Dynamic Topic Adaptation for Phrase-based MT

Eva Hasler[1], **Phil Blunsom**[2], **Philipp Koehn**[1], **Barry Haddow**[1]
[1]School of Informatics, University of Edinburgh
[2]Dept. of Computer Science, University of Oxford

Abstract

Translating text from diverse sources poses a challenge to current machine translation systems which are rarely adapted to structure beyond corpus level. We explore topic adaptation on a diverse data set and present a new bilingual variant of Latent Dirichlet Allocation to compute topic-adapted, probabilistic phrase translation features. We dynamically infer document-specific translation probabilities for test sets of unknown origin, thereby capturing the effects of document context on phrase translations. We show gains of up to 1.26 BLEU over the baseline and 1.04 over a domain adaptation benchmark. We further provide an analysis of the domain-specific data and show additive gains of our model in combination with other types of topic-adapted features.

1 Introduction

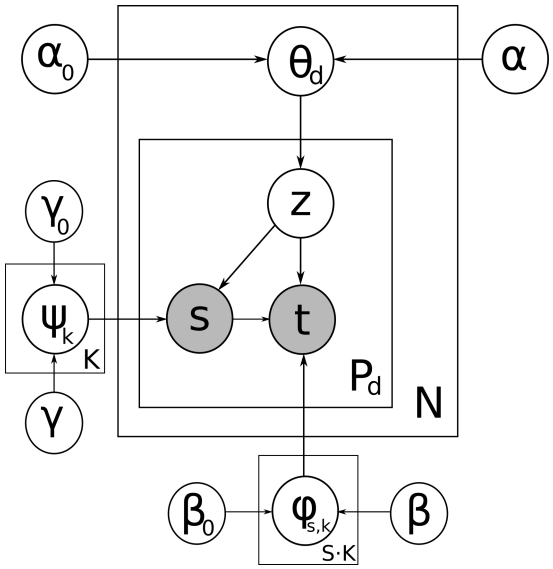
In statistical machine translation (SMT), there has been a lot of interest in trying to incorporate information about the provenance of training examples in order to improve translations for specific target domains. A popular approach are mixture models (Foster and Kuhn, 2007) where each component contains data from a specific genre or domain. Mixture models can be trained for *cross-domain* adaption when the target domain is known or for *dynamic* adaptation when the target domain is inferred from the source text under translation. More recent domain adaptation methods employ corpus or instance weights to promote relevant training examples (Matsoukas et al., 2009; Foster et al., 2010) or do more radical data selection based on language model perplexity (Axelrod et al., 2011). In this work, we are interested in the dynamic adaptation case, which is challenging because we cannot tune our model towards any specific domain.

In previous literature, domains have often been loosely defined in terms of corpora, for example, news texts would be defined as belonging to the news domain, ignoring the specific content of news documents. It is often assumed that the data within a domain is homogeneous in terms of style and vocabulary, though that is not always true in practice. The term *topic* on the other hand can describe the thematic content of a document (e.g. politics, economy, medicine) or a latent cluster in a topic model. Topic modelling for machine translation aims to find a match between thematic context and topic clusters. We view topic adaptation as fine-grained domain adaptation with the implicit assumption that there can be multiple distributions over translations within the same data set. If these distributions overlap, then we expect topic adaptation to help separate them and yield better translations than an unadapted system. Topics can be of varying granularity and are therefore a flexible means to structure data that is not uniform enough to be modelled in its entirety. In recent years there have been several attempts to integrating topical information into SMT either by learning better word alignments (Zhao and Xing, 2006), by adapting translation features cross-domain (Su et al., 2012), or by dynamically adapting lexical weights (Eidelman et al., 2012) or adding sparse topic features (Hasler et al., 2012).

We take a new approach to topic adaptation by estimating probabilistic phrase translation features in a completely Bayesian fashion. The motivation is that automatically identifying topics in the training data can help to select the appropriate translation of a source phrase in the context of a document. By adapting a system to automatically induced topics we do not have to trust data from a given domain to be uniform. We also overcome the problem of defining the level of granularity for domain adaptation. With more and more training data automatically extracted from the web and little knowledge about its content, we believe this is an important area to focus on. Translation of web sites is already a popular application for MT systems and could be helped by dynamic model adaptation. We present results on a mixed data set of the TED corpus, parts of the Commoncrawl corpus which contains crawled web data and parts of the News Commentary corpus which contains

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 328–337,
Gothenburg, Sweden, April 26-30 2014.

Figure 1: Phrasal LDA model for inference on training data.



documents about politics and economics. We believe that the broad range of this data set makes it a suitable testbed for topic adaptation. We focus on translation model adaptation to learn how words and phrases translate in a given document-context without knowing the origin of the document. By learning translations over latent topics and combining several topic-adapted features we achieve improvements of more than 1 BLEU point.

