Hierarchical Phrase Table Combination for Machine Translation

Conghui Zhu[1] **Taro Watanabe**[2] **Eiichiro Sumita**[2] **Tiejun Zhao**[1]
[1]School of Computer Science and Technology
Harbin Institute of Technology (HIT), Harbin, China
[2]National Institute of Information and Communication Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan
{chzhu,tjzhao}@mtlab.hit.edu.cn
{taro.watanabe,Sumita}@nict.go.jp

Abstract

Typical statistical machine translation systems are batch trained with a given training data and their performances are largely influenced by the amount of data. With the growth of the available data across different domains, it is computationally demanding to perform batch training every time when new data comes. In face of the problem, we propose an efficient phrase table combination method. In particular, we train a Bayesian phrasal inversion transduction grammars for each domain separately. The learned phrase tables are hierarchically combined as if they are drawn from a hierarchical Pitman-Yor process. The performance measured by BLEU is at least as comparable to the traditional batch training method. Furthermore, each phrase table is trained separately in each domain, and while computational overhead is significantly reduced by training them in parallel.

1 Introduction

Statistical machine translation (SMT) systems usually achieve 'crowd-sourced' improvements with batch training. Phrase pair extraction, the key step to discover translation knowledge, heavily relies on the scale of training data. Typically, the more parallel corpora used, the more phrase pairs and more accurate parameters will be learned, which can obviously be beneficial to improving translation performances. Today, more parallel sentences are drawn from divergent domains, and the size keeps growing. Consequently, how to effectively use those data and improve translation performance becomes a challenging issue.

Batch retraining is not acceptable for this case, since it demands serious computational overhead when training on a large data set, and it requires us to re-train every time new training data is available. Even if we can handle the large computation cost, improvement is not guaranteed every time we perform batch tuning on the newly updated training data obtained from divergent domains. Traditional domain adaption methods for SMT are also not adequate in this scenario. Most of them have been proposed in order to make translation systems perform better for resource-scarce domains when most training data comes from resource-rich domains, and ignore performance on a more generic domain without domain bias (Wang et al., 2012). As an alternative, incremental learning may resolve the gap by incrementally adding data sentence-by-sentence into the training data. Since SMT systems trend to employ very large scale training data for translation knowledge extraction, updating several sentence pairs each time will be annihilated in the existing corpus.

This paper proposes a new phrase table combination method. First, phrase pairs are extracted from each domain without interfering with other domains. In particular, we employ the non-parametric Bayesian phrasal inversion transduction grammar (ITG) of Neubig et al. (2011) to perform phrase table extraction. Second, extracted phrase tables are combined as if they are drawn from a hierarchical Pitman-Yor process, in which the phrase tables represented as tables in the Chinese restaurant process (CRP) are hierarchically chained by treating each of the previously learned phrase tables as prior to the current one. Thus, we can easily update the chain of phrase tables by appending the newly extracted phrase table and by treating the chain of the previous ones as its prior.

Experiment results indicate that our method can achieve better translation performance when there exists a large divergence in domains, and can

This joint work was done while the first author visited NICT.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 802–810,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

achieve at least comparable results to batch training methods, with a significantly less computational overhead.

The rest of the paper is organized as follows. In Section 2, we introduce related work. In section 3, we briefly describe the translation model with phrasal ITGs and Pitman-Yor process. In section 4, we explain our hierarchical combination approach and give experiment results in section 5. We conclude the paper in the last section.

2 Related Work

Bilingual phrases are cornerstones for phrase-based SMT systems (Och and Ney, 2004; Koehn et al., 2003; Chiang, 2005) and existing translation systems often get 'crowd-sourced' improvements (Levenberg et al., 2010). A number of approaches have been proposed to make use of the full potential of the available parallel sentences from various domains, such as domain adaptation and incremental learning for SMT.

