Vladimir Eidelman, **Jordan Boyd-Graber**, and Philip Resnik. **Topic Models for Dynamic Translation Model Adaptation**. *Association for Computational Linguistics*, 2012.

```
    @inproceedings{Eidelman:Boyd-Graber:Resnik-2012,
Title = {Topic Models for Dynamic Translation Model Adaptation},
Booktitle = {Association for Computational Linguistics},
Author = {Vladimir Eidelman and Jordan Boyd-Graber and Philip Resnik},
Year = {2012},
Location = {Jeju, South Korea},
}
```

Topic Models for Dynamic Translation Model Adaptation

Vladimir Eidelman
Computer Science
and UMIACS
University of Maryland
College Park, MD
vlad@umiacs.umd.edu

Jordan Boyd-Graber
iSchool
and UMIACS
University of Maryland
College Park, MD
jbg@umiacs.umd.edu

Philip Resnik
Linguistics
and UMIACS
University of Maryland
College Park, MD
resnik@umd.edu

Abstract

We propose an approach that biases machine translation systems toward relevant translations based on topic-specific contexts, where topics are induced in an unsupervised way using topic models; this can be thought of as inducing subcorpora for adaptation without any human annotation. We use these topic distributions to compute topic-dependent lexical weighting probabilities and directly incorporate them into our translation model as features. Conditioning lexical probabilities on the topic biases translations toward topic-relevant output, resulting in significant improvements of up to 1 BLEU and 3 TER on Chinese to English translation over a strong baseline.

1 Introduction

The performance of a statistical machine translation (SMT) system on a translation task depends largely on the suitability of the available parallel training data. Domains (e.g., newswire vs. blogs) may vary widely in their lexical choices and stylistic preferences, and what may be preferable in a general setting, or in one domain, is not necessarily preferable in another domain. Indeed, sometimes the domain can change the meaning of a phrase entirely.

In a food related context, the Chinese sentence "粉丝很多 " ("fěnsī hěnduō") would mean "They have a lot of vermicelli"; however, in an informal Internet conversation, this sentence would mean "They have a lot of fans". Without the broader context, it is impossible to determine the correct translation in otherwise identical sentences.

This problem has led to a substantial amount of recent work in trying to bias, or adapt, the translation model (TM) toward particular domains of interest (Axelrod et al., 2011; Foster et al., 2010; Snover et al., 2008).[1] The intuition behind TM adaptation is to increase the likelihood of selecting relevant phrases for translation. Matsoukas et al. (2009) introduced assigning a pair of binary features to each training sentence, indicating sentences' *genre and collection* as a way to capture domains. They then learn a mapping from these features to sentence weights, use the sentence weights to bias the model probability estimates and subsequently learn the model weights. As sentence weights were found to be most beneficial for lexical weighting, Chiang et al. (2011) extends the same notion of conditioning on provenance (i.e., the origin of the text) by removing the separate mapping step, directly optimizing the weight of the genre and collection features by computing a separate word translation table for each feature, estimated from only those sentences that comprise that genre or collection.

The common thread throughout prior work is the concept of a *domain*. A domain is typically a hard constraint that is externally imposed and hand labeled, such as genre or corpus collection. For example, a sentence either comes from newswire, or weblog, but not both. However, this poses several problems. First, since a sentence contributes its counts only to the translation table for the source it came from, many word pairs will be unobserved for a given table. This sparsity requires smoothing. Second, we may not know the (sub)corpora our training

[1] Language model adaptation is also prevalent but is not the focus of this work.

data come from; and even if we do, "subcorpus" may not be the most useful notion of domain for better translations.

We take a finer-grained, flexible, unsupervised approach for lexical weighting by domain. We induce unsupervised domains from large corpora, and we incorporate soft, probabilistic domain membership into a translation model. Unsupervised modeling of the training data produces naturally occurring subcorpora, generalizing beyond corpus and genre. Depending on the model used to select subcorpora, we can bias our translation toward any arbitrary distinction. This reduces the problem to identifying what automatically defined subsets of the training corpus may be beneficial for translation.