2 Bilingual topic models over phrase pairs

Our model is based on LDA and infers topics as distributions over phrase pairs instead of over words. It is specific to machine translation in that the conditional dependencies between source and target phrases are modelled explicitly, and therefore we refer to it as phrasal LDA. Topic distributions learned on a training corpus are carried over to tuning and test sets by running a modified inference algorithm on the source side text of those sets. Translation probabilities are adapted separately to each source text under translation which makes this a dynamic topic adaptation approach. In the following we explain our approach to topic modelling with the objective of estimating better phrase translation probabilities for data sets that exhibit a heterogeneous structure in terms of vocabulary and style. The advantage from a modelling point of view is that unlike with mixture models, we avoid sparsity problems that would arise if we treated documents or sets of documents as domains and learned separate models for them.

2.1 Latent Dirichlet Allocation (LDA)

LDA is a generative model that learns latent topics in a document collection. In the original formulation, topics are multinomial distributions over words of the vocabulary and each document is assigned a multinomial distribution over topics (Blei et al., 2003). Our goal is to learn topic-dependent phrase translation probabilities and hence we modify this formulation by replacing words with phrase pairs. This is straightforward when both source and target phrases are observed but requires a modified inference approach when only source phrases are observed in an unknown test set. Different from standard LDA and previous uses of LDA for MT, we define a bilingual topic model that learns topic distributions over phrase pairs. This allows us to model the units of interest in a more principled way, without the need to map per-word or per-sentence topics to phrase pairs. Figure 1 shows a graphical representation of the following generative process.

For each of N documents in the collection

1. Choose topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$.
2. Choose the number of phrases pairs P_d in the document, $P_d \sim \textit{Poisson}(\zeta)$.
3. For every position d_i in the document corresponding to a phrase pair $p_{d,i}$ of source and target phrase s_i and t_i[1]:
   (a) Choose a topic $z_{d,i} \sim \text{Multinomial}(\theta_d)$.
   (b) Conditioned on topic $z_{d,i}$, choose a source phrase $s_{d,i} \sim \text{Multinomial}(\psi_{z_{d,i}})$.
   (c) Conditioned on $z_{d,i}$ and $s_{d,i}$, choose target phrase $t_{d,i} \sim \text{Multinomial}(\phi_{s_{d,i},z_{d,i}})$.

α, β and γ are parameters of the Dirichlet distributions, which are asymmetric for $k = 0$. Our inference algorithm is an implementation of collapsed variational Bayes (CVB), with a first-order Gaussian approximation (Teh et al., 2006). It has been shown to be more accurate than standard VB and to converge faster than collapsed Gibbs sampling (Teh et al., 2006; Wang and Blunsom, 2013), with little loss in accuracy. Because we have to do inference over a large number of phrase pairs, CVB is more practical than Gibbs sampling.

2.2 Overview of training strategy

Ultimately, we want to learn translation probabilities for all possible phrase pairs that apply to a given test document during decoding. Therefore, topic modelling operates on phrase pairs as they will be seen during decoding. Given word-aligned parallel corpora from several domains, we extract lists of per-document phrase pairs produced by the extraction algorithm in the Moses toolkit (Koehn et al., 2007) which contain all phrase pairs consistent with the word alignment. We run CVB on the set of all training documents to learn latent topics without providing information about the domains.

[1]Parallel documents are modelled as bags of phrase pairs.

Using the trained model, CVB with modified inference is run on all test documents with the set of possible phrase translations that a decoder would load from a phrase table before decoding. When test inference has finished, we compute adapted translation probabilities at the document-level by marginalising over topics for each phrase pair.

3 Bilingual topic inference

3.1 Inference on training documents

The aim of inference on the training data is to find latent topics in the distributions over phrase pairs in each document.This is done by repeatedly visiting all phrase pair positions in all documents, computing conditional topic probabilities and updating counts. To bias the model to cluster stop word phrases in one topic, we place an asymmetric prior over the hyperparameters[2] as described in (Wallach et al., 2009) to make one of the topics a priori more probable in every document. We use a fixed-point update (Minka, 2012) to update the hyperparameters after every iteration. For CVB the conditional probability of topic $z_{d,i}$ given the current state of all variables except $z_{d,i}$ is

$$P(z_{d,i}=k|\mathbf{z}^{-(d,i)},\mathbf{s},\mathbf{t},d,\alpha,\beta,\gamma) \propto \frac{(\mathbb{E}_{\hat{q}}[n_{.,k,s,t}^{-(d,i)}]+\beta)}{(\mathbb{E}_{\hat{q}}[n_{.,k,s,.}^{-(d,i)}]+T_s\cdot\beta)} \frac{(\mathbb{E}_{\hat{q}}[n_{.,k,s,.}^{-(d,i)}]+\gamma)}{(\mathbb{E}_{\hat{q}}[n_{.,k,.}^{-(d,i)}]+S\cdot\gamma)} \cdot(\mathbb{E}_{\hat{q}}[n_{d,k,.}^{-(d,i)}]+\alpha) \quad (1)$$