The translation model and language model are primary components in SMT. Previous work proved successful in the use of large-scale data for language models from diverse domains (Brants et al., 2007; Schwenk and Koehn, 2008). Alternatively, the language model is incrementally updated by using a succinct data structure with a interpolation technique (Levenberg and Osborne, 2009; Levenberg et al., 2011).

In the case of the previous work on translation modeling, mixed methods have been investigated for domain adaptation in SMT by adding domain information as additional labels to the original phrase table (Foster and Kuhn, 2007). Under this framework, the training data is first divided into several parts, and phase pairs are extracted with some sub-domain features. Then all the phrase pairs and features are tuned together with different weights during decoding. As a way to choose the right domain for the domain adaption, a classifier-based method and a feature-based method have been proposed. Classification-based methods must at least add an explicit label to indicate which domain the current phrase pair comes from. This is traditionally done with an automatic domain classifier, and each input sentence is classified into its corresponding domain (Xu et al., 2007). As an alternative to the classification-based approach, Wang et al. (2012) employed a feature-based approach, in which phrase pairs are enriched by a feature set to potentially reflect the domain information. The similarity calculated by a information retrieval system between the training subset and the test set is used as a feature for each parallel sentence (Lu et al., 2007). Monolingual topic information is taken as a new feature for a domain adaptive translation model and tuned on the development set (Su et al., 2012). Regardless of underlying methods, either classifier-based or feature-based method, the performance of current domain adaptive phrase extraction methods is more sensitive to the development set selection. Usually the domain similar to a given development data is usually assigned higher weights.

Incremental learning in which new parallel sentences are incrementally updated to the training data is employed for SMT. Compared to traditional frequent batch oriented methods, an online EM algorithm and active learning are applied to phrase pair extraction and achieves almost comparable translation performance with less computational overhead (Levenberg et al., 2010; González-Rubio et al., 2011). However, their methods usually require numbers of hyperparameters, such as mini-batch size, step size, or human judgment to determine the quality of phrases, and still rely on a heuristic phrase extraction method in each phrase table update.

3 Phrase Pair Extraction with Unsupervised Phrasal ITGs

Recently, phrase alignment with ITGs (Cherry and Lin, 2007; Zhang et al., 2008; Blunsom et al., 2008) and parameter estimation with Gibbs sampling (DeNero and Klein, 2008; Blunsom and Cohn, 2010) are popular. Here, we employ a method proposed by Neubig et al. (2011), which uses parametric Bayesian inference with the phrasal ITGs (Wu, 1997). It can achieve comparable translation accuracy with a much smaller phrase table than the traditional GIZA++ and heuristic phrase extraction methods. It has also been proved successful in adjusting the phrase length granularity by applying character-based SMT with more sophisticated inference (Neubig et al., 2012).

ITG is a synchronous grammar formalism which analyzes bilingual text by introducing inverted rules, and each ITG derivation corresponds to the alignment of a sentence pair (Wu, 1997). Translation probabilities of ITG phrasal alignments can be estimated in polynomial time by slightly limiting word reordering (DeNero and Klein, 2008).

More formally, $P\big(\langle e, f\rangle; \theta_x, \theta_t\big)$ are the probability of phrase pairs $\langle e, f\rangle$, which is parameterized by a phrase pair distribution θ_t and a symbol distribution θ_x. θ_x is a Dirichlet prior, and θ_t is estimated with the Pitman-Yor process (Pitman and Yor, 1997; Teh, 2006), which is expressed as

$$\theta_t \sim PY\big(d, s, P_{dac}\big) \tag{1}$$

where d is the discount parameter, s is the strength parameter, and , and P_{dac} is a prior probability which acts as a fallback probability when a phrase pair is not in the model.