In this work, we consider the underlying *latent topics* of the documents (Blei et al., 2003). Topic modeling has received some use in SMT, for instance Bilingual LSA adaptation (Tam et al., 2007), and the BiTAM model (Zhao and Xing, 2006), which uses a bilingual topic model for learning alignment. In our case, by building a topic distribution for the source side of the training data, we abstract the notion of domain to include automatically derived subcorpora with probabilistic membership. This topic model infers the topic distribution of a test set and biases sentence translations to appropriate topics. We accomplish this by introducing topic dependent lexical probabilities directly as features in the translation model, and interpolating them log-linearly with our other features, thus allowing us to discriminatively optimize their weights on an arbitrary objective function. Incorporating these features into our hierarchical phrase-based translation system significantly improved translation performance, by up to 1 BLEU and 3 TER over a strong Chinese to English baseline.

2 Model Description

Lexical Weighting Lexical weighting features estimate the quality of a phrase pair by combining the lexical translation probabilities of the words in the phrase[2] (Koehn et al., 2003). Lexical conditional probabilities $p(e|f)$ are obtained with maximum likelihood estimates from relative frequencies $c(f,e)/\sum_e c(f,e)$. Phrase pair probabilities $p(\overline{e}|\overline{f})$ are computed from these as described in Koehn et al. (2003).

Chiang et al. (2011) showed that is it beneficial to condition the lexical weighting features on provenance by assigning each sentence pair a set of features, $f_s(\overline{e}|\overline{f})$, one for each domain s, which compute a new word translation table $p_s(e|f)$ estimated from only those sentences which belong to s: $c_s(f,e)/\sum_e c_s(f,e)$, where $c_s(\cdot)$ is the number of occurrences of the word pair in s.

Topic Modeling for MT We extend provenance to cover a set of automatically generated topics z_n. Given a parallel training corpus T composed of documents d_i, we build a source side topic model over T, which provides a topic distribution $p(z_n|d_i)$ for $z_n = \{1, \ldots, K\}$ over each document, using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Then, we assign $p(z_n|d_i)$ to be the topic distribution for every sentence $x_j \in d_i$, thus enforcing topic sharing across sentence pairs in the same document instead of treating them as unrelated. Computing the topic distribution over a document and assigning it to the sentences serves to tie the sentences together in the document context.

To obtain the lexical probability conditioned on topic distribution, we first compute the expected count $e_{z_n}(e,f)$ of a word pair under topic z_n:

$$e_{z_n}(e,f) = \sum_{d_i \in T} p(z_n|d_i) \sum_{x_j \in d_i} c_j(e,f) \quad (1)$$

where $c_j(\cdot)$ denotes the number of occurrences of the word pair in sentence x_j, and then compute:

$$p_{z_n}(e|f) = \frac{e_{z_n}(e,f)}{\sum_e e_{z_n}(e,f)} \quad (2)$$

Thus, we will introduce $2{\cdot}K$ new word translation tables, one for each $p_{z_n}(e|f)$ and $p_{z_n}(f|e)$, and as many new corresponding features $f_{z_n}(\overline{e}|\overline{f})$, $f_{z_n}(\overline{f}|\overline{e})$. The actual feature values we compute will depend on the topic distribution of the document we are translating. For a test document V, we infer topic assignments on V, $p(z_n|V)$, keeping the topics found from T fixed. The feature value then becomes $f_{z_n}(\overline{e}|\overline{f}) = -\log\left\{p_{z_n}(\overline{e}|\overline{f}) \cdot p(z_n|V)\right\}$, a combination of the topic dependent lexical weight and the

[2]For hierarchical systems, these correspond to translation rules.

topic distribution of the sentence from which we are extracting the phrase. To optimize the weights of these features we combine them in our linear model with the other features when computing the model score for each phrase pair[3]:

$$\underbrace{\sum_{p} \lambda_p h_p(e,f)}_{\text{unadapted features}} + \underbrace{\sum_{z_n} \lambda_{z_n} f_{z_n}(\overline{e}|\overline{f})}_{\text{adapted features}} \quad (3)$$

Combining the topic conditioned word translation table $p_{z_n}(e|f)$ computed from the training corpus with the topic distribution $p(z_n|V)$ of the test sentence being translated provides a probability on how relevant that translation table is to the sentence. This allows us to bias the translation toward the topic of the sentence. For example, if topic k is dominant in T, $p_k(\overline{e}|\overline{f})$ may be quite large, but if $p(k|V)$ is very small, then we should steer away from this phrase pair and select a competing phrase pair which may have a lower probability in T, but which is more relevant to the test sentence at hand.