where $\mathbf{s}$ and $\mathbf{t}$ are all source and target phrases in the collection. $n_{.,k,s,t}^{-(d,i)}$, $n_{.,k,s,.}^{-(d,i)}$ and $n_{d,k,.}^{-(d,i)}$ are cooccurrence counts of topics with phrase pairs, source phrases and documents respectively. $\mathbb{E}_{\hat{q}}$ is the expectation under the variational posterior and in comparison to Gibbs sampling where the posterior would otherwise look very similar, counts are replaced by their means. $n_{.,k,.}^{-(d,i)}$ is a topic occurrence count, T_s is the number of possible target phrases for a given source phrase and S is the total number of source phrases. By modelling phrase translation probabilities separately as $P(t_i|s_i,z_i=k,..)$ and $P(s_i|z_i=k,..)$, we can put different priors on these distributions. For example, we want a sparse distribution over target phrases for a given source phrase and topic to express our translation preference under each topic. The algorithm stops when the variational posterior has converged for all documents or after a maximum of 100 iterations.

3.2 Inference on tuning and test documents

To compute translation probabilities for tuning and test documents where target phrases are not observed, the variational posterior is adapted as shown in Equation 2

$$P(z_{d,i}=k,t_{i,j}|\mathbf{z}^{-(d,i)},\mathbf{s},\mathbf{t}^{-(d,i)},d,\alpha,\beta,\gamma) \propto \frac{(\mathbb{E}_{\hat{q}}[n_{.,k,s,t_j}^{-(d,i)}]+\beta)}{(\mathbb{E}_{\hat{q}}[n_{.,k,s,.}^{-(d,i)}]+T_s\cdot\beta)} \frac{(\mathbb{E}_{\hat{q}}[n_{.,k,s,.}^{-(d,i)}]+\gamma)}{(\mathbb{E}_{\hat{q}}[n_{.,k,.}^{-(d,i)}]+S\cdot\gamma)} \cdot(\mathbb{E}_{\hat{q}}[n_{d,k,.}^{-(d,i)}]+\alpha) \quad (2)$$

which now computes the joint conditional probability of a topic k and a target phrase $t_{i,j}$, given the source phrase s_i and the test document d. Therefore, the size of the support changes from K to $K\cdot T_s$. While during training inference we compute a distribution over topics for each source-target pair, in test inference we can use the posterior to marginalise out the topics and get a distribution over target phrases for each source phrase.

We use the Moses decoder to produce lists of translation options for each document in the tuning and test sets. These lists comprise all phrase pairs that will enter the search space at decoding time. By default, only 20 target phrases per source phrase are loaded from the phrase table, so in order to allow for new phrase pairs to enter the search space and for translation probabilities to be computed more accurately, we allow for up to 200 target phrases per source. For each source sentence, we consider all possible phrase segmentations and applicable target phrases. Unlike in training, we do not iterate over all phrase pairs in the list but over blocks of up to 200 target phrases for a given source phrase. The algorithm stops when all marginal translation probabilities have converged though in practice we stopped earlier to avoid overfitting.

3.3 Phrase translation probabilities

After topic inference on the tuning and test data, the forward translation probabilities $P(t|s,d)$ are computed. This is done separately for every document d because we are interested in the translation probabilities that depend on the inferred topic proportions for a given document. For every document, we iterate over source positions $p_{d,i}$ and use the current variational posterior to compute $P(t_{i,j}|s_i,d)$ for all possible target phrases by marginalizing over topics:

$$P(t_{i,j}|s_i,d)=\textstyle\sum_k P(z_i=k,t_{i,j}|\mathbf{z}^{-(d,i)},\mathbf{s},\mathbf{t}^{-(d,i)},d)$$

This is straightforward because during test inference the variational posterior is normalised to a distribution over topics and target phrases for a given source phrase. If a source phrase occurs multiple times in the same document, the probabilities are averaged over all occurrences. The inverse translation probabilities can be computed analogously except that in cases where we

[2]Omitted from the following equations for simplicity.

do not have variational posteriors for a given pair of source and target phrases, an approximation is needed. We omit the results here since our experiments so far did not indicate improvements with the inverse features included.