Under this model, the probability for a phrase pair found in a bilingual corpus $\langle E, F\rangle$ can be represented by the following equation using the Chinese restaurant process (Teh, 2006):

$$P\big(\langle e_i, f_i\rangle; \langle E, F\rangle\big) = \frac{1}{C+s}(c_i - d \times t_i) + \frac{1}{C+s}(s + d \times T) \times P_{dac}(\langle e_i, f_i\rangle) \tag{2}$$

where

1. c_i and t_i are the customer and table count of the ith phrase pair $\langle e_i, f_i\rangle$ found in a bilingual corpus $\langle E, F\rangle$;

2. C and T are the total customer and table count in corpus $\langle E, F\rangle$;

3. d and s are the discount and strengthen hyperparameters.

The prior probability P_{dac} is recursively defined by breaking a longer phrase pair into two through the recursive ITG's generative story as follows (Neubig et al., 2011):

1. Generate symbol x from $P_x(x; \theta_x)$ with three possible values: $Base$, REG, or INV.

2. Depending on the value of x take the following actions.

   a. If $x = Base$, generate a new phrase pair directly from P_{base}.

   b. If $x = REG$, generate $\langle e_1, f_1\rangle$ and $\langle e_2, f_2\rangle$ from $P\big(\langle e, f\rangle; \theta_x, \theta_t\big)$, and concatenate them into a single phrase pair $\langle e_1e_2, f_1f_2\rangle$.

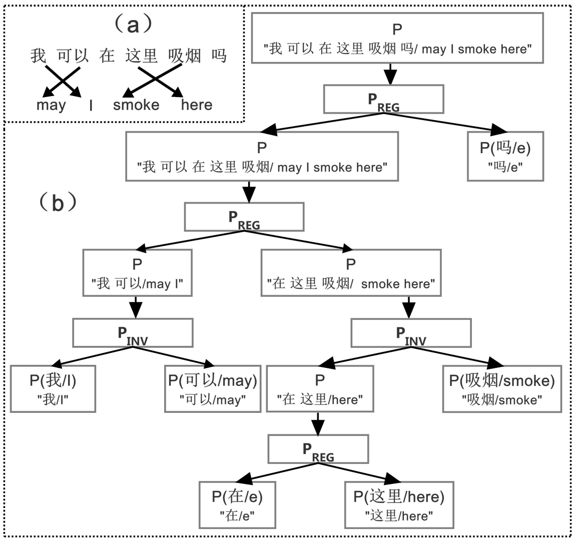

Figure 1: A word alignment (a), and its hierarchical derivation (b).

   c. If $x = INV$, follow a similar process as b, but concatenate f_1 and f_2 in reverse order $\langle e_1e_2, f_2f_1\rangle$.

Note that the P_{dac} is recursively defined through the binary branched P, which in turns employs P_{dac} as a prior probability. P_{base} is a base measure defined as a combination of the IBM Models in two directions and the unigram language models in both sides. Inference is carried out by a heuristic beam search based block sampling with an efficient look ahead for a faster convergence (Neubig et al., 2012).

Compared to GIZA++ with heuristic phrase extraction, the Bayesian phrasal ITG can achieve competitive accuracy under a smaller phrase table size. Further, the fallback model can incorporate phrases of all granularity by following the ITG's recursive definition. Figure 1 (b) illustrates an example of the phrasal ITG derivation for word alignment in Figure 1 (a) in which a bilingual sentence pair is recursively divided into two through the recursively defined generative story.

4 Hierarchical Phrase Table Combination

We propose a new phrase table combination method, in which individually learned phrase table are hierarchically chained through a hierarchical Pitman-Yor process.

