Corpus	Sentences	Tokens	
		En	Zh
FBIS	269K	10.3M	7.9M
NIST	1.6M	44.4M	40.4M

Table 1: Corpus statistics

2010) as our decoder, and tuned the parameters of the system to optimize BLEU (Papineni et al., 2002) on the NIST MT06 tuning corpus using the Margin Infused Relaxed Algorithm (MIRA) (Crammer et al., 2006; Eidelman, 2012). Topic modeling was performed with Mallet (Mccallum, 2002), a standard implementation of LDA, using a Chinese stoplist and setting the per-document Dirichlet parameter $\alpha = 0.01$. This setting of was chosen to encourage sparse topic assignments, which make induced subdomains consistent within a document.

Results Results for both settings are shown in Table 2. GTM models the latent topics at the document level, while LTM models each sentence as a separate document. To evaluate the effect topic granularity would have on translation, we varied the number of latent topics in each model to be 5, 10, and 20. On FBIS, we can see that both models achieve moderate but consistent gains over the baseline on both BLEU and TER. The best model, LTM-10, achieves a gain of about 0.5 and 0.6 BLEU and 2 TER. Although the performance on BLEU for both the 20 topic models LTM-20 and GTM-20 is suboptimal, the TER improvement is better. Interestingly, the difference in translation quality between capturing document coherence in GTM and modeling purely on the sentence level is not substantial.[5] In fact, the opposite is true, with the LTM models achieving better performance.[6]

On the NIST corpus, LTM-10 again achieves the best gain of approximately 1 BLEU and up to 3 TER. LTM performs on par with or better than GTM, and provides significant gains even in the NIST data setting, showing that this method can be effectively applied directly on the sentence level to large training corpora which have no document markings. Depending on the diversity of training corpus, a varying number of underlying topics may be appropriate. However, in both settings, 10 topics performed best.

Model	MT03		MT05	
	↑BLEU	↓TER	↑BLEU	↓TER
BL	28.72	65.96	27.71	67.58
GTM-5	28.95^{ns}	65.45	27.98^{ns}	67.38^{ns}
GTM-10	29.22	64.47	**28.19**	66.15
GTM-20	29.19	**63.41**	28.00^{ns}	**64.89**
LTM-5	29.23	64.57	**28.19**	66.30
LTM-10	**29.29**	63.98	**28.18**	65.56
LTM-20	29.09^{ns}	63.57	27.90^{ns}	65.17
Model	MT03		MT05	
	↑BLEU	↓TER	↑BLEU	↓TER
BL	34.31	61.14	30.63	65.10
MERT	34.60	60.66	30.53	64.56
LTM-5	35.21	59.48	31.47	62.34
LTM-10	**35.32**	**59.16**	**31.56**	**62.01**
LTM-20	33.90^{ns}	60.89^{ns}	30.12^{ns}	63.87

Table 2: Performance using FBIS training corpus (top) and NIST corpus (bottom). Improvements are significant at the p <0.05 level, except where indicated (^{ns}).

4 Discussion and Conclusion

Applying SMT to new domains requires techniques to inform our algorithms how best to adapt. This paper extended the usual notion of domains to finer-grained topic distributions induced in an unsupervised fashion. We show that incorporating lexical weighting features conditioned on soft domain membership directly into our model is an effective strategy for dynamically biasing SMT towards relevant translations, as evidenced by significant performance gains. This method presents several advantages over existing approaches. We can construct a topic model once on the training data, and use it infer topics on any test set to adapt the translation model. We can also incorporate large quantities of additional data (whether parallel or not) in the source language to infer better topics without relying on collection or genre annotations. Multilingual topic models (Boyd-Graber and Resnik, 2010) would provide a technique to use data from multiple languages to ensure consistent topics.

[5] An avenue of future work would condition the sentence topic distribution on a document distribution over topics (Teh et al., 2006).

[6] As an empirical validation of our earlier intuition regarding feature representation, presenting the features in the form of F_1 caused the performance to remain virtually unchanged from the baseline model.