4 More topic-adapted features

Inspired by previous work on topic adaptation for SMT, we add three additional topic-adapted features to our model. All of these features make use of the topic mixtures learned by our bilingual topic model. The first feature is an adapted lexical weight, similar to the features in the work of Eidelman et al. (2012). Our feature is different in that we marginalise over topics to produce a single adapted feature where $v[k]$ is the k^{th} element of a document topic vector for document d and w(t|s,k) is a topic-dependent word translation probability:

$$lex(\bar{t}|\bar{s},d) = \prod_{i}^{|t|} \frac{1}{\{j|(i,j) \in \mathbf{a}\}} \sum_{\forall(i,j)\in\mathbf{a}} \underbrace{\sum_{k} w(t|s,k) \cdot v[k]}_{w(t|s)} \quad (3)$$

The second feature is a target unigram feature similar to the lazy MDI adaptation of Ruiz and Federico (2012). It includes an additional term that measures the relevance of a target word w_i by comparing its document-specific probability P_{doc} to its probability under the asymmetric topic 0:

$$trgUnigrams_t = \prod_{i=1}^{|t|} \underbrace{f(\frac{P_{doc}(w_i)}{P_{baseline}(w_i)})}_{lazy\ MDI} \cdot \underbrace{f(\frac{P_{doc}(w_i)}{P_{topic0}(w_i)})}_{relevance} \quad (4)$$

$$f(x) = \frac{2}{1+\frac{1}{x}}, \quad x > 0 \quad (5)$$

The third feature is a document similarity feature, similar to the semantic feature described by Banchs and Costa-jussà (2011):

$$docSim_t = \max_i (1 - JSD(v_{train\ doc_i}, v_{test\ doc})) \quad (6)$$

where $v_{train\_doc_i}$ and v_{test_doc} are document topic vector of training and test documents. Because topic 0 captures phrase pairs that are common to many documents, we exclude it from the topic vectors before computing similarities.

4.1 Feature combination

We tried integrating the four topic-adapted features separately and in all possible combinations. As we will see in the results section, while all features improve over the baseline in isolation, the adapted translation feature P(t|s,d) is the strongest feature. For the features that have a counterpart in the baseline model (p(t|s,d) and lex(t|s,d)), we experimented with either adding or replacing them in the log-linear model. We found that while adding the features worked well and yielded close to zero weights for their baseline counterparts after tuning, replacing them yielded better results in combination with the other adapted features. We believe the reason could be that fewer phrase table features in total are easier to optimise.

Data	Mixed	CC	NC	TED
Train	354K (6450)	110K	103K	140K
Dev	2453 (39)	818	817	818
Test	5664 (112)	1892	1878	1894

Table 1: Number of sentence pairs and documents (in brackets) in the French-English data sets. The training data has 2.7M English words per domain.

5 Experimental setup

5.1 Data and baselines

Our experiments were carried out on a mixed data set, containing the TED corpus (Cettolo et al., 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. We were guided by two constraints in chosing our data set. 1) the data has document boundaries and the content of each document is assumed to be topically related, 2) there is some degree of topical variation within each data set. In order to compare to domain adaptation approaches, we chose a setup with data from different corpora. We want to abstract away from adaptation effects that concern tuning of length penalties and language models, so we use a mixed tuning set containing data from all three domains and train one language model on the concatenation of (equally sized) target sides of the training data. Word alignments are trained on the concatenation of all training data and fixed for all models.

Our baseline (ALL) is a phrase-based French-English system trained on the concatenation of all parallel data. It was built with the Moses toolkit (Koehn et al., 2007) using the 14 standard core features including a 5gram language model. Translation quality is evaluated on a large test set, 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. As domain-aware benchmark systems, we use the phrase table fill-up method (FILLUP) of Bisazza et al. (2011) which preserves the translation scores of phrases from the IN model and the linear mixture models (LIN-TM) of Sennrich (2012b) (both available in the Moses toolkit). For both systems, we build separate phrase tables for each domain and use a wrapper to decode tuning and test sets with domain-specific tables. Both benchmarks have an advan-

Model	Mixed	CC	NC	TED
IN	26.77	18.76	**29.56**	**32.47**
ALL	26.86	**19.61**	29.42	31.88

Table 2: BLEU of in-domain and baseline models.

Model	Avg JSD	Rank1-diff
Ted-IN vs ALL	0.15	10.8%
CC-IN vs ALL	0.17	18.4%
NC-IN vs ALL	0.13	13.3%

Table 3: Average JSD of IN vs. ALL models. Rank1-diff: % PT entries where preferred translation changes.

tage over our model because they are aware of domain boundaries in the test set. Further, LIN-TM adapts phrase table features in both translation directions while we only adapt the forward features.

Table 2 shows BLEU scores of the baseline system as well as the performance of three in-domain models (IN) tuned under the same conditions. For the IN models, every portion of the test set is decoded with a domain-specific model. Results on the test set are broken down by domain but also reported for the entire test set (mixed). For Ted and NC, the in-domain models perform better than ALL, while for CC the all-domain model improves quite significantly over IN.

