Topic Models for Translation Domain Adaptation

Yuening Hu[1], **Ke Zhai**[1], **Vladimir Edelman**[1], **Jordan Boyd-Graber**[2]
[1]Computer Science, [2]iSchool and UMIACS
[1,2]University of Maryland
{ynhu,zhaike}@cs.umd.edu, {vlad,jbg}@umiacs.umd.edu

Abstract

Topic models have been successfully applied in domain adaptation for translation models. However, previous works applied topic models only on source side and ignored the relations between source and target languages in machine translation. This paper corrects this omission by learning models that can also use target-side information to discover more distinct topics: tree-based topic models and polylingual topic models. We evaluate the models using translation quality.

1 Introduction

Domain knowledge plays an important role in improving the performance of *statistical machine translation* (SMT). Often, translations within a domain are consistently better than across domains, since domains (e.g., newswire vs. blogs) may vary widely in their word choice and style; what may be correct usage in a general setting, or in one domain, is not necessarily preferable in a specialized domain. Indeed, sometimes the domain can change the meaning of a phrase entirely.

This problem has led to recent adapting the *translation model* toward particular domains of interest [1, 2, 3]. The intuition behind translation model adaptation is to increase the likelihood of selecting relevant phrases for translation. Matsoukas et al. [4] introduced assigning a pair of binary features to each training sentence, indicating sentences' genre and collection as a way to capture domains. Chiang et al. [5] further extended the idea to directly optimize the weight of the genre and collection features by computing a separate word translation table for each feature. However, these approaches consider domains as a hard constraint: externally imposed and hand labeled.

We are interested in more flexible unsupervised approaches for domain adaptation. Eidelman et al. [6] applied topic models, exemplified by *latent Dirichlet allocation* [7], to perform unsupervised domain induction and bias machine translation systems toward relevant translations based on topic-specific contexts. Introducing adaptation features based on probabilistic domain membership into the translation model obtained a significant gain on BLEU score [8] of translation results.

However, vanilla LDA cannot connect the source and target sentences effectively. Eidelman et al. [6] trained vanilla topic models on the source side of a parallel corpus, and assumed that the whole sentence pair shared the same topic distribution from the source sentence. This topic model ignores the knowledge of the target language nor does it incorporate knowledge that may aid translations.

This paper corrects this omission of [6] by introducing tree-based topic models [9, 10, 11] and polylingual topic models [12] to learn topics **across languages** and **model language-specific variations in different domains**. While a source sentence has various meanings in domains, the target translation would help to figure out the correct domain that this sentence pair should belong to.

In this paper, we review how the vanilla topic models are applied to the problem of domain adaptation for translation in [6] in Section 2. We further introduce tree-based topic models and polylingual topic models to translation domain adaption in Section 3. Section 4 evaluates our models on BLEU and TER scores based on topic-based domain features against the results using vanilla topic models [6].

2 Background: Topic Models for Translation Domain Adaption

In machine translation, lexical weighting features estimate the phrase pair quality by combining lexical translation probabilities of words in the phrase[1] [13]. Lexical conditional probabilities $p(e|f)$ are 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 [13].

Chiang et al. [5] showed that is it beneficial to condition the lexical weighting features on provenance by assigning each sentence pair a set of features, $f_s(e|f)$, 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.

Eidelman et al. [6] extended provenance to cover a set of automatically generated topics z_n. Given a parallel training corpus composed of documents d_i, they build a source-side topic model over the training data, which provides a topic distribution $p(z_n|d_i)$ for $z_n = \{1,\ldots,K\}$ over each document, using Latent Dirichlet Allocation (LDA) [7]. To obtain the lexical probability conditioned on topic distribution, they first compute the expected count $e_{z_n}(e,f)$ of a word pair under topic z_n:

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

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

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

Thus there are totally $2{\cdot}K$ new word translation tables, one for each $p_{z_n}(e|f)$ and $p_z(f|e)$, and as many new corresponding features $f_{z_n}(e|f)$, $f_{z_n}(f|e)$.

The document topic distribution of test data $p(z_n|d)$ is inferred based on the topics of T. The adapted 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 topic distribution of the sentence from which we are extracting the phrase. [6] computes the resulting model score by combining these features in a linear model with other standard MT features and optimizing their weights using an online large-margin learner [14].