Firstly, we assume that the whole training data $\langle E, F\rangle$ can be split into J domains, $\{\langle E^1, F^1\rangle, \ldots, \langle E^J, F^J\rangle\}$. Then phrase pairs are

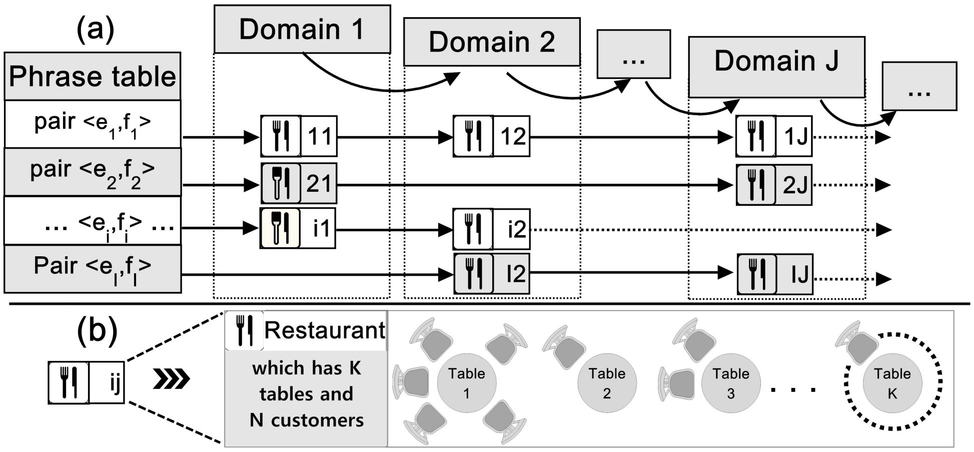

Figure 2: A hierarchical phrase table combination (a), and a basic unit of a Chinese restaurant process with K tables and N customers.

extracted from each domain j ($1 \leq j \leq J$) separately with the method introduced in Section 3. In traditional domain adaptation approaches, phrase pairs are extracted together with their probabilities and/or frequencies so that the extracted phrase pairs are merged uniformly or after scaling.

In this work, we extract the table counts for each phrase pair under the Chinese restaurant process given in Section 3. In Figure 2 (b), a CRP is illustrated which has K tables and N customers with each chair representing a customer. Meanwhile there are two parameters, discount and strength for each domain similar to the ones in Equation (1).

Our proposed hierarchical phrase table combination can be formally expressed as following:

$$\begin{aligned}
&\theta^1 \sim PY(d^1, s^1, P^2)\\
&\cdots\ \cdots\\
&\theta^j \sim PY(d^j, s^j, P^{j+1})\\
&\cdots\ \cdots\\
&\theta^J \sim PY\big(d^J, s^J, P^J_{base}\big)
\end{aligned} \tag{3}$$

Here the $(j+1)$th layer hierarchical Pitman-Yor process is employed as a base measure for the jth layer hierarchical Pitman-Yor process. The hierarchical chain is terminated by the base measure from the Jth domain P^J_{base}. The hierarchical structure is illustrated in Figure 2 (a) in which the solid lines implies a fall back using the table counts from the subsequent domains, and the dotted lines means the final fallback to the base measure P^J_{base}. When we query a probability of a phrase pair $\langle e, f\rangle$, we first query the probability of the first layer $P^1(\langle e, f\rangle)$. If $\langle e, f\rangle$ is not in the model, we will fallback to the next level of $P^2(\langle e, f\rangle)$. This process continues until we reach the J_{th} base measure of $P^J(\langle e, f\rangle)$. Each fallback can be viewed as a translation knowledge integration process between subsequent domains.

For example in Figure 2 (a), the ith phrase pair $\langle e_i, f_i\rangle$ appears only in the domain 1 and domain 2, so its translation probability can be calculated by substituting Equation (3) with Equation (2):

$$\begin{aligned}
P\big(\langle e_i, f_i\rangle; \langle E, F\rangle\big) &= \frac{1}{C^1 + s^1}(c^1_i - d^1 \times t^1_i)\\
&+ \frac{s^1 + d^1 \times T^1}{(C^1 + s^1) \times (C^2 + s^2)}(c^2_i - d^2 \times t^2_i)\\
&+ \prod_{j=1}^{J}\big(\frac{s^j + d^j \times T^j}{C^j + s^j}\big) \times P^J_{base}(\langle e_i, f_i\rangle)
\end{aligned} \tag{4}$$

