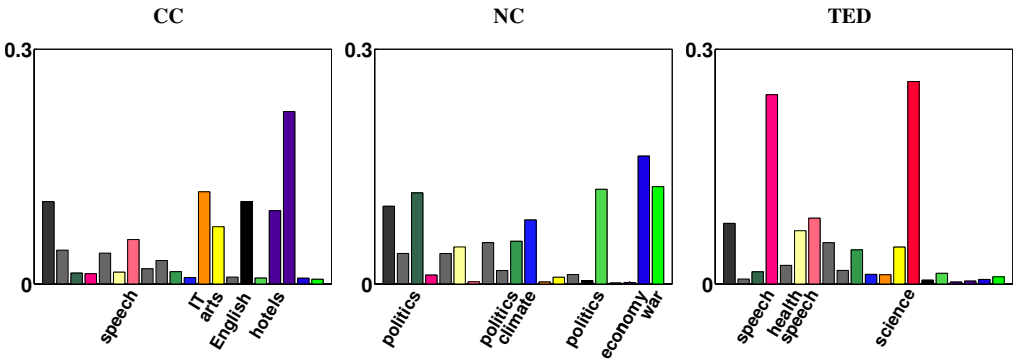

Figure 1: Average domain vectors (20 topics) for Commoncrawl, News Commentary and TED corpus.

We further observe in Table 3 that the rate of predicting *unknown* for CC documents increases with the number of topics. The reason for this is that we use the same classification threshold for all k while the cosine similarities of higher-dimensional topic vectors are typically lower than those of low-dimensional vectors[6]. For the experiments in the next section, we used the single-prototype-threshold classifier with $k = 50$.

Model	**CC**			**NC**	**TED**
# dev+test docs	88			39	24
	corr	other	unk	corr	corr
k=10	0.68	0.30	0.02	1.0	1.0
k=20	0.76	0.15	0.09	1.0	1.0
k=50	0.60	0.19	0.21	1.0	1.0
k=100	0.55	0.12	0.33	1.0	1.0

Table 3: Accuracy of domain prediction using single-prototype vectors with a threshold of 0.35 and different numbers of topics (k). *Corr*: correct domain predicted, *other*: wrong domain predicted, *unk*: no domain predicted.

6 Experimental setup

All of the test corpora contain document boundaries which allows us to consider document context during translation and switch translation and language models at document boundaries. While the domain-adapted baselines use gold domain labels, we use automatically predicted domains when combining domain-adapted and topic-adapted models[7]. We use a tuning set containing data from all three test domains and tune a single set of feature weights for all portions of the test set. Translation quality is evaluated using the average feature weights of three optimisation runs with PRO (Hopkins and May, 2011). We use the mteval-v13a.pl script to compute case-insensitive BLEU scores and use bootstrap resampling (Koehn, 2004) to measure significance of the BLEU scores on the mixed test set.

[6]This trend was observed by Banchs and Costa-jussà (2011) for vectors derived from Latent Semantic Indexing.
[7]Note that topic adaptation does not rely on domain labels.

6.1 Unadapted baseline system

Our baseline is a phrase-based French-English system trained on the concatenation of all parallel data for condition 1 and 2, respectively. It was built with the Moses toolkit (Koehn et al., 2007) using the 14 standard core features including a 5-gram language model, trained on the concatenation of the target sides of the training data.

6.2 Domain-adapted systems

We use the linear mixture model (DA-TM) of Sennrich (2012) (available in the Moses toolkit) to adapt the translation model to each of the three test domains CC, NC and TED. The domain labels of the documents are used to group documents of the same domain together. We build adapted tables for each domain by treating the remaining documents as out-of-domain data. For development and test, the domain labels are used to select the respective adapted model for decoding. We also use domain-adapted language models (DA-LM) which are linear interpolations of separate language models for each training domain, tuned to minimise perplexity on an in-domain development set per domain.

6.3 Topic-adapted systems

In order to integrate document-specific features into decoding, we build a (filtered) phrase table with topic-adapted features for each test document which is loaded before decoding each document. It would be straightforward to achieve a tighter integration with the SMT system by setting up feature functions that have access to document-level information, but for now we use a simple architecture where a wrapper script runs the decoder for each document.