3 Topic Models Extensions for Translation Domain Adaption

While Eidelman et al. [6] has successfully improved the translation results by introducing topic models for domain adaption, they extracted topics only based on source side. This section discusses why we need topic models on both target and source sides, and introduces tree-based topic models and polylingual topic models to extract topics across languages to further improve the translation.

3.1 Topic Models Across Languages

A common problem in translation systems is word choice: there are many possible translations for a given word, and a common problem is choosing the right word that preserves the meaning, is idiomatic, and is consistent with the context. Machine translation systems have improved their ability to model local context correctly, but global context (the domain information captured by topic models) is still a challenge.

While Eidelman et al. [6] introduced the topic models to learn the global context, they built monolingual topic models and ignored the target side, which may also provide important information for domain disambiguation. For example, the Chinese sentence "很多粉丝" has two meanings: either "a lot of noodles" or "a lot of fans". If we do topic modeling only on the source, without any other context or information, either interpretation is plausible. However, if we have access to the English translation, this can disambiguate the meaning; i.e., "a lot of fans" implies an entertainment topic.

As a result, it is important to do topic modeling on both source and target languages. Specifically, different languages can complement each other to assuage the ambiguity problems. We will introduce two different topic models to enforce a relationship between the source and target languages.

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

3.2 Tree-based Topic Models

Vanilla LDA uses a symmetric Dirichlet prior for all words, and it ignores the potential relations between words. Instead, tree-based topic models [9, 10, 11] use a tree-structured prior to model the relations between words, which can be further extended to model the relations between words across languages. Thus tree-based topic models can be used for extracting multilingual topics [15].

To see how correlations can occur, consider the generative process. Start with a rooted tree structure that contains internal nodes and leaf nodes. For the tree of topic k, internal nodes have a distribution $\pi_{k,i}$ over children, where $\pi_{k,i}$ comes from per-node Dirichlet parameterized by β_i. Each leaf node is associated with a word, and each word must appear in at least (possibly more than) one leaf node.

To generate a word from topic k, start at the root. Select a child $x_0 \sim \text{Mult}(\pi_{k,\text{ROOT}})$, and traverse the tree until reaching a leaf node. Then emit the leaf's associated word. This walk replaces the draw from a topic's multinomial distribution over words. The rest of the generative process for LDA remains the same, with θ, the per-document topic multinomial, and z, the topic assignment.

This tree structure encodes correlations. The closer types are in the tree, the more correlated they are. Because types can appear in multiple leaves, this encodes polysemy. The path that generates a token is an additional latent variable we must sample. These correlations can come from WordNet [16, 15], domain experts [10], or normal users [11], based on which we can build up the prior tree.

Note that the correlations are not limited to one language; it can be across languages. In the application of machine translation, we can extract correlations from bilingual dictionaries of both source and target languages, and apply the tree-based topic models. Once we obtain the document topic distribution θ, we can apply the same domain adaption model as in [6] for machine translation.

3.3 Polylingual Topic Models

While tree-based topic models connect the different languages by topic word distributions, polylingual topic models [12] extend LDA to parallel corpus. It assumes aligned documents share the same topic distribution, while maintaining different topic word distributions. These aligned documents usually come from different languages, each of which has its unique vocabulary.

The generative story of a document based on polylingual topic models is similar to vanilla LDA, except the words are generated for each language. For each document set, we sample a Dirichlet distribution θ as document topic distribution, then for each language l, we sample a topic and generate a word based the topic word distribution of language l. The correlation is achieved via the shared document topic distribution θ. It is further propagated to the word distribution.

Similarly to tree-based topic models (Section 3.2), once we obtain the document topic distributions θ, we can apply the same idea as in [6] for machine translation.

4 Experiments

To evaluate our approaches, we performed experiments on Chinese to English translation. Our parallel training corpus is taken from the NIST MT evaluation, and includes 1.6M sentence pairs with 44.4M English tokens and 40.4M Chinese tokens, excluding both the non-UN and non-HK Hansards portions. The MT pipeline is the same as [6]. We tuned the parameters to optimize BLEU [8] on the NIST MT06 using MIRA [14], with results reported on two unseen sets, MT03 and MT05.

