Learning Probabilistic Synchronous CFGs for Phrase-based Translation

Markos Mylonakis
ILLC
University of Amsterdam
m.mylonakis@uva.nl

Khalil Sima'an
ILLC
University of Amsterdam
k.simaan@uva.nl

Abstract

Probabilistic phrase-based synchronous grammars are now considered promising devices for statistical machine translation because they can express reordering phenomena between pairs of languages. Learning these hierarchical, probabilistic devices from parallel corpora constitutes a major challenge, because of multiple latent model variables as well as the risk of data overfitting. This paper presents an effective method for learning a family of particular interest to MT, binary Synchronous Context-Free Grammars with inverted/monotone orientation (a.k.a. Binary ITG). A second contribution concerns devising a lexicalized phrase reordering mechanism that has complimentary strengths to Chiang's model. The latter conditions reordering decisions on the surrounding lexical context of phrases, whereas our mechanism works with the lexical content of phrase pairs (akin to standard phrase-based systems). Surprisingly, our experiments on French-English data show that our learning method applied to far simpler models exhibits performance indistinguishable from the Hiero system.

1 Introduction

A fundamental problem in phrase-based machine translation concerns the learning of a probabilistic synchronous context-free grammar (SCFG) over phrase pairs from an input parallel corpus. Chiang's Hiero system (Chiang, 2007) exemplifies the gains to be had by combining phrase-based translation (Och and Ney, 2004) with the hierarchical reordering capabilities of SCFGs, particularly originating from Binary Inversion Transduction Grammars (BITG) (Wu, 1997). Yet, existing empirical work is largely based on successful heuristic techniques, and the learning of Hiero-like BITG/SCFG remains an unsolved problem,

The difficulty of this problem stems from the need for simultaneously learning of two kinds of preferences (see Fig.1) (1) lexical translation probabilities ($P(\langle e, f\rangle \mid X)$) of source ($f$) and target ($e$) phrase pairs, and (2) phrase reordering preferences of a target string relative to a source string, expressed in synchronous productions probabilities (for monotone or switching productions). Theoretically speaking, both kinds of preferences may involve latent structure relative to the parallel corpus. The mapping between source-target sentence pairs can be expressed in terms of latent phrase segmentations and latent word/phrase-alignments, and the hierarchical phrase reordering can be expressed in terms of latent binary synchronous hierarchical structures (cf. Fig. 1). But each of these three kinds of latent structures may be made explicit using external resources: word-alignment could be considered solved using Giza++ (Och and Ney, 2003)), phrase pairs can be obtained from these word-alignments (Och and Ney, 2004), and the hierarchical synchronous structure can be grown over source/target linguistic syntactic trees output by an existing parser.

The Joint Phrase Translation Model (Marcu and Wong, 2002) constitutes a specific case, albeit without the hierarchical, synchronous reordering

Start	$S \rightarrow X_{\boxed{1}} \,/\, X_{\boxed{1}}$	(1)
Monotone	$X \rightarrow X_{\boxed{1}}\, X_{\boxed{2}} \,/ X_{\boxed{1}}\, X_{\boxed{2}}$	(2)
Switching	$X \rightarrow X_{\boxed{1}}\, X_{\boxed{2}} \,/ X_{\boxed{2}}\, X_{\boxed{1}}$	(3)
Emission	$X \rightarrow e \,/\, f$	(4)

Figure 1: A phrase-pair SCFG (BITG)

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 117–125,
Uppsala, Sweden, 15-16 July 2010.

component. Other existing work, e.g. (Chiang, 2007), assumes the word-alignments are given in the parallel corpus, but the problem of learning phrase translation probabilities is usually avoided by using surface counts of phrase pairs (Koehn et al., 2003). The problem of learning the hierarchical, synchronous grammar reordering rules is oftentimes addressed as a learning problem in its own right assuming all the rest is given (Blunsom et al., 2008b).

A small number of efforts has been dedicated to the simultaneous learning of the probabilities of phrase translation pairs as well as hierarchical reordering, e.g., (DeNero et al., 2008; Zhang et al., 2008; Blunsom et al., 2009). Of these, some concentrate on evaluating word-alignment, directly such as (Zhang et al., 2008) or indirectly by evaluating a heuristically trained hierarchical translation system from sampled phrasal alignments (Blunsom et al., 2009). However, very few evaluate on actual translation performance of induced synchronous grammars (DeNero et al., 2008). In the majority of cases, the Hiero system, which constitutes the yardstick by which hierarchical systems are measured, remains superior in translation performance, see e.g. (DeNero et al., 2008).

This paper tackles the problem of learning *generative BITG models* as translation models assuming latent segmentation and latent reordering: this is the most similar setting to the training of Hiero. Unlike all other work that heuristically selects a subset of phrase pairs, we start out from an SCFG that works with *all* phrase pairs in the training set and concentrate on the aspects of learning. This learning problem is fraught with the risks of overfitting and can easily result in inadequate reordering preferences (see e.g. (DeNero et al., 2006)).

Almost instantly, we find that the translation performance of all-phrase probabilistic SCFGs learned in this setting crucially depends on the interplay between two aspects of learning:

- Defining a more constrained parameter space, where the reordering productions are phrase-lexicalised and made sensitive to neighbouring reorderings, and
- Defining an objective function that effectively smoothes the maximum-likelihood criterion.

One contribution of this paper is in devising an effective, data-driven smoothed Maximum-Likelihood that can cope with a model working with *all* phrase pair SCFGs. This builds upon our previous work on estimating parameters of a "bag-of-phrases" model for Machine Translation (Mylonakis and Sima'an, 2008). However, learning SCFGs poses significant novel challenges, the core of which lies on the hierarchical nature of a stochastic SCFG translation model and the relevant additional layer of latent structure. We address these issues in this work. Another important contribution is in defining a lexicalised reordering component within BITG that captures order divergences orthogonal to Chiang's model (Chiang, 2007) but somewhat akin to Phrase-Based Statistical Machine Translation reordering models (Koehn et al., 2003).

Our analysis shows that the learning difficulties can be attributed to a rather weak generative model. Yet, our best system exhibits Hiero-level performance on French-English Europarl data using an SCFG-based decoder (Li et al., 2009). Our findings should be insightful for others attempting to make the leap from shallow phrase-based systems to hierarchical SCFG-based translation models using learning methods, as opposed to heuristics.

The rest of the paper is structured as follows. Section 2 briefly introduces the SCFG formalism and discusses its adoption in the context of Statistical Machine Translation (SMT). In section 3, we consider some of the pitfalls of stochastic SCFG grammar learning and address them by introducing a novel learning objective and algorithm. In the section that follows we browse through latent translation structure choices, while in section 5 we present our empirical experiments on evaluating the induced stochastic SCFGs on a translation task and compare their performance with a hierarchical translation baseline. We close with a comparison of related work and a final discussion including future research directions.

2 Synchronous Grammars for Machine Translation

Synchronous Context Free Grammars (SCFGs) provide an appealing formalism to describe the translation process, which explains the generation of parallel strings recursively and allows capturing long-range reordering phenomena. Formally, an SCFG $\mathbf{G}$ is defined as the tuple (N, E, F, R, S),

where N is the finite set of non-terminals with $S \in N$ the start symbol, F and E are finite sets of words for the source and target language and R is a finite set of rewrite rules. Every rule expands a left-hand side non-terminal to a right-hand side pair of strings, a source language string over the vocabulary $F \cup N$ and a target language string over $E \cup N$. The number of non-terminals in the two strings is equal and the rule is complemented with a mapping between them.

String pairs in the language of the SCFG are those with a valid derivation, consisting of a sequence of rule applications, starting from S and recursively expanding the linked non-terminals at the right-hand side of rules. *Stochastic* SCFGs augment every rule in R with a probability, under the constraint that probabilities of rules with the same left-hand side sum up to one. The probability of each derived string pair is then the product of the probabilities of rules used in the derivation. Unless otherwise stated, for the rest of the paper when we refer to SCFGs we will be pointing to their stochastic extension.

The *rank* of an SCFG is defined as the maximum number of non-terminals in a grammar's rule right-hand side. Contrary to monolingual Context Free Grammars, there does not always exist a conversion of an SCFG of a higher rank to one of a lower rank with the same language of string pairs. For this, most machine translation applications focus on SCFGs of rank two (binary SCFGs), or *binarisable* ones witch can be converted to a binary SCFG, given that these seem to cover most of the translation phenomena encountered in language pairs (Wu, 1997) and the related processing algorithms are less demanding computationally.

Although SCFGS were initially introduced for machine translation as a stochastic *word-based* translation process in the form of the Inversion-Transduction Grammar (Wu, 1997), they were actually able to offer state-of-the-art performance in their latter *phrase-based* implementation by Chiang (Chiang, 2005). Chiang's Hiero hierarchical translation system is based on a synchronous grammar with a single non-terminal X covering all learned phrase-pairs. Beginning from the start symbol S, an initial phrase-span structure is constructed monotonically using a simple 'glue grammar':

$$S \rightarrow S_{\boxed{1}}\ X_{\boxed{2}}\ /\ S_{\boxed{1}}\ X_{\boxed{2}}$$
$$S \rightarrow X_{\boxed{1}}\ /\ X_{\boxed{1}}$$

The true power of the system lies in expanding these initial phrase-spans with a set of hierarchical translation rules, which allow conditioning reordering decisions based on lexical context. For the French to English language pair, some examples would be:

$$S \rightarrow \quad X_{\boxed{1}}\ économiques\ /\ financial\ X_{\boxed{1}}$$
$$S \rightarrow \quad cette\ X_{\boxed{1}}\ de\ X_{\boxed{2}}\ /\ this\ X_{\boxed{1}}\ X_{\boxed{2}}$$
$$S \rightarrow \quad politique\ X_{\boxed{1}}\ commune\ de\ X_{\boxed{2}}\ /\ X_{\boxed{2}}\ '\ s\ common\ X_{\boxed{1}}\ policy$$

Further work builds on the Hiero grammar to expand it with constituency syntax motivated non-terminals (Zollmann and Venugopal, 2006).

3 Synchronous Grammar Learning

The learning of phrase-based stochastic SCFGs with a Maximum Likelihood objective is exposed to overfitting as other *all-fragment models* such as Phrase-Based SMT (PBSMT) (Marcu and Wong, 2002; DeNero et al., 2006) and Data-Oriented Parsing (DOP) (Bod et al., 2003; Zollmann and Sima'an, 2006). Maximum Likelihood Estimation (MLE) returns degenerate grammar estimates that memorise well the parallel training corpus but generalise poorly to unseen data.

The bias-variance decomposition of the generalisation error Err sheds light on this learning problem. For an estimator $\hat{p}$ with training data $\mathcal{D}$, Err can be expressed as the expected Kullback-Leibler (KL) divergence between the target distribution q and that the estimate $\hat{p}$. This error decomposes into bias and variance terms (Heskes, 1998):

$$Err = \overbrace{KL(q, \bar{p})}^{bias} + \overbrace{E_{\mathcal{D}} KL(\bar{p}, \hat{p})}^{variance} \qquad (5)$$

Bias is the KL-divergence between q and the mean estimate over all training data $\bar{p} = E_{\mathcal{D}} \hat{p}(\mathcal{D})$. Variance is the expected divergence between the average estimate and the estimator's actual choice. MLE estimators for all-fragment models are *zero-biased* with zero divergence between the average estimate and the true data distribution. In contrast, their variance is unboundedly large, leading to unbounded generalisation error on unseen cases.

3.1 Cross Validated MLE

A well-known method for estimating generalisation error is k-fold *Cross-Validation* (CV) (Hastie et al., 2001). By partitioning the training data $\mathcal{D}$ into k parts H_1^k, we estimate Err as the expected error over all $1 \leq i \leq k$, when testing on H_i with a model trained by MLE on the rest of the data $D^{-i} = \cup_{j \neq i} H_j$.

Here we use CV to leverage the bias-variance trade-off for learning stochastic all-phrase SCFGs. Given an input all-phrase SCFG grammar with phrase-pairs extracted from the training data, we maximise training data likelihood (MLE) subject to CV smoothing: for each data part H_i ($1 \leq i \leq k$), we consider only derivations which employ grammar rules extracted from the rest of the data $\mathcal{D}^{-i}$. Other work (Mylonakis and Sima'an, 2008) has also explored MLE under CV for a "bag-of-phrases model" that does not deal with reordering preferences, does not employ latent hierarchical structure and works with a non-hierarchical decoder, and partially considers the sparsity issues that arise within CV training. The present paper deals with these issues.

Because of the latent segmentation and hierarchical variables, CV-smoothed MLE cannot be solved analytically and we devise a CV instance of the Expectation-Maximization (EM) algorithm, with an implementation based on a synchronous version of the Inside-Outside algorithm (see Fