Synchronous Constituent Context Model for Inducing Bilingual Synchronous Structures

Xiangyu Duan **Min Zhang**[*] **Qiaoming Zhu**
School of Computer Science & Technology, Soochow University
{xiangyuduan;minzhang;qmzhu}@suda.edu.cn

Abstract

Traditional Statistical Machine Translation (SMT) systems heuristically extract synchronous structures from word alignments, while synchronous grammar induction provides better solutions that can discard heuristic method and directly obtain statistically sound bilingual synchronous structures. This paper proposes Synchronous Constituent Context Model (SCCM) for synchronous grammar induction. The SCCM is different to all previous synchronous grammar induction systems in that the SCCM does not use the Context Free Grammars to model the bilingual parallel corpus, but models bilingual constituents and contexts directly. The experiments show that valuable synchronous structures can be found by the SCCM, and the end-to-end machine translation experiment shows that the SCCM improves the quality of SMT results.

1 Introduction

Traditional Statistical Machine Translation (SMT) learns translation model from bilingual corpus that is sentence aligned. No large-scale hand aligned structures inside the parallel sentences are usually available to the SMT community, while the aligned structures are essential for training the translation model. Thus, various unsupervised methods had been explored to automatically obtain aligned structures inside the parallel sentences. Currently, the dominant method is a two step pipeline that obtains word alignments by unsupervised learning (Brown et al., 1993) at the first step, then obtains aligned structures at the second step by heuristically extracting all bilingual structures that are consistent with the word alignments.

The second step in this two step pipeline is problematic due to its obtained aligned structures, whose counts are heuristically collected and violate valid translation derivations, while most SMT decoders perform translation via valid translation derivations. This problem leads to researches on synchronous grammar induction that discards the heuristic method and the two separate steps pipeline.

Synchronous grammar induction aims to directly obtain aligned structures by using one statistically sound model. The aligned structures in synchronous grammar induction are hierarchical/syntax level (Cohn and Blunsom, 2009) synchronous structures, which can be modeled by Synchronous Context Free Grammars (SCFGs) (Cohn and Blunsom, 2009; Levenberg et al., 2012; Xiao et al., 2012; Xiao and Xiong, 2013) or a kind of SCFGs variant - Inversion Transduction Grammars (ITGs) (Neubig et al., 2011; Cohn and Haffari, 2013). Both SCFGs and ITGs are studied in recent years by using generative or discriminative modeling.

This paper departs from using the above two traditional CFGs-based grammars, and proposes Synchronous Constituent Context Model (SCCM) which models synchronous constituents and contexts directly so that bilingual translational equivalences can be directly modeled. The proposed SCCM is inspired by researches on monolingual grammar induction, whose experience is valuable to the synchronous grammar induction community due to its standard evaluation on released monolingual treebanks, while no hand annotated bilingual synchronous treebank is available for evaluating synchronous

[*] Corresponding Author

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1865–1874, Dublin, Ireland, August 23-29 2014.

grammar induction. According to the evaluation results, the state-of-the-art monolingual grammar induction was achieved by Bayesian modeling of the Constituent Context Model (CCM) (Duan et al., 2013; Klein and Manning, 2002), while traditional CFGs based monolingual grammar induction methods perform well below the CCM.

In view of the significant achievements of the CCM in monolingual grammar induction, we propose the SCCM to apply the CCM to the bilingual case. The tremendous possible constituents and contexts incurred in this bilingual case put a challenge for the SCCM to model such kind of sparse variables. We further propose a non-parametric Bayesian Modeling of the SCCM to cope with the sparse variables. Experiments on Chinese-English machine translation show that meaningful synchronous phrases can be detected by our SCCM, and the performance of the end-to-end SMT is significantly improved.

The rest of the paper is structured as follows: we propose the SCCM in Section 2. The non-parametric Bayesian modeling of the SCCM is presented in Section 3, followed by the presentation of posterior inference for the Bayesian SCCM. Then experiments and results are presented. Conclusion are presented in the final section.