In many cases, document delineations may not be readily available for the training corpus. Furthermore, a document may be too broad, covering too many disparate topics, to effectively bias the weights on a phrase level. For this case, we also propose a local LDA model (LTM), which treats each sentence as a separate document.

While Chiang et al. (2011) has to *explicitly* smooth the resulting $p_s(e|f)$, since many word pairs will be unseen for a given domain s, we are already performing an *implicit* form of smoothing (when computing the expected counts), since each document has a distribution over all topics, and therefore we have some probability of observing each word pair in every topic.

Feature Representation After obtaining the topic conditional features, there are two ways to present them to the model. They could answer the question F_1: What is the probability under topic 1, topic 2, etc., or F_2: What is the probability under the most probable topic, second most, etc.

A model using F_1 learns whether a *specific* topic is useful for translation, i.e., feature f_1 would be $f_1 := p_{z=1}(\overline{e}|\overline{f}) \cdot p(z = 1|V)$. With F_2, we are learning how useful knowledge of the topic distribution is, i.e., $f_1 := p_{(\arg\max_{z_n}(p(z_n|V)))}(\overline{e}|\overline{f}) \cdot p(\arg\max_{z_n}(p(z_n|V))|V)$.

Using F_1, if we restrict our topics to have a one-to-one mapping with genre/collection[4] we see that our method fully recovers Chiang (2011).

F_1 is appropriate for *cross-domain* adaptation when we have advance knowledge that the distribution of the tuning data will match the test data, as in Chiang (2011), where they tune and test on web. In general, we may not know what our data will be, so this will overfit the tuning set.

F_2, however, is intuitively what we want, since we do not want to bias our system toward a specific distribution, but rather learn to utilize information from *any* topic distribution if it helps us create topic relevant translations. F_2 is useful for *dynamic* adaptation, where the adapted feature weight changes based on the source sentence.

Thus, F_2 is the approach we use in our work, which allows us to tune our system weights toward having topic information be useful, not toward a specific distribution.

3 Experiments

Setup To evaluate our approach, we performed experiments on Chinese to English MT in two settings. First, we use the FBIS corpus as our training bitext. Since FBIS has document delineations, we compare local topic modeling (LTM) with modeling at the document level (GTM). The second setting uses the non-UN and non-HK Hansards portions of the NIST training corpora with LTM only. Table 1 summarizes the data statistics. For both settings, the data were lowercased, tokenized and aligned using GIZA++ (Och and Ney, 2003) to obtain bidirectional alignments, which were symmetrized using the `grow-diag-final-and` method (Koehn et al., 2003). The Chinese data were segmented using the Stanford segmenter. We trained a trigram LM on the English side of the corpus with an additional 150M words randomly selected from the non-NYT and non-LAT portions of the Gigaword v4 corpus using modified Kneser-Ney smoothing (Chen and Goodman, 1996). We used `cdec` (Dyer et al.,

[3]The unadapted lexical weight $p(\overline{e}|\overline{f})$ is included in the model features.

[4]By having as many topics as genres/collections and setting $p(z_n|d_i)$ to 1 for every sentence in the collection and 0 to everything else.

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.

Acknowledgments

Vladimir Eidelman is supported by a National Defense Science and Engineering Graduate Fellowship. This work was also supported in part by NSF grant #1018625, ARL Cooperative Agreement W911NF-09-2-0072, and by the BOLT and GALE programs of the Defense Advanced Research Projects Agency, Contracts HR0011-12-C-0015 and HR0011-06-2-001, respectively. Any opinions, findings, conclusions, or recommendations expressed are the authors' and do not necessarily reflect those of the sponsors.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of Emperical Methods in Natural Language Processing*.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:2003.

Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of Emperical Methods in Natural Language Processing*.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318.

David Chiang, Steve DeNeefe, and Michael Pust. 2011. Two easy improvements to lexical weighting. In *Proceedings of the Human Language Technology Conference*.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of ACL System Demonstrations*.

Vladimir Eidelman. 2012. Optimization strategies for online large-margin learning in machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of Emperical Methods in Natural Language Processing*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Stroudsburg, PA, USA.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of Emperical Methods in Natural Language Processing*.

A. K. Mccallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29(21), pages 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 311–318.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of Emperical Methods in Natural Language Processing*.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

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