5.2 General properties of the data sets

In this section we analyse some internal properties of our three data sets that are relevant for adaptation. All of the scores were computed on the sets of source side tokens of the test set which were limited to contain content words (nouns, verbs, adjectives and adverbs). The test set was tagged with the French TreeTagger (Schmid, 1994). The top of Table 3 shows the average Jensen-Shannon divergence (using $\log_2$, JSD $\in [0,1]$) of each in-domain model in comparison to the all-domain model, which is an indicator of how much the distributions in the IN model change when adding out-of-domain data. Likewise, Rank1-diff gives the percentage of word tokens in the test set where the preferred translation according to $p(e|f)$ changes between IN and ALL. These are the words that are most affected by adding data to the IN model. Both numbers show that for Commoncrawl the IN and ALL models differ more than in the other two data sets. According to the JS divergence between NC-IN and ALL, translation distibutions in the NC phrase table are most similar to the ALL phrase table. Table 4 shows the average JSD for each IN model compared to a model trained on half of its in-domain data. This score gives an idea of how diverse a data set is, measured by comparing distributions over translations for source words in the test set. According to this score, Commoncrawl is the most diverse data set and Ted the most uniform. Note however, that these divergence scores do not provide information about the relative quality of the systems under comparison. For CC, the ALL model yields a much higher BLEU score than the IN model and it is likely that this is due to noisy data in the CC corpus. In this case, the high divergence is likely to mean that distributions are corrected by out-of-domain data rather than being shifted away from in-domain distributions.

Model	Avg JSD
Ted-half vs Ted-full	0.07
CC-half vs CC-full	0.17
NC-half vs NC-full	0.09

Table 4: Average JSD of in-domain models trained on half vs. all of the data.

5.3 Topic-dependent decoding

The phrase translation probabilities and additional features described in the last two sections are used as features in the log-linear translation model in addition to the baseline translation features. When combining all four adapted features, we replace $P(t|s)$ and $lex(t|s)$ by their adapted counterparts. We construct separate phrase tables for each document in the development and test sets and use a wrapper around the decoder to ensure that each input document is paired with a configuration file pointing to its document-specific translation table. Documents are decoded in sequence so that only one phrase table needs to be loaded at a time. Using the wrapped decoder we can run parameter optimisation (PRO) in the usual way to get one set of tuned weights for all test documents.

6 Results

In this section we present experimental results with phrasal LDA. We show BLEU scores in comparison to a baseline system and two domain-aware benchmark systems. We also evaluate the adapted translation distributions by looking at translation probabilities under specific topics and inspect translations of ambiguous source words.

6.1 Analyis of bilingual topic models

We experimented with different numbers of topics for phrasal LDA. The diagrams in Figure 2 shows blocks of training and test documents in each of the three domains for a model with 20 topics. Darker shading means that documents have a higher proportion of a particular topic in their document-topic distribution. The first topic is the one that was affected by the asymmetric prior and inspecting its most probable phrase pairs showed that it had 'collected' a large number of stop word phrases. This explains why it is the topic that is most shared across documents and domains.

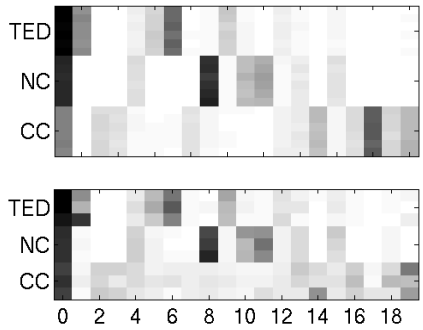


Figure 2: Document-topic distributions for training (top) and test (bottom) documents, grouped by domain and averaged into blocks for visualisation.

Topic 8	**Topic 11**
européenne → european	crise → crisis
politiques → political	taux → rate
politique → policy	financière →financial
intérêts → interests	monétaire → monetary
Topic 14	**Topic 19**
hôtel → hotel	web → web
plage → beach	utiliser → use
situé → located	logiciel → software
chambres → bedrooms	données → data

Figure 3: Frequent phrase pairs in learned topics.

There is quite a clear horizontal separation between documents of different domains, for example, topics 6, 8, 19 occur mostly in Ted, NC and CC documents respectively. The overall structure is very similar between training (top) and test (bottom) documents, which shows that test inference was successful in carrying over the information learned on training documents. There is also some degree of topic sharing across domains, for example topics 4 and 15 occur in documents of all three domains. Figure 3 shows examples of latent topics found during inference on the training data. Topic 8 and 11 seem to be about politics and economy and occur frequently in documents from the NC corpus. Topic 14 contains phrases related to hotels and topic 19 is about web and software, both frequent themes in the CC corpus.

6.2 Comparison according to BLEU

In Table 5 we compare our topic-adapted features when added separately to the baseline phrase table. The inclusion of each feature improves over the concatenation baseline but the combination of all four features gives the best overall results. Though the relative performance differs slightly for each domain portion in the test set, overall the adapted lexical weight is the weakest feature and the adapted translation probability is the strongest feature. We also performed feature ablation tests and found that no combination of features was superior to combining all four features. This confirms that the gains of each feature lead to additive improvements in the combined model.