where the superscript indicates the domain for the corresponding counts, i.e. c^j_i for the customer count in the jth domain. The first term in Equation (4) is the phrase probability from the first domain, and the second one comes from the second domain, but weighted by the fallback weight of the 1st domain. Since $\langle e_i, f_i\rangle$ does not appear in the rest of the layers, the last term is taken from all the fallback weight from the second layer to the Jth layer with the final P^J_{base}. All the parameters θ^j and hyperparameters d^j and s^j, are obtained by learning on the jth domain. Returning the hyperparameters again when cascading another domain may improve the performance of the combination weight, but we will leave it for future work. The hierarchical process can be viewed as an instance of adapted integration of translation knowledge from each sub-domain.

Algorithm 1 Translation Probabilities Estimation

```
Input: $c_i^j$, $t_i^j$, $P_{base}^j$, $C^j$, $T^j$, $d^j$ and $s^j$
Output: The translation probabilities for each pair
for all phrase pair $\langle e_i, f_i \rangle$ do
    Initialize the $P(\langle e_i, f_i \rangle) = 0$ and $w_i = 1$
    for all domain $\langle E_j, F_j \rangle$ such that $1 \leqslant j \leqslant J - 1$ do
        if $\langle e_i, f_i \rangle \in \langle E_j, F_j \rangle$ then
            $P(\langle e_i, f_i \rangle)$ += $w_i \times (C_i^j - d^j \times t_i^j)/(C^j + s^j)$
        end if
        $w_i = w_i \times (s^j + d^j \times T^j)/(C^j + s^j)$
    end for
    $P(\langle e_i, f_i \rangle)$ += $w_i \times (C_i^J - d^J \times t_i^J + (s^J + d^J \times T^J) \times P_{base}^J(\langle e_i, f_i \rangle))/(C^J + s^J)$
end for
```

Our approach has several advantages. First, each phrase pair extraction can concentrate on a small portion of domain-specific data without interfering with other domains. Since no tuning stage is involved in the hierarchical combination, we can easily include a new phrase table from a new domain by simply chaining them together. Second, phrase pair phrase extraction in each domain is completely independent, so it is easy to parallelize in a situation where the training data is too large to fit into a small amount of memory. Finally, new domains can be integrated incrementally. When we encounter a new domain, and if a phrase pair is completely new in terms of the model, the phrase pair is simply appended to the current model, and computed without the fallback probabilities, since otherwise, the phrase pair would be boosted by the fallback probabilities. Pitman-Yor process is also employed in n-gram language models which are hierarchically represented through the hierarchical Pitman-Yor process with switch priors to integrate different domains in all the levels (Wood and Teh, 2009). Our work incrementally combines the models from different domains by directly employing the hierarchical process through the base measures.

5 Experiment

We evaluate the proposed approach on the Chinese-to-English translation task with three data sets with different scales.

Data set	Corpus	#sent. pairs
IWSLT	HIT	$52,603$
	BTEC	$19,975$
FBIS	Domain 1	$47,993$
	Domain 2	$30,272$
	Domain 3	$49,509$
	Domain 4	$38,228$
	Domain 5	$55,913$
LDC	News	$221,915$
	News	$95,593$
	Magazine	$98,335$
	Magazine	$254,488$
	Finance	$86,112$

Table 1: The sentence pairs used in each data set.

5.1 Experiment Setup

The first data set comes from the IWSLT2012 OLYMPICS task consisting of two training sets: the HIT corpus, which is closely related to the Beijing 2008 Olympic Games, and the BTEC corpus, which is a multilingual speech corpus containing tourism-related sentences. The second data set, the FBIS corpus, is a collection of news articles and does not have domain information itself, so a Latent Dirichlet Allocation (LDA) tool, PLDA[1], is used to divide the whole corpus into 5 different sub-domains according to the concatenation of the source side and target side as a single sentence (Liu et al., 2011). The third data set is composed of 5 corpora[2] from LDC with various domains, including news, magazine, and finance. The details are shown in Table 1.