2 Synchronous Constituent Context Model (SCCM)

We propose the SCCM to model synchronous structures explicitly. Unlike Synchronous Context Free Grammars (SCFGs) which are defined on latent production rules of parallel corpus, the SCCM deals with both synchronous tree spans ($syn\ spans$) and non-synchronous spans ($non\text{-}syn\ spans$). All spans are represented by two kinds of strings: bilingual constituents and bilingual contexts. The SCCM is a generative model defined over such representations.

2.1 Bilingual Constituents and Contexts

By extending the concept of constituents and contexts introduced in (Klein and Manning, 2002), we define bilingual constituents and contexts as follows. Bilingual constituents are pairs of contiguous surface strings of sentence spans (bilingual subsequences), bilingual contexts are tokens preceding and following the bilingual constituents. In the SCCM, each bi-span in a sentence pair, either a $syn\ span$ or a $non\text{-}syn\ span$, is represented by a bilingual constituent and a bilingual context.

Fig. 1 gives an illustration of the bilingual constituents and contexts. In Fig. 1-(a), a latent synchronous tree over the example sentence pair is illustrated. With the word alignments shown in the sentence pair, the latent tree over the target sentence "$e_1\ e_2\ e_3$" can be inferred. For the ease of presentation, the latent target side tree is neglected in Fig. 1-(a).

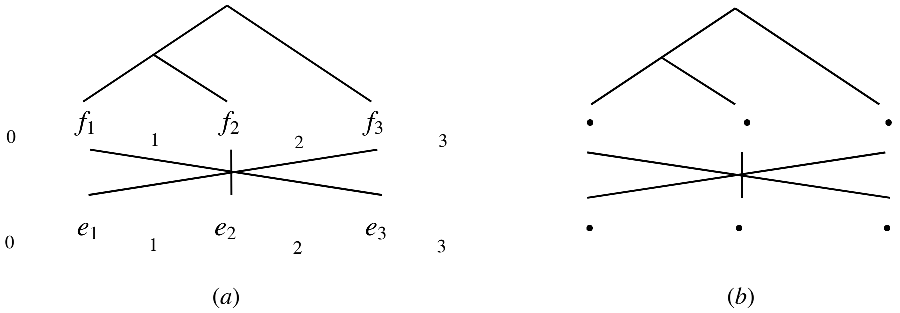
Given the synchronous tree, two sets of bilingual constituents and contexts can be extracted as shown in the two tables of Fig. 1. One is about $syn\ span$s, the other is about $non\text{-}syn\ span$s. $\diamond$ appearing in the contexts denotes a sentence boundary. *nil* appearing in the constituents of the non-tree spans denotes an empty span, which is actually a space between two terminals (or between a terminal and $\diamond$).

2.2 Generative Model

The SCCM computes the joint probability of a sentence pair S and its synchronous tree T as below:

$$\begin{aligned} P(S,T) &= P(S|T)P(T) = P(S|T)P(B)P(T|B) \\ &= P(S|T)P(B) \prod_{\substack{0\leq i\leq j\leq m \\ 0\leq p\leq q\leq n}} P(\alpha_{ij,pq}|B_{ij,pq})P(\beta_{ij,pq}|B_{ij,pq}) \end{aligned} \tag{1}$$

where B denotes a synchronous bracketing skeleton, in which no words are populated. Fig. 1-(b) shows the skeleton of Fig. 1-(a). The skeleton B is considered being filled by the synchronous tree T, and $P(T|B)$ is decomposed into conditional probabilities of bilingual constituents α and contexts β conditioning on $B_{ij,pq}$, a Boolean variable indicating whether the under-consideration bi-span $<i,j><p,q>$ is a $syn\ span$ or not. In particular, $\alpha_{ij,pq}$ denotes the bilingual constituent spanning from i to j on source side sentence, and spanning from p to q on target side sentence. $\beta_{ij,pq}$ denotes the context of $\alpha_{ij,pq}$.