In Table 6 we compare topic-adapted models with varying numbers of topics to the concatenation baseline. We see a consistent gain on all domains when increasing the number of topics from three to five and ten topics. This is evidence that the number of domain labels is in fact smaller than the number of underlying topics. The optimal number of latent topics varies for each domain and reflects our insights from section 5.2. The CC domain was shown to be the most diverse and the best performance on the CC portion of the test set is achieved with 100 topics. Likewise, the TED domain was shown to be least diverse and here the best performance is achieved with only 10 topics. The best performance on the entire test set is achieved with 50 topics, which is also the optimal number of topics for the NC domain. The botton row of the table indicates the relative improvement of the best topic-adapted model per domain over the ALL model. Using all four topic-adapted features yields an improvement of 0.81 BLEU on the mixed test set. The highest improvement on a given domain is achieved for TED with an increase of 1.26 BLEU. The smallest improvement is measured on the NC domain. This is in line with the observation that distributions in the NC in-domain table are most similar to the ALL table, therefore we would expect the smallest improvement for domain or topic adaptation. We used bootstrap resampling (Koehn, 2004) to measure significance on the mixed test set and marked all statistically significant results compared to the respective baselines with asterisk (*: $p \leq 0.01$).

To demonstrate the benefit of topic adaptation over more standard domain adaptation approaches for a diverse data set, we show the performance

Model	**Mixed**	**CC**	**NC**	**TED**
lex(e|f,d)	26.99	19.93	29.34	32.19
trgUnigrams	27.15	19.90	29.54	32.50
docSim	27.22	20.11	29.63	32.40
p(e|f,d)	27.31	20.23	29.52	32.58
All features	**27.67**	**20.40**	**30.04**	**33.08**

Table 5: BLEU scores of pLDA features (50 topics), separately and combined.

Model	**Mixed**	**CC**	**NC**	**TED**
ALL	26.86	19.61	29.42	31.88
3 topics	26.95	19.83	29.46	32.02
5 topics	*27.48	19.98	29.94	33.04
10 topics	*27.65	20.34	29.99	**33.14**
20 topics	*27.63	20.39	29.93	33.09
50 topics	***27.67**	20.40	**30.04**	33.08
100 topics	*27.65	**20.54**	30.00	32.90
>ALL	+0.81	+0.93	+0.62	+1.26

Table 6: BLEU scores of baseline and topic-adapted systems (pLDA) with all 4 features and largest improvements over baseline.

Model	Mixed	CC	NC	TED
FILLUP	27.12	19.36	29.78	32.71
LIN-TM	27.24	19.61	29.87	32.73
pLDA	***27.67**	**20.40**	**30.04**	**33.08**
>FILLUP	+0.55	+1.04	+0.26	+0.37
>LIN-TM	+0.43	+0.79	+0.17	+0.35

Table 7: Comparison of best pLDA system with two domain-aware benchmark systems.

Model	Mixed	CC	NC	TED
LIN-LM				
+ ALL	27.16	19.71	29.77	32.46
+ FILLUP	27.20	19.37	29.84	32.90
+ LIN-TM	27.34	19.59	29.92	33.02
+ pLDA	***27.84**	**20.48**	**30.03**	**33.57**
>ALL	+0.68	+0.77	+0.26	+1.11

Table 8: Combination of all models with additional LM adaptation (pLDA: 50 topics).

of two state-of-the-art domain-adapted systems in Table 7. Both FILLUP and LIN-TM improve over the ALL model on the mixed test set, by 0.26 and 0.38 BLEU respectively. The largest improvement is on TED while on the CC domain, FILLUP decreases in performance and LIN-TM yields no improvement either. This shows that relying on in-domain distributions for adaptation to a noisy and diverse domain like CC is problematic. The pLDA model yields the largest improvement over the domain-adapted systems on the CC test set, with in increase of 1.04 BLEU over FILLUP and 0.79 over LIN-TM. The improvements on the other two domains are smaller but consistent.

We also compare the best model from Table 6 to all other models in combination with linearly interpolated language models (LIN-LM), interpolated separately for each domain. Though the improvements are slightly smaller than without adapted language models, there is still a gain over the concatenation baseline of 0.68 BLEU on the mixed test set and similar improvements to before over the benchmarks (on TED the improvements are actually even larger). Thus, we have shown that topic-adaptation is effective for test sets of diverse documents and that we can achieve substantial improvements even in comparison with domain-adapted translation and language models.

6.3 Properties of adapted distributions and topic-specific translations

The first column of Table 9 shows the average entropy of phrase table entries in the adapted models according to $p(t|s,d)$ versus the all-domain model, computed over source tokens in the test set that are content words. The entropy decreases in the adapted tables in all cases which is an indicator that the distributions over translations of content

Set	Model	Avg entropy	Avg perplexity
CC	pLDA	**3.74**	**9.21**
	ALL	3.99	10.13
NC	pLDA	**3.42**	**6.96**
	ALL	3.82	7.51
TED	pLDA	**3.33**	**9.17**
	ALL	4.00	9.71

Table 9: Average entropy of translation distributions and test set perplexity of the adapted model.

régime		
topic 6	diet = 0.79	diet aids = 0.04
topic 8	regime* = 0.82	rule = 0.05
topic 19	restrictions = 0.53	diplomats = 0.10
noyau		
topic 9	nucleus* = 0.89	core = 0.01
topic 11	core* = 0.93	inner = 0.03
topic 19	kernel = 0.58	core = 0.11
démon		
topic 6	devil = 0.89	demon = 0.07
topic 8	demon* = 0.98	devil = 0.01
topic 19	daemon = 0.95	demon = 0.04

Table 10: The two most probable translations of *régime*, *noyau* and *démon* and probabilities under different latent topics (*: preferred by ALL).

words have become more peaked. The second column shows the average perplexity of target tokens in the test set which is a measure of how likely a model is to produce words in the reference translation. We use the alignment information between source and reference and therefore limit our analysis to pairs of aligned words, but nevertheless this shows that the adapted translation distributions model the test set distributions better than the baseline model. Therefore, the adapted distributions are not just more peaked but also more often peaked towards the correct translation.

Table 10 shows examples of ambiguous French words that have different preferred translations depending on the latent topic. The word *régime* can be translated as *diet*, *regime* and *restrictions* and the model has learned that the probability over translations changes when moving from one topic to another (preferred translations under the ALL model are marked with *). For example, the translation to *diet* is most probable under topic 6 and the translation to *regime* which would occur in a political context is most probable under topic 8. Topic 6 is most prominent among Ted documents while topic 8 is found most frequently in News Commentary documents which have a high percentage of politically related text. The French word *noyau* can be translated to *nucleus* (physics), *core* (generic) and *kernel* (IT) among other translations and the topics that exhibit these preferred translations can be attributed to Ted (which contains many talks about physics), NC and CC (with

Src: "il suffit d'éjecter le *noyau* et d'en insérer un autre, comme ce qu'on fait pour le clônage."
BL: "it is the **nucleus** eject and insert another, like what we do to the clônage."
pLDA: "he just eject the **nucleus** and insert another, like what we do to the clônage." (nucleus = **0.77**)
Ref: "you can just pop out the **nucleus** and pop in another one, and that's what you've all heard about with cloning."

Src: "pourtant ceci obligerait les contribuables des pays de ce *noyau* à fournir du capital au sud"
BL: "but this would force western taxpayers to provide the nucleus of capital in the south"
pLDA: "but this would force western taxpayers to provide the **core** of capital in the south" (core = **0.78**)
Ref: "but this would unfairly force taxpayers in the **core** countries to provide capital to the south"

Src: "le *noyau* contient de nombreux pilotes, afin de fonctionner chez la plupart des utilisateurs."
BL: "the nucleus contains many drivers, in order to work for most users."
pLDA: "the **kernel** contains many drivers, to work for most users." (kernel = **0.53**)
Ref: "the precompiled **kernel** includes a lot of drivers, in order to work for most users."

Figure 4: pLDA correctly translates *noyau* in test docs from Ted, NC and CC (adapted probabilities in brackets). The baseline (nucleus = 0.27, core = 0.27, kernel = 0.23) translates all instances to *nucleus*.

many IT-related documents). The last example, *démon*, has three frequent translations in English: *devil*, *demon* and *daemon*. The last translation refers to a computer process and would occur in an IT context. The topic-phrase probabilities reveal that its mostly likely translation as *daemon* occurs under topic 19 which clusters IT-related phrase pairs and is frequent in the CC corpus. These examples show that our model can disambiguate phrase translations using latent topics.

As another motivating example, in Figure 4 we compare the output of our adapted models to the output produced by the all-domain baseline for the word *noyau* from Table 10. While the ALL baseline translates each instance of *noyau* to *nucleus*, the adapted model translates each instance differently depending on the inferred topic mixtures for each document and always matches the reference translation. The probabilities in brackets show that the chosen translations were indeed the most likely under the respective adapted model. While the ALL model has a flat distribution over possible translations, the adapted models are peaked towards the correct translation. This shows that topic-specific translation probabilities are necessary when the translation of a word shifts between topics or domains and that peaked, adapted distributions can lead to more correct translations.