7 Results

In this section we present results of different combinations of the baseline model, the domain-adapted and the topic-adapted models. Results are reported separately per test domain as well as on the entire mixed test set. We first describe results for training condition 1 in Sections 7.1 and 7.2, before showing results on training condition 2 in Section 7.3. We also provide qualitative evaluation of translation outputs in Section 7.4.

7.1 Overlapping topic feature set

Table 4 shows the results when adding the overlapping topic feature set (containing probabilistic and non-probabilistic translation features) on top of unadapted and domain-adapted systems. Adding topic-adapted features always yields improvements over the respective baseline system, even though the amount of previous adaptation has an influence on the relative gain.

Topic adaptation works best for TED documents but we observe that the improvement decreases with increasing domain-adaptation. Depending on the amount of previous adaptation, the BLEU improvements range between 1.44 and 0.31. These results add to our observations from Section 5 that on top of showing the characteristics of a domain, TED documents exhibit a further layer of structure that can be exploited with topic adaptation. For CC, the improvement of topic adaptation is quite stable at between 0.6 and 0.7 BLEU because domain adaptation has almost no effect on performance here. This is in line with our observations from Section 5 that CC behaves least like a domain in comparison with the other test corpora. For NC documents, the topic-adapted features yield a small improvement of 0.24 BLEU over the unadapted system but no further improvement over the domain-adapted models. A possible explanation is that because of the close relation between the dominant topics in the NC corpus (politics/economy), domain adaptation methods are sufficient to capture the important characteristics of the documents.

Model	Mixed	CC	NC	TED
Baseline	26.86	19.61	29.42	31.88
+ topics	*27.57	20.35	29.68	33.22
	+0.71	+0.74	+0.26	+1.34
DA-TM	27.24	19.61	29.87	32.73
+ topics	***27.73**	**20.33**	**29.88**	**33.55**
	+0.49	+0.69	+0.01	+0.82
DA-LM	27.16	19.71	29.77	32.46
+ topics	*27.60	20.37	29.80	33.20
	+0.44	+0.63	+0.03	+0.74
DA-TM+LM	27.34	19.59	29.92	33.02
+ topics	*27.63	20.22	29.90	33.33
	+0.29	+0.60	-0.02	+0.31
Gain of best system over baseline	+0.87	+0.72	+0.46	+1.67

Table 4: BLEU scores of unadapted/adapted baseline models (training condition 1) and additional topic-adapted features (Overlap) with their gain over the respective baseline (bottom of each block). The best system on the mixed test set is marked in bold. *: $p \leq 0.001$ marks significantly better scores compared to the respective baseline.

Overall, the best results on the mixed test set are achieved with a combination of domain and topic adaptation of the translation model (DA-TM + topics). This system yields a 0.82 BLEU improvement over the DA-TM model and a 1.67 improvment BLEU over the unadapted baseline, on TED documents. On the mixed test set, the gain over the DA-TM model is 0.49 and the overall gain is 0.87 BLEU.

7.2 Smaller topic feature sets

While the results from the previous section show that topic adaptation is beneficial at all levels of domain adaptation as long as the test documents are "topic-adaptable", the role of domain adaptation is not that clear yet as the difference between the best topic-adapted system with and without domain-adapted features is relatively small (27.73 vs. 27.57 BLEU on the mixed test set). Therefore, we study the effect of adding domain-adapted features to already topic-adapted systems with smaller topic feature sets, thereby avoiding feature overlap between the systems. Another goal is to measure the contribution of particular topic features and find the best feature combination.

Table 5 shows the results when adding the domain-adapted features to the topic feature sets that do not contain probabilistic features. The upper part of the table shows the performance with single topic features, the lower part shows combinations of two or three topic features. In all experiments, the topic features improve over the unadapted baseline and the additional domain-adapted features improve over the topic-adapted model. Among the single topic features, the **Sim-phrasePair** feature yields the best performance on the mixed test set (27.32) and this trend persists when adding the domain-adapted features (27.53).

However, the best overall performance is achieved with the **Sim-combine** feature in combination with the domain-adapted features. For this setup, both adaptation methods yield a gain of $\sim$0.4 BLEU on the mixed test set, adding up to a total gain of 0.83 as shown at the bottom of the table. The performance of this model on the mixed test set is close to the performance of the best model in Table 4, which indicates that we can achieve good performance with a small set

Model	Mixed	CC	NC	TED
Baseline	26.86	19.61	29.42	31.88
+ TrgUnigrams	27.04	19.86	29.25	32.57
+ DA-TM	**27.50	19.96	29.77	33.34
+ Sim-phrasePair	27.32	20.19	29.31	32.66
+ DA-TM	27.53	20.04	30.05	32.98
+ Sim-targetPhrase	27.21	19.92	29.39	32.58
+ DA-TM	**27.52	19.96	29.94	33.20
+ Sim-targetWord	26.99	19.89	29.16	32.12
+ DA-TM	**27.44	19.91	29.98	32.94
+ Sim-combine	27.29	20.10	29.49	32.60
+ DA-TM	**27.69	**20.13**	**29.90**	**33.37**
+ Sim-combine-loglin	27.18	20.13	29.55	32.34
+ DA-TM	*27.41	19.93	29.86	32.97
+ Sim-combine+trgUnigrams	27.21	20.05	29.36	32.78
+ DA-TM	**27.47	19.87	29.76	33.36
DA gain of best system	+0.40	+0.03	+0.41	+0.77
Gain of best system over baseline	+0.83	+0.52	+0.48	+1.49

Table 5: BLEU scores of smaller topic feature sets with added domain-adapted features (training condition 1). The best system on the mixed test set is marked in bold and its improvements over the topic-adapted system and the baseline are shown at the bottom of the table. **: $p \leq 0.01$, *: $p \leq 0.05$ mark significant improvements over the topic-adapted systems.

of topic-adapted features that encode information not captured by the domain-adapted features. As topic adaptation requires dynamic computation at test time, an architecture where part of the adaptation is done offline, as is the case for domain adaptation, could reduce the computational effort at test time.

7.3 Results for training condition 2

In this section, we evaluate the topic modelling approach on the same test set but with a larger amount of training data for the translation model, language model and topic model (condition 2 in Table 1). The results are shown in Table 6. First we note that the baseline system yields lower overall performance after the addition of the Europarl corpus. This is mostly due to a big hit in performance on the TED test set[8]. The domain-adapted models are able to balance the additional data and the combination of phrase table and language model domain adaptation yields an improvement of $\sim$2 BLEU on the mixed test set compared to the baseline. As for training condition 1, adding topic-adapted features always improves the performance, depending on the amount of previous adaptation. This can be seen most clearly on the TED test set where topic adaptation yields an improvement of 2.15 BLEU over the unadapted baseline, improvements of 1.10 and 0.91 over domain adaptation of the translation or language model, and an improvement of 0.28 over the domain-adapted model with both translation and language model adaptation.

The best overall performance is achieved with a combination of all three adaptation methods, as marked in bold in Table 6. While for CC and NC, the performance of the larger model

[8]The performance on TED in comparison to training condition 1 drops to 30.01 when adding Europarl data to the translation model and to 30.61 when adding Europarl data to the language model, respectively.

Model	Mixed	CC	NC	TED
Baseline	25.74	20.01	29.01	27.82
+ topics	**26.54	20.30	29.55	29.97
	+0.80	+0.29	+0.54	+2.15
DA-TM	26.74	20.13	29.53	30.86
+ topics	**27.21	20.35	29.74	31.96
	+0.47	+0.22	+0.21	+1.10
DA-LM	27.01	20.26	30.48	30.43
+ topics	*27.36	20.34	30.62	31.34
	+0.35	+0.08	+0.14	+0.91
DA-TM+LM	27.70	20.10	30.68	32.70
+ topics	***27.91**	**20.38**	**30.80**	**32.98**
	+0.21	+0.28	+0.12	+0.28
Total gain over baseline	+2.17	+0.37	+1.79	+5.16

Table 6: BLEU scores of unadapted/adapted baseline models (training condition 2) and additional topic-adapted features (Overlap) with their gain over the respective baseline (bottom of each block). The best system on the mixed test set is marked in bold.. **: $p \leq 0.01$, *: $p \leq 0.05$ mark significantly better scores compared to the respective baseline.

is equal or better than the best model in Table 4, the performance on TED falls short of that model by ∼0.6 BLEU. This is likely due to the fact that adding Europarl data is particularly harmful for translating TED documents. Therefore, in future work we will look at combining the adaptation approaches studied here with data selection methods such as the work of Axelrod et al. (2011).

7.4 Qualitative evaluation

In this section, we analyse some concrete output examples that visualise the differences in the translations produced by the different models for training condition 1. Figure 2 shows two input and reference sentences with their translations under the unadapted baseline, the domain-adapted model and the model with both domain-adapted and topic-adapted features[9]. In the first example, the baseline system does not translate the source verb *remontent* appropriately. This is fixed by the domain-adapted model and in addition, the topic-adapted model finds a contextually better translation that matches the reference. In the second example, the domain-adapted model fixes the wrong lexical choice of the baseline model and the topic-adapted model maintains the same translation. Thus, these are examples where domain adaptation is doing most of the adaptation work.

Figure 3 shows two examples where all models make different lexical choices and only the addition of the topic-adapted model yields the correct lexical selection. In these examples, both the baseline and the domain-adapted model choose a translation that corresponds to a different sense of the French source word (*speed*/*bitrate*/*throughput*), while the topic-adapted model selects a translation capturing the same sense as the reference translation (*flow*/*flows*).

The example in Figure 4 shows an incremental improvement from the unadapted model to the domain-adapted model and the topic-adapted model. Here, the domain-adapted model improves slightly over the baseline model by producing a more fluent translation. However, the underlined segments are still translated incorrectly, for example *historique de recherche*

[9]The outputs correspond to the models in the first line and the second block of Table 4.

Input	elles représentent les étendues de l'imagination humaine qui *remontent* à l'aube du temps.
BL	they represent the bodies of the human imagination *back* at the dawn of time.
+DA-TM	they represent the bodies of the human imagination **that** *date* **back to** the dawn of time.
+topics	they represent the bodies of the human imagination **that go back** the dawn of time.
Reference	they represent branches of the human imagination **that go back to** the dawn of time.

Input	ils l'ont fait en *drainant* les terres.
BL	they did in *drawing* the land.
+DA-TM	they did in **draining** the land.
+topics	they did in **draining** the land.
Reference	they did it by **draining** the land.

Figure 2: Comparison of translation output of different models: domain adaptation yields most of the improvement in quality.

Input	le *débit* est en augmentation très rapide.	le *débit* a augmenté.
BL	the *speed* is growing very rapidly.	the *bitrate* has increased.
+DA-TM	the *throughput* is rising very fast.	the *throughput* has increased.
+topics	the **flow** is growing very rapidly.	the **flow** has increased.
Reference	these **flows** are increasing very rapidly.	the **flows** have increased.

Figure 3: Comparison of translation output of different models: topic adaptation yields the translation that captures the correct sense of the French source word *débit*, while domain adaptation does not.

Input	et, si je veux m'éloigner et tout regarder je peux décortiquer mon *historique* peut-être mon *historique de recherche*.
BL	and, if i want to move me and look at everything i can go into my *historical historic* perhaps my *research*.
+DA-TM	and, if i want to get away from and look at everything i can go into my maybe *historical record of my research*.
+topics	and, if i want to get away from it and look at everything i can go into my **history** can be my **search history**.
Reference	and, if i want to step back and look at everything, i can slice and dice my **history** *perhaps by my* **search history**.

Figure 4: Comparison of translation output of different models: here we observe an incremental improvement from domain adaptation to topic adaptation.

is translated as *record of my research*. The topic-adapted model fixes the translations of the underlined segments and finds the correct translations *history* and *search history*.

These examples show that domain and topic adaptation both contribute to the improved translation quality and that depending on the input example, the contribution of one of the two models may be more important. While we cannot draw any definite conclusions about the kind

of improvement each model makes, there seems to be a tendency that the topic-adapted model contributes more to improved lexical choice. We assume that the difference between domain and topic adaptation lies in the granularity of the modelled distributions over translations rather than a clearly defined difference in the level of adaptation, such as style or genre versus topic. However, this would mean that combining models of different granularity implicitly accounts for these levels of adaptation.

8 Conclusion

We have presented an empirical study on the effect of combining domain adaptation and topic adaptation within the same translation system. We have measured the relative benefit of both types of adaptation on a diverse set of test documents and found that the two approaches can be complementary depending on the text type and the amount of overlap between their feature sets. We have shown that the improvements gained by our topic modelling approach hold for domain-adapted models with smaller or larger amounts of training data and are particularly prominent when the training and test domains diverge strongly. We have further shown that the domain of a test document can be predicted accurately by using trained topic models to build domain vector prototypes. Combining domain adaptation, topic adaptation and automatic domain prediction is useful when translating documents from unknown origin and could also help to reduce the load of test time computations while still benefitting from dynamic topic adaptation. Our best combined model yields BLEU improvements of up to 1.67 over an unadapted baseline system and 0.82 over a domain-adapted system, measured on the training condition that yields the stronger baseline system.

Acknowledgements

This work was supported by funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE). We thank Phil Blunsom and the anonymous reviewers for helpful comments.

References

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. *Proceedings of the International Conference on Machine Learning*, pages 25–32.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*.

Banchs, R. E. and Costa-jussà, M. R. (2011). A Semantic Feature for Statistical Machine Translation. In *SSST-5 Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Banerjee, P., Du, J., Li, B., Naskar, S. K., Way, A., and Genabith, J. V. (2010). Combining Multi-Domain Statistical Machine Translation Models using Automatic Classifiers. In *Proceedings of AMTA*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of WMT 2013*.

Cettolo M., G. C. and Federico, M. (2012). Witˆ3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*.

Cui, L., Chen, X., Zhang, D., Liu, S., Li, M., and Zhou, M. (2013). Multi-domain Adaptation for SMT Using Multi-task Learning. In *Proceedings of EMNLP*, pages 1055–1065.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (2012). Topic Models for Dynamic Translation Model Adaptation. In *Proceedings of ACL*.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In *Proceedings of EMNLP*.

Foster, G. and Kuhn, R. (2007). Mixture-Model Adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Gong, Z., Zhang, M., and Guodong, Z. (2011). Cache-based Document-Level Statistical Machine Translation. In *Proceedings of EMNLP 2011*.

Hasler, E., Blunsom, P., Koehn, P., and Haddow, B. (2014a). Dynamic Topic Adaptation for Phrase-based MT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Hasler, E., Haddow, B., and Koehn, P. (2014b). Dynamic Topic Adaptation for SMT using Distributional Profiles. In *Proceedings of the 9th Workshop on Statistical Machine Translation*.

Hewavitharana, S., Mehay, D. N., and Ananthakrishnan, S. (2013). Incremental Topic-Based Translation Model Adaptation for Conversational Spoken Language Translation. In *Proceedings of ACL*, pages 697–701.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for SMT. In *Proceedings of ACL: Demo and poster sessions*.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Ruiz, N. and Federico, M. (2012). MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation. In *Proceedings of IWSLT 2012*.

Sennrich, R. (2012). Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of EACL*.

Sennrich, R., Schwenk, H., and Aransa, W. (2013). A Multi-Domain Translation Model Framework for Statistical Machine Translation. In *Proceedings of ACL*, pages 832–840.

Su, J., Wu, H., Wang, H., Chen, Y., Shi, X., Dong, H., and Liu, Q. (2012). Translation Model Adaptation for Statistical Machine Translation with Monolingual Topic Information. In *Proceedings of ACL*.

Wang, W., Macherey, K., Macherey, W., Och, F., and Xu, P. (2012). Improved Domain Adaptation for Statistical Machine Translation. In *Proceedings of AMTA*.

Xu, J., Deng, Y., Gao, Y., and Ney, H. (2007). Domain Dependent Statistical Machine Translation. In *Proceedings of MT Summit XI*, pages 2–7.