syn span	<0,1><2,3>	<1,2> <1,2>	<2,3><0,1>	<0,2><1,3>	<0,3><0,3>
constituent	(f_1)(e_3)	(f_2)(e_2)	(f_3)(e_1)	($f_1 f_2$)(e_2 e_3)	($f_1 f_2 f_3$)(e_1 e_2 e_3)
context	(◇-f_2)(e_2-◇)	(f_1-f_3)(e_1- e_3)	(f_2-◇)(◇- e_2)	(◇-f_3)(e$_1$-◇)	(◇-◇)(◇-◇)

non-syn span	<1,3><0,1>	<1,1><1,2>	…
constituent	($f_2 f_3$)(e_1)	(*nil*)(e_2)	…
context	(f_1-◇)(◇-e_2)	(f_1-f_2)(e_1- e_3)	…

Figure 1: Illustration of bilingual constituents and contexts over a sentence pair which consists of a source side sentence "f_1 f_2 f_3" and a target side sentence "e_1 e_2 e_3". In (a), the bottom numbers around each word are indexes for denoting spans. A synchronous tree is illustrated in (a), based on which two sets of bilingual constituents and contexts are extracted as shown in the two tables below the tree. Take a *syn span* <1,2><1,2> for example, the source side span <1,2> is "f_2" and the target side span <1,2> is "e_2". They constitutes a bilingual constituent "(f_2)(e_2)", whose context is "(f_1-f_3)(e_1-e_3)" that is preceding and following the bilingual constituent. Figure (b) shows the skeleton of figure (a).

$B_{ij,pq}$ is defined as below:

$$B_{ij,pq} = \begin{cases} 1 & if\ bispan < i,j >< p,q > is\ a\ syn\ span \\ 0 & otherwise \end{cases}$$

In the SCCM, skeletons Bs are restricted to be binary branching and are distributed uniformly. Furthermore, since T and S are consistent, $P(S|T)$ is always equal to 1 in Eq. (1). Therefore, we can infer (with the expansion of the continued multiplication operator of Eq. (1)):

$$P(S,T) \propto \prod_{<i,j><p,q>\in T} (P(\alpha_{ij,pq}|B_{ij,pq}=1)P(\beta_{ij,pq}|B_{ij,pq}=1)) \prod_{<i,j><p,q>\notin T} (P(\alpha_{ij,pq}|B_{ij,pq}=0)P(\beta_{ij,pq}|B_{ij,pq}=0)) \quad (2)$$

where $<i,j><p,q> \in T$ indicates that bi-span $<i,j><p,q>$ is a *syn span* contained in T, $<i,j><p,q> \notin T$ indicates otherwise case. Formula (2) is the basis for Bayesian modeling of the SCCM and the posterior inference that are proposed in the following sections.

3 Bayesian Modeling for the SCCM

For the SCCM, the posterior of a synchronous tree T given the observation of a sentence pair S is: $P(T|S) \propto P(S,T)$. As shown in formula (2), it turns out that the posterior $P(T|S)$ depends on the four kinds of distributions:

$$P(\alpha_{ij,pq}|B_{ij,pq}=1) \qquad P(\beta_{ij,pq}|B_{ij,pq}=1)$$
$$P(\alpha_{ij,pq}|B_{ij,pq}=0) \qquad P(\beta_{ij,pq}|B_{ij,pq}=0)$$

We propose to define two kinds of Bayesian priors over the constituents related variables $\alpha_{ij,pq}|B_{ij,pq}$ and the contexts related variables $\beta_{ij,pq}|B_{ij,pq}$ respectively. Since constituents exhibits richer appearances than contexts, the proposed Bayesian prior over $\alpha_{ij,pq}|B_{ij,pq}$ is more complicate than that over $\beta_{ij,pq}|B_{ij,pq}$.

Specifically, one of the non-parametric Bayesian priors, the Pitman-Yor-Process (PYP) prior, is defined on $\alpha_{ij,pq}|B_{ij,pq}$. The PYP prior can produce the power-law distribution (Goldwater et al., 2009) that is commonly observed in natural languages, and can flexibly model distributions on layer structures due to its defined distribution on distribution hierarchy. The PYP prior had been successfully applied on many NLP tasks such as language modeling (YeeWhye, 2006), word segmentation (Johnson et al., 2007b; Goldwater et al., 2011), dependency grammar induction (Cohen et al., 2008; Cohn et al., 2010), grammar refinement (Liang et al., 2007; Finkel et al., 2007) and Tree-Substitution Grammar induction (Cohn et al., 2010). We use the PYP to model the constituents' layered structure by using the PYP's distribution hierarchy. On $\beta_{ij,pq}|B_{ij,pq}$, we use the Dirichlet distribution for its simplicity because contexts appear in much fewer kinds of surface strings than those of constituents.