7 Related work

There has been a lot of previous work using topic information for SMT, most of it using monolingual topic models. For example, Gong and Zhou (2011) use the topical relevance of a target phrase, computed using a mapping between source and target side topics, as an additional feature in decoding. Axelrod et al. (2012) build topic-specific translation models from the TED corpus and select topic-relevant data from the UN corpus to improve coverage. Su et al. (2012) perform phrase table adaptation in a setting where only monolingual in-domain data and parallel out-of-domain data are available. Eidelman et al. (2012) use topic-dependent lexical weights as features in the translation model, which is similar to our work in that topic features are tuned towards usefulness of topic information and not towards a target domain. Hewavitharana et al. (2013) perform dynamic adaptation with monolingual topics, encoding topic similarity between a conversation and training documents in an additional feature. This is similar to the work of Banchs and Costa-jussà (2011), both of which inspired our document similarity feature. Also related is the work of Sennrich (2012a) who explore mixture-modelling on unsupervised clusters for domain adaptation and Chen et al. (2013) who compute phrase pair features from vector space representations that capture domain similarity to a development set. Both are cross-domain adaptation approaches, though. Instances of multilingual topic models outside the field of MT include Boyd-Graber and Blei (2009; Boyd-Graber and Resnik (2010) who learn cross-lingual topic correspondences (but do not learn conditional distributions like our model does). In terms of model structure, our model is similar to BiTAM (Zhao and Xing, 2006) which is an LDA-style model to learn topic-based word alignments. The work of Carpuat and Wu (2007) is similar to ours in spirit, but they predict the most probable translation in a context at the token level while our adaptation operates at the type level of a document.

8 Conclusion

We have presented a novel bilingual topic model based on LDA and applied it to the task of translation model adaptation on a diverse French-English data set. Our model infers topic distributions over phrase pairs to compute document-specific translation probabilities and performs dynamic adaptation on test documents of unknown origin. We have shown that our model outperforms a concatenation baseline and two domain-adapted benchmark systems with BLEU gains of up to 1.26 on domain-specific test set portions and 0.81 overall. We have also shown that a combination of topic-adapted features performs better than each feature in isolation and that these gains are additive. An analysis of the data revealed that topic adaptation compares most favourably to domain adaptation when the domain in question is rather diverse.

Acknowledgements

This work was supported by funding from the Scottish Informatics and Computer Science Alliance (Eva Hasler) and funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE) and grant agreement 288769 (ACCEPT). Thanks to Chris Dyer for an initial discussion about the phrasal LDA model.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, Li Deng, Alex Acero, and Mei-Yuh Hwang. 2012. New methods and evaluation experiments on translating TED talks in the IWSLT benchmark. In *Proceedings of ICASSP*. IEEE.

Rafael E. Banchs and Marta R. Costa-jussà. 2011. A semantic feature for statistical machine translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5. Association for Computational Linguistics.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proceedings of IWSLT*.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *JMLR*.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of WMT 2013. Association for Computational Linguistics.

Jordan Boyd-Graber and David Blei. 2009. Multilingual Topic Models for Unaligned Text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. How phrase sense disambiguation outperforms word sense disambiguation for SMT. In *International Conference on Theoretical and Methodological Issues in MT*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*.

Boxing Chen, Roland Kuhn, and George Foster. 2013. Vector space model for adaptation in SMT. In *Proceedings of ACL*. Association for Computational Linguistics.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of ACL*. Association for Computational Linguistics.

G. Foster and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *Proceedings of WMT*. Association for Computational Linguistics.

G. Foster, C. Goutte, and R. Kuhn. 2010. Discriminative instance weighting for domain adaptation in SMT. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Zhengxian Gong and Guodong Zhou. 2011. Employing topic modeling for SMT. In *Proceedings of IEEE (CSAE)*, volume 4.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2012. Sparse lexicalised features and topic adaptation for SMT. In *Proceedings of IWSLT*.

S. Hewavitharana, D. Mehay, S. Ananthakrishnan, and P. Natarajan. 2013. Incremental topic-based TM adaptation for conversational SLT. In *Proceedings of ACL*. Association for Computational Linguistics.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for SMT. In *ACL 2007: Demo and poster sessions*. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*. Association for Computational Linguistics.

S. Matsoukas, A. Rosti, and B. Zhang. 2009. Discriminative corpus weight estimation for MT. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Thomas P Minka. 2012. Estimating a Dirichlet distribution. Technical report.

Nick Ruiz and Marcello Federico. 2012. MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation. In *Proceedings of IWSLT*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Rico Sennrich. 2012a. Mixture-modeling with unsupervised clusters for domain adaptation in SMT. In *Proceedings of EAMT*.

Rico Sennrich. 2012b. Perplexity Minimization for Translation Model Domain Adaptation in SMT. In *Proceedings of EACL*. Association for Computational Linguistics.

J. Su, H. Wu, H. Wang, Y. Chen, X. Shi, H. Dong, and Q. Liu. 2012. Translation model adaptation for SMT with monolingual topic information. In *Proceedings of ACL*. Association for Computational Linguistics.

Yee Whye Teh, David Newman, and Max Welling. 2006. A collapsed variational Bayesian inference algorithm for LDA. In *Proceedings of NIPS*.

Hanna M. Wallach, David M. Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of NIPS*.

Pengyu Wang and Phil Blunsom. 2013. Collapsed variational Bayesian inference for Hidden Markov Models. In *AISTATS*, volume 31 of *JMLR Proceedings*, pages 599–607.

Bing Zhao and Eric P. Xing. 2006. Bilingual topic admixture models for word alignment. In *Proceedings of ACL*. Association for Computational Linguistics.