We use 10 topics. The baseline vanilla topic modeling was performed with Mallet [18] with a Chinese stoplist and setting the per-document Dirichlet parameter $\alpha = 0.01$. We use Tree-TM [19] for tree-based topic modeling, and the implementation in Mallet for polylingual topic models.

Table 1 shows our results of "Tree-TM" and "Poly-TM". "MT-Baseline" and "MERT" are the two baselines in [6], "Vanilla-TM [6]" are taken from [6], which we repeated in the fourth column (with a slightly better outcome). Thus far, our results favorably compare with the baselines. "Poly-TM" performs very similar to "Vanilla-TM", and "Tree-TM" achieves a slight improvement on TER.

Model		MT-Baseline	MERT	Vanilla-TM [6]	Vanilla-TM	Tree-TM	Poly-TM
MT03	↑BLEU	34.31	34.60	35.32	**35.80**	35.73	35.68
	↓TER	61.14	60.66	59.16	58.49	**57.99**	58.44
MT05	↑BLEU	30.63	30.53	31.56	**32.35**	32.13	32.30
	↓TER	65.10	64.56	62.01	61.14	**61.00**	61.24

Table 1: Results on NIST corpus. Tree-TM improved TER a little bit, but a little bit worse on BLEU.

5 Conclusion

While topic models can be used for translation domain adaption, we argue that it is not appropriate to only focus on the source side. We introduce tree-based topic models and polylingual topic models to extract topics across languages. The preliminary results improved the MT baselines, but did not outperform the vanilla topic models.

There are several directions for further exploration. We currently only consider both source and target sides on training data, and we can also apply these more general topic models on test data, incorporating topic models into the decoding or reranking.

Second, we can explore more flexible models that remove the constraints from modeling only on parallel corpora; allowing us to use as much data as we care to, using scalable computing platforms such as Hadoop MapReduce [20]. This requires scalable inference techniques for multilingual topic models, for example, by extending variational inference on MapReduce [21] to these general models.

References

[1] Axelrod, A., X. He, J. Gao. Domain adaptation via pseudo in-domain data selection. In *EMNLP*. 2011.

[2] Foster, G., C. Goutte, R. Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*. 2010.

[3] Snover, M., B. Dorr, R. Schwartz. Language and translation model adaptation using comparable corpora. In *EMNLP*. 2008.

[4] Matsoukas, S., A.-V. I. Rosti, B. Zhang. Discriminative corpus weight estimation for machine translation. In *EMNLP*. 2009.

[5] Chiang, D., S. DeNeefe, M. Pust. Two easy improvements to lexical weighting. In *HLT*. 2011.

[6] Eidelman, V., J. Boyd-Graber, P. Resnik. Topic models for dynamic translation model adaptation. In *ACL*. 2012.

[7] Blei, D. M., A. Ng, M. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

[8] Papineni, K., S. Roukos, T. Ward, et al. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. 2002.

[9] Boyd-Graber, J., D. M. Blei, X. Zhu. A topic model for word sense disambiguation. In *EMNLP*. 2007.

[10] Andrzejewski, D., X. Zhu, M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML*. 2009.

[11] Hu, Y., J. Boyd-Graber, B. Satinoff, et al. Interactive topic modeling. In *MLJ*. To Appear.

[12] Mimno, D., H. Wallach, J. Naradowsky, et al. Polylingual topic models. In *EMNLP*. 2009.

[13] Koehn, P., F. J. Och, D. Marcu. Statistical phrase-based translation. In *NAACL*. 2003.

[14] Boyd-Graber, J., P. Resnik. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *EMNLP*. 2010.

[15] Miller, G. A. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264, 1990.

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

[17] McCallum, A. K. Mallet: A machine learning for language toolkit, 2002. Http://www.cs.umass.edu/ mccallum/mallet.

[18] Hu, Y., J. Boyd-Graber. Efficient tree-based topic modeling. In *ACL*. 2012.

[19] Dean, J., S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI*. 2004.

[20] Zhai, K., J. Boyd-Graber, N. Asadi, et al. Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *WWW*. 2012.