3.1 The PYP Prior over Bilingual Constituents

Constituents consist of both words and POS tags. Though in much monolingual grammar induction works, only POS tag sequences were used as the observed constituents for their significant hints of phrases (Klein and Manning, 2002; Cohn et al., 2010), our work needs considering raw words as observation data too because word alignments encode the important translation correspondence and contribute to synchronous bi-spans. But it causes severe data sparse problem due to the quite large number of unique constituents consisting of both words and POS tags. Besides, constituents can be extremely long which intensify the data sparse problem. So, solely using the surface strings of constituents is impractical.

In this section, we propose a hierarchical representation of constituents to overcome the data sparse problem and use the PYP prior on this kind of representation. From top to bottom, the hierarchical representation encodes the information of a bilingual constituent from fine-grained level to coarse-grained levels. The probability of a fine-grained level can be backed-off to the probabilities of coarse-grained levels.

The first (top) level of the hierarchical representation is the bilingual constituent itself. The second level is composed of two sequences: one is word sequence, the other is POS tags sequence. The third level mainly decomposes the second level into boundaries and middle words/POSs. Since the target of inducing synchronous structures in this paper is to induce the latent phrasal equivalences of a parallel sentence, boundaries of bilingual constituents play the key role of identifying phrasal equivalences. The third level is the function to make use of boundaries. Fig. 2 gives an illustration of the hierarchical representation.

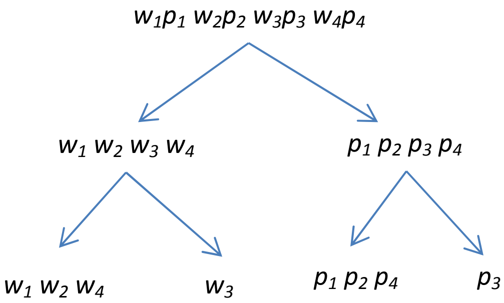


Figure 2: Illustration of the hierarchical representation of a bilingual constituent "w_1p_1 w_2p_2 w_3p_3 w_4p_4". Here w and p denote word and POS respectively, and the suffixes denote positions. Note that both w and p are composite, w denotes a source side word and a target side word, and p denotes the POS case. The second level decomposes the first level into a word sequence and a POS sequence, and the third level decomposes further into boundaries and middle words/POSs. The boundary width in this figure is two for left side boundary and one for right side boundary.

The PYP prior encodes distribution on distribution. Recursively using the PYP prior can create a distribution hierarchy, which is appropriate for modeling the distribution over the hierarchical representations of constituents. Smoothing is fulfilled through backing-off fine-grained level distributions to coarse-grained level distributions.

3.1.1 The PYP Hierarchy

We define the PYP hierarchy over the hierarchical representation of bilingual constituents in a top-down manner. For the topmost (first) level:

$$\begin{aligned} \alpha_{ij,pq}|B_{ij,pq} = b &\sim G_b^{first} \\ G_b^{first} &\sim PYP(d_b^{first}, \theta_b^{first}, P_{word-pos}(.|B_{ij,pq} = b)) \end{aligned}$$

The PYP has three parameters: $(d_b^{first}, \theta_b^{first}, P_{word-pos})$. $P_{word-pos}(.|B_{ij,pq} = b)$ is a $base$ $distribution$ over infinite space of bilingual constituents conditioned on span type b, which provides the back-off probability of $P(\alpha_{ij,pq}|B_{ij,pq} = b)$. The remaining parameters d_b^{first} and θ_b^{first} control the strength of the base distribution.

The back-off probability $P_{word-pos}(\alpha_{ij,pq} = x|B_{ij,pq} = b)$ is defined as below:

$$P_{word-pos}(\alpha_{ij,pq} = x|B_{ij,pq} = b)) = P_{word}(Rw(x)|b) \times P_{pos}(Rp(x)|b)$$

where $Rw(x)$ is the function returning a word sequence of a bilingual constituent x, $Rp(x)$ returning the correspondent POS sequence. This is the second level of the hierarchical representation of bilingual constituents as illustrated in Fig. 2. Further, $Rw(x)$ and $Rp(x)$ are decomposed into the third level of the hierarchy. Taking $Rw(x)$ for example:

$$P_{word}(Rw(x)|B_{ij,pq} = b)) = P_{word-bound}(Rwb(x)|b) \times \frac{1}{|W|^{|Rw(x)|-|Rwb(x)|}}$$

where Rwb is a function returning a word sequence's boundary representation, $|W|$ is the vocabulary size, $|Rw(x)| - |Rwb(x)|$ is the number of the words in $Rw(x)$ excluding those in the boundary representation. The above equation models the generation of a word sequence with surface string $Rw(x)$ (given b) by first generating its boundary representation $Rwb(x)$, then generating its middle words from a uniform distribution over the vocabulary. $P_{pos}(Rp(x)|B_{ij,pq} = b))$ is defined similarly.

We put the Dirichlet prior over $P_{word-bound}(Rwb(x)|b)$:

$$\begin{aligned} Rwb(x)|b &\sim Discrete(G_b^{Rwb}) \\ G_b^{Rwb} &\sim Dirichlet(\tau_b) \end{aligned}$$

For $P_{pos-bound}(Rpb(x)|b)$, similar definition to $P_{word-bound}(Rwb(x)|b)$ is applied.

3.2 The Dirichlet Prior over Bilingual Contexts

The Dirichlet prior is defined as below:

$$\begin{aligned} \beta_{ij,pq}|B_{ij,pq} = b &\sim Discrete(G_b^{Dir}) \\ G_b^{Dir} &\sim Dirichlet(\tau_b) \end{aligned}$$

A context $\beta_{ij,pq}$ (given the specific span type b) is drawn $i.i.d$ according to a multinomial parameter G_b^{Dir}, which is drawn from the Dirichlet distribution with a real value parameter τ_b.

4 MCMC Sampling for Inferring the Latent Synchronous Trees

We approximate the distribution over latent synchronous trees by sampling them from the posterior $P(T|S)$, where T is a latent synchronous tree of a sentence pair S. As presented in the beginning of section 3, the posterior depends on $P(\alpha_{ij,pq}|B_{ij,pq} = b)$ and $P(\beta_{ij,pq}|B_{ij,pq} = b)$, on which we put the PYP prior and the Dirichlet prior respectively. Because of integrating out all Gs in all of the priors, interdependency between samples of $\alpha_{ij,pq}|B_{ij,pq} = b$ or $\beta_{ij,pq}|B_{ij,pq} = b$ is introduced, resulting in simultaneously obtaining multiple samples impractical. On the other hand, blocked sampling, which obtains sentence-level samples simultaneously (Blunsom and Cohn, 2010; Cohn et al., 2010; Johnson et al., 2007a) is attractive for the fast mixing speed and the easy application of standard dynamic programming algorithms.

4.1 Metropolis-Hastings (MH) Sampler

We apply a MH sampler similar to (Johnson et al., 2007a) to overcome the difficulty of obtaining multiple samples simultaneously from posterior. The MH sampler is a MCMC technique that draws samples from a true distribution by first drawing samples simultaneously from a proposal distribution, and then correcting the samples to the true distribution by using an accept/reject test. In practical, the proposal distribution is designed to facilitate the use of blocked sampling that applies standard dynamic programming, and the resulting samples are corrected by the accept/reject test to the true distribution.