In order to evaluate our approach, four phrase pair extraction methods are performed:

1. GIZA-linear: Phase pairs are extracted in each domain by GIZA++ (Och and Ney, 2003) and the ”grow-diag-final-and” method with a maximum length 7. The phrase tables from various domains are linearly combined by averaging the feature values.

2. Pialign-linear: Similar to GIZA-linear, but we employed the phrasal ITG method described in Section 3 using the pialign toolkit [3] (Neubig et

[1] http://code.google.com/p/plda/

[2] In particular, they come from LDC catalog number: LDC2002E18, LDC2002E58, LDC2003E14, LDC2005E47, LDC2006E26, in this order.

[3] http://www.phontron.com/pialign/

Methods	IWSLT		FBIS		LDC	
	BLEU	Size	BLEU	Size	BLEU	Size
GIZA-linear	19.222	1,200,877	29.342	15,369,028	30.67	77,927,347
Pialign-linear	19.534	876,059	29.858	7,235,342	31.12	28,877,149
GIZA-batch	19.616	1,185,255	**31.38**	13,737,258	**32.06**	63,606,056
Pialign-batch	19.506	841,931	31.104	6,459,200		
Pialign-adaptive	19.624	841,931	30.926	6,459,200		
Hier-combin	**20.32**	876,059	31.29	7,235,342	32.03	28,877,149

Table 2: BLEU scores and phrase table size by alignment method and probabilities estimation method. Pialign was run with five samples. Because of computational overhead, the baseline Pialign-batch and Pialign-adaptive were not run on the largest data set.

al., 2011). Extracted phrase pairs are linearly combined by averaging the feature values.

3. GIZA-batch: Instead of splitting into each domain, the data set is merged as a single corpus and then a heuristic GZA-based phrase extraction is performed, similar as GIZA-linear.

4. Pialign-batch: Similar to the GIZA-batch, a single model is estimated from a single, merged corpus. Since pialign cannot handle large data, we did not experiment on the largest LDC data set.

5. Pialign-adaptive: Alignment and phrase pairs extraction are same to Pialign-batch, while translation probabilities are estimated by the adaptive method with monolingual topic information (Su et al., 2012). The method established the relationship between the out-of-domain bilingual corpus and in-domain monolingual corpora via topic distribution to estimate the translation probability.

$$\begin{aligned} \o(\tilde{e}|\tilde{f}) &= \sum_{t_f} \o(\tilde{e}, t_f|\tilde{f}) \\ &= \sum_{t_f} \o(\tilde{e}|t_f, \tilde{f}) \cdot P(t_f|\tilde{f}) \end{aligned} \tag{5}$$

where $\o(\tilde{e}|t_f, \tilde{f})$ is the probability of translating $\tilde{f}$ into $\tilde{e}$ given the source-side topic $\tilde{f}$, $P(t_f|\tilde{f})$ is the phrase-topic distribution of f.

The method we proposed is named Hier-combin. It extracts phrase pairs in the same way as the Pialign-linear. In the phrase table combination process, the translation probability of each phrase pair is estimated by the Hier-combin and the other features are also linearly combined by averaging the feature values. Pialign is used with default parameters. The parameter 'samps' is set to 5, which indicates 5 samples are generated for a sentence pair.

The IWSLT data consists of roughly $2,000$ sentences and $3,000$ sentences each from the HIT and BTEC for development purposes, and the test data consists of $1,000$ sentences. For the FBIS and LDC task, we used NIST MT 2002 and 2004 for development and testing purposes, consisting of 878 and $1,788$ sentences respectively. We employ Moses, an open-source toolkit for our experiment (Koehn et al., 2007). SRILM Toolkit (Stolcke, 2002) is employed to train 4-gram language models on the Xinhua portion of Gigaword corpus, while for the IWLST2012 data set, only its training set is used. We use batch-MIRA (Cherry and Foster, 2012) to tune the weight for each feature and translation quality is evaluated by the case-insensitive BLEU-4 metric (Papineni et al., 2002). The BLEU scores reported in this paper are the average of 5 independent runs of independent batch-MIRA weight training, as suggested by (Clark et al., 2011).

5.2 Result and Analysis

5.2.1 Performances of various extraction methods

We carry out a series of experiments to evaluate translation performance. The results are listed in Table 2. Our method significantly outperforms the baseline Pialign-linear. Except for the translation probabilities, the phrase pairs of two methods are exactly same, so the number of phrase pairs are equal in the two methods. Further more, the performance of the baseline Pialign-adaptive is also higher than the baseline Pialign-linear's and lower than ours. This proves that the adaptive method

Methods	Task	Time(minute)
Batch	Retraining	536.9
Hierarchical Combination	Parallel Extraction	122.55
	Integrating	1.5
	Total	124.05

Table 3: Minutes used for alignment and phase pair extraction in the FBIS data set.

with monolingual topic information is useful in the tasks, but our approach with the hierarchical Pitman-Yor process can estimate more accurate translation probabilities based on all the data from various domains.

Compared with the GIZA-batch, our approach achieves competitive performance with a much smaller phrase table. The number of phase pairs generated by our method is only 73.9%, 52.7%, and 45.4% of the GIZA-batch's respectively. In the IWLST2012 data set, there is a huge difference gap between the HIT corpus and the BTEC corpus, and our method gains 0.814 BLEU improvement. While the FBIS data set is artificially divided and no clear human assigned differences among subdomains, our method loses 0.09 BLEU.

In the framework we proposed, phrase pairs are extracted from each domain completely independent of each other, so those tasks can be executed on different machines, at different times, and of course in parallel when we assume that the domains are not incrementally added in the training data. The runtime of our approach and the batch-based ITGs sampling method in the FBIS data set is listed in Table 3 measured on a 2.7 GHz E5-2680 CPU and 128 Gigabyte memory. When comparing the hier-combin with the pialign-batch, the BLEU scores are a little higher while the time spent for training is much lower, almost one quarter of the pialign-batch.

Even the performance of the pialign-linear is better than the Baseline GIZA-linear's, which means that phrase pair extraction with hierarchical phrasal ITGs and sampling is more suitable for domain adaptation tasks than the combination GIZA++ and a heuristic method.

Generally, the hierarchical combination method exploits the nature of a hierarchical Pitman-Yor process and gains the advantage of its smoothing effect, and our approach can incrementally generate a succinct phrase table based on all the data from various domains with more accurate probabilities. Traditional SMT phrase pair extraction is batch-based, while our method has no obvious shortcomings in translation accuracy, not to mention efficiency.

5.2.2 Effect of Integration Order

Here, we evaluate whether our hierarchical combination is sensitive to the order of the domains when forming a hierarchical structure. Through Equation (3), in our experiments, we chained the domains in the order listed in Table 1, which is in almost chronological order. Table 4 shows the BLEU scores for the three data sets, in which the order of combining phrase tables from each domain is alternated in the ascending and descending of the similarity to the test data. The similarity between the data from each domain and the test data is calculated using the perplexity measure with 5-gram language model. The model learned from the domain more similar to the test data is placed in the front so that it can largely influence the parameter computation with less backoff effects. There is a big difference between the two opposite order in IWSLT 2012 data set, in which more than one point of decline in BLEU score when taking the BTEC corpus as the first layer. Note that the perplexity of BTEC was 344.589 while that of HIT was 107.788. The result may indicate that our hierarchical phrase combination method is sensitive to the integration order when the training data is small and there exists large gap in the similarity. However, if most domains are similar (FBIS data set) or if there are enough parallel sentence pairs (NIST data set) in each domain, then the translation performances are almost similar even with the opposite integrating orders.

	IWSLT	FBIS	LDC
Descending	20.154	30.491	31.268
Ascending	19.066	30.388	31.254
Difference	**1.088**	**0.103**	**0.014**

Table 4: BLEU scores for the hierarchical model with different integrating orders. Here Pialign was run without multi-samples.

6 Conclusion and Future Work

In this paper, we present a novel hierarchical phrase table combination method for SMT, which can exploit more of the potential from all of data coming from various fields and generate a succinct phrase table with more accurate translation probabilities. The method assumes that a combined model is derived from a hierarchical Pitman-Yor process with each prior learned separately in each domain, and achieves BLEU scores competitive with traditional batch-based ones. Meanwhile, the framework has natural characteristics for parallel and incremental phrase pair extraction. The experiment results on three different data sets indicate the effectiveness of our approach.

In future work, we will also introduce incremental learning for phase pair extraction inside a domain, which means using the current translation probabilities already obtained as the base measure of sampling parameters for the upcoming domain. Furthermore, we will investigate any tradeoffs between the accuracy of the probability estimation and the coverage of phrase pairs.

Acknowledgments

We would like to thank our colleagues in both HIT and NICT for insightful discussions, and three anonymous reviewers for many invaluable comments and suggestions to improve our paper. This work is supported by National Natural Science Foundation of China (61100093, 61173073, 61073130, 61272384), and the Key Project of the National High Technology Research and Development Program of China (2011AA01A207).

References

Phil Blunsom and Trevor Cohn. 2010. Inducing synchronous grammars with slice sampling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 238–241, Los Angeles, California, June. Association for Computational Linguistics.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL*, pages 200–208, Columbus, Ohio, June. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June. Association for Computational Linguistics.

Colin Cherry and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 17–24.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28, Columbus, Ohio, June. Association for Computational Linguistics.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135.

Jesús González-Rubio, Daniel Ortiz-Martinez, and Francisco Casacuberta. 2011. Fast incremental active learning for statistical machine translation. *A-VANCES EN INTELIGENCIA ARTIFICIAL.*

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*, pages 45–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abby Levenberg and Miles Osborne. 2009. Stream-based randomised language models for smt. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 756–764. Association for Computational Linguistics.

Abby Levenberg, Chris Callison-Burch, and Miles Osborne. 2010. Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 394–402, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abby Levenberg, Miles Osborne, and David Matthews. 2011. Multiple-stream language models for statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 177–186, Edinburgh, Scotland, July. Association for Computational Linguistics.

Zhiyuan Liu, Yuzhou Zhang, Edward Y Chang, and Maosong Sun. 2011. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–18.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 343–350, Prague, Czech Republic, June. Association for Computational Linguistics.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 632–641, Portland, Oregon, USA, June. Association for Computational Linguistics.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 165–174, Jeju Island, Korea, July. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449, December.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Jim Pitman and Marc Yor. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Holger Schwenk and Philipp Koehn. 2008. Large and diverse language models for statistical machine translation. In *International Joint Conference on Natural Language Processing*, pages 661–668.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. of ICSLP*.

Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 459–468.

Yee Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics.

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved domain adaptation for statistical machine translation. In *Proceedings of the Conference of the Association for Machine translation, Americas*.

F. Wood and Y. W. Teh. 2009. A hierarchical nonparametric Bayesian approach to statistical language model domain adaptation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.

Jia Xu, Yonggang Deng, Yuqing Gao, and Hermann Ney. 2007. Domain dependent statistical machine translation. In *Proceedings of the MT Summit XI*.

Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of noncompositional phrases with synchronous parsing. In *Proceedings of ACL-08: HLT*, pages 97–105, Columbus, Ohio, June. Association for Computational Linguistics.