In our case, the proposal distribution is the Maximum-a-Posteriori (MAP) estimate of $P(\alpha_{i,j}|B_{i,j} = b)$ and $P(\beta_{i,j}|B_{i,j} = b)$, and the blocked sampling of T applies a dynamic programming algorithm that is based on the inside chart derived from a transformation of Eq. (1):

$$\begin{aligned} P(S,T) &= K(S) \prod_{<i,j><p,q>\in T} \phi(ij,pq) \\ where\ \phi(ij,pq) &= \frac{P(\alpha_{ij,pq}|B_{ij,pq}=1)P(\beta_{ij,pq}|B_{ij,pq}=1)}{P(\alpha_{ij,pq}|B_{ij,pq}=0)P(\beta_{ij,pq}|B_{ij,pq}=0)} \end{aligned}$$

$K(S)$ is a constant given S. The inside chart I can be constructed recursively as below:

$$I_{ij,pq} = \begin{cases} \phi(ij,pq) & if\ j-i \leq 1\ and\ q-p \leq 1 \\ \phi(ij,pq) \sum_{\substack{i\leq u\leq j \\ p\leq v\leq q}} (I_{iu,pv} I_{uj,vq} + I_{iu,vq} I_{uj,pv}) & otherwise \end{cases}$$

Based on this inside chart, a synchronous tree can be top-down sampled (Johnson et al., 2007a), then is accepted or rejected by the MH-test to correct to the true distribution.

5 Experiments

The experiments were conducted on both a pilot word alignment task and an end-to-end Chinese-to-English machine translation task to test the quality of the learned synchronous structures by the SCCM. The bi-side monolingual gold bracketings contained in Penn treebanks were not used for evaluating the quality of the learned synchronous structures because of great syntactic divergence between source tree and target tree, which results in that gold monolingual syntactic trees on both sides are asynchronous (large number of tree nodes can not be aligned). The synchronous grammar induction community assumes the existence of synchronous grammar for MT, and do not evaluate synchronous grammar induction on monolingual gold treebanks because of their asynchronous property. The synchronous grammar induction community is not the same with the multilingual grammar induction community, which targets at inducing bi-side monolingual syntactic trees. Due to the same reason, our synchronous bracketing induction method was not evaluated on bi-side monolingual bracketing trees which are asynchronous.

5.1 Sampler Configuration

Our sampler was initialised with trees through a random split process. Firstly, we used GIZA++ model 4 to get source-to-target and target-to-source word alignments, and used grow-diag-final-and (gdfa) heuristic to extract reliable word alignments for each sentence pair. Secondly, we randomly split each sentence pair in a top-down manner, and make sure that each split is consistent with the GIZA++ gdfa word alignments. For example, given a sentence pair of m source words and n target words, we randomly choose a split point at each side and the alignment type (straight alignment or inverted alignment), then recursively build bi-spans further on each new split. Finally, a synchronous binary tree is built at the end of this process [1]. Note that all splits must be consistent with the GIZA++ gdfa word alignments. When a piece of word alignments (such as non-ITG alignment structure) do not permit binary split, we keep this structure unsplitted and continue split only on its sub-structures that are ITG derivable.

Our sampler ran 200 iterations for all data. After each sampling iteration, we resample all the hyperparameters using slice-sampling, with the following priors: $d \sim Beta(1,1)$, $\theta \sim Gamma(10, 0.1)$.

The time complexity of our inference algorithm is $O(n^6)$, which is not practical in applications. We reduce the time complexity by only considering bi-spans that do not violate GIZA++ intersection word alignments (intersection of source-to-target and target-to-source word alignments) (Cohn and Haffari, 2013).

5.2 Word Alignment Task

5.2.1 Experimental Setting

Since there are no annotated synchronous treebanks, we evaluate the SCCM indirectly by evaluating its output word alignments on a gold standard English Chinese parallel tree bank with hand aligned word alignments referred as HIT corpus[2]. The HIT corpus, which was collected from English learning text books in China as well as example sentences in dictionaries, was originally designed for annotating bilingual tree node alignments. The annotation strictly reserves the semantic equivalence of the aligned sub-tree pair. The byproduct of this corpus is the hand aligned word alignments, which was utilized to evaluate word alignment performance[3]. The word segmentation, tokenization and parse-tree in the corpus were manually constructed or checked. The statistics of the HIT corpus are shown in table 1.

Table 1: Corpus statistics of the HIT corpus.

	ch	en
sent	16131	
word	210k	209k
avg. len.	13.06	13.0

5.2.2 Results

We adopt the commonly used metric: the alignment error rate (AER) to evaluate our proposed alignments (a) against hand-annotated alignments, which are marked with sure (s) and possible (p) alignments. The AER is given by (the lower the better):

$$AER(a, s, p) = 1 - \frac{|a \cap s| + |a \cap p|}{|a| + |s|}$$

In the HIT corpus, only sure alignments were annotated, possible alignments were bypassed because of the strict annotation standard of semantic equivalence.

The word alignments evaluation results are reported in Table 2. The baseline was GIZA++ model 4 in both directions with symmetrization by the grow-diag-final-and heuristic (Koehn et al., 2003). A

[1] The initialization with different random split bi-trees results in marginal variance of performances.

[2] HIT corpus is designed and constructed by HIT-MITLAB. http://mitlab.hit.edu.cn/index.php/resources.html

[3] We did not use annotated tree node alignments for synchronous structure evaluation because the coverage of tree nodes that can be aligned is quite low. The reason of low coverage is that Chinese and English exhibit great syntax divergences from monolingual treebank point of view.

released induction system - PIALIGN (Neubig et al., 2011)[4] was also experimented to compare with our proposed induction system - SCCM.

PIALIGN is a model that generalizes adaptor grammars for machine translation (MT), while our model is to generalize CCM for MT. Adaptor grammars has been successfully applied on shallow unsupervised tasks such as morphlogical/word analysis, while CCM has obtained state-of-the-art performance on the more complex unsupervised task - inducing syntactic trees. In view of CCM's successful monolingual application, we generalize it to bilingual case. In depth comparison: our SCCM deals with both consituents and distituents, and contexts of them, while PIALIGN only deals with constituents. Furthermore, SCCM does not model non-terminal rewriting rules, while PIALIGN model those rules which can rewrite a non-terminal into a complete subtree as adaptor grammars does. In addition, PIALIGN adopts a beam search algorithm of (Saers et al., 2009). Through setting small beam size, PIALIGN's time complexity is almost $O(n^3)$. But as critisized by (Cohn and Haffari, 2013), their heuristic beam search algorithm does not meet either of the Markov Chain Monte Carlo (MCMC) criteria of ergodicity or detailed balance. Our method adopts MCMC sampling (Johnson et al., 2007a) which meets the MCMC criteria.

We can see that the two induction systems perform significantly better than GIZA++, and our proposed SCCM performs better than PIALIGN. Manual evaluation for the quality of the phrase pairs generated from word alignments is also reported in Table 2. We considered the top-100 high frequency phrase pairs that are beyond word level and less than six words on both sides, and report the proportion of reasonably well phrase pairs through manual check. We found that more good phrase pairs can be derived from the SCCM's word alignments than from others.

Table 2: Quality of word alignments and their generated phrase pairs.

	AER	good phrase pairs proportion
GIZA++	0.322	0.493
PIALIGN	0.263	0.531
SCCM	0.255	0.534

5.3 Machine Translation Task

5.3.1 Experimental Setting

A released tourism-related domain machine translation data was used in our experiment. It consists of a parallel corpus extracted from the $Basic\ Travel\ Expression\ Corpus$ (BTEC), which had been used in evaluation campaigns of the yearly International Workshop on Spoken Language Translation (IWSLT). Table 3 lists statistics of the corpus used in the experiment.

Table 3: Statistics of the corpus used by IWSLT

	ch	en
sent	23k	
word	190k	213k
avg. len.	8.3	9.2

We used CSTAR03 as development set, used IWSLT04 and IWSLT05 official test set for test. A 4-gram language model with modified Kneser-Ney smoothing was trained on English side of parallel corpus. We use minimum error rate training (Och, 2003) with nbest list size 100 to optimize the feature weights for maximum development BLEU. Experimental results were evaluated by case-insensitive BLEU-4 (Papineni et al., 2001). Closest reference sentence length was used for brevity penalty.

5.3.2 Results

Following (Levenberg et al., 2012; Neubig et al., 2011; Cohn and Haffari, 2013), we evaluate our model by using the SCCM's output word alignments to construct a phrase table. As a baseline, we train a phrase-based model using the moses toolkit [5] based on the word alignments obtained using GIZA++

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

[5] http://www.statmt.org/moses

model 4 in both directions and symmetrized using the grow-diag-final-and heuristic (Koehn et al., 2003). For comparison with CFG-based induction systems, word alignments generated by the PIALIGN were also used to train a phrase-based model.

In the end-to-end MT evaluation, we used the standard set of features: relative-frequency and lexical translation model probabilities in both directions; distance-based distortion model; language model and word count. The evaluation results are reported in table 4. Word alignments derived by the two induction systems can be more helpful to obtain better translations than GIZA++ derived word alignments. The SCCM, while departing from traditional CFG-based methods, achieves comparable translation performance to the PIALIGN.

Table 4: BLEU on both the development set: CSTAR03, and the two test sets: IWSLT04 and IWSLT05.

	CSTAR03	IWSLT04	IWSLT05
GIZA++	0.4304	0.4190	0.4866
PIALIGN	0.4661	0.4556	0.5248
SCCM	0.4560	0.4469	0.5193

6 Conclusion

A new model for synchronous structure induction is proposed in this paper. Different to all the previous works that are based on Context Free Grammars, our proposed SCCM deals with bilingual constituents and contexts explicitely so that bilingual translational equivalences can be directly modeled. A non-parametric Bayesian modeling of the SCCM is applied to cope with the sparse representations of bilingual constituents and contexts. Both intrinsic evaluation on word alignments and extrinsic evaluation on end-to-end machine translation were conducted. The intrinsic evaluation show that the highest quality word alignments were obtained by our proposed SCCM. Such statistically sound word alignments of the SCCM were used in the extrinsic evaluation on machine translation, showing that significantly better translations were achieved than those obtained by using the word alignments of GIZA++, the widely used word aligner in the two-step pipeline.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant No. 61273319, and grant No. 61373095. Thanks for the helpful advices of anonymous reviewers.

References

Phil Blunsom and Trevor Cohn. 2010. Unsupervised induction of tree substitution grammars for dependency parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Shay B Cohen, Kevin Gimpel, and Noah A Smith. 2008. Logistic normal priors for unsupervised probabilistic grammar induction. In *Proceedings of the Advances in Neural Information Processing Systems*.

Trevor Cohn and Phil Blunsom. 2009. A bayesian model of syntax-directed tree to string grammar induction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 352–361. Association for Computational Linguistics.

Trevor Cohn and Gholamreza Haffari. 2013. An infinite hierarchical bayesian model of phrasal translation. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*.

Trevor Cohn, Phil Blunsom, and Sharon Goldwater. 2010. Inducing tree-substitution grammars. *Journal of Machine Learning Research*, 11:3053–3096.

Xiangyu Duan, Zhang Min, and Chen Wenliang. 2013. Smoothing for bracketing induction. In *Proceedings of 23rd International Joint Conference on Artificial Intelligence*. AAAI Press/International Joint Conferences on Artificial Intelligence.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2007. The infinite tree. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 272.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12:2335–2382.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007a. Bayesian inference for pcfgs via markov chain monte carlo. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146.

Mark Johnson, Thomas L Griffiths, and Sharon Goldwater. 2007b. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Proceedings of Advances in neural information processing systems*, 19:641.

Dan Klein and Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 128–135. Association for Computational Linguistics.

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*, pages 48–54. Association for Computational Linguistics.

Abby Levenberg, Chris Dyer, and Phil Blunsom. 2012. A bayesian model for learning scfgs with discontiguous rules. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 223–232. Association for Computational Linguistics.

P. Liang, S. Petrov, M. I. Jordan, and D. Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

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-Volume 1*, pages 632–641. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Markus Saers, Joakim Nivre, and Dekai Wu. 2009. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 29–32. Association for Computational Linguistics.

Xinyan Xiao and Deyi Xiong. 2013. Max-margin synchronous grammar induction for machine translation. In *EMNLP*.

Xinyan Xiao, Deyi Xiong, Yang Liu, Qun Liu, and Shouxun Lin. 2012. Unsupervised discriminative induction of synchronous grammar for machine translation. In *COLING*, pages 2883–2898.

Teh YeeWhye. 2006. A bayesian interpretation of interpolated kneser-ney. In *Technical Report TRA2/06*. School of Computing, National University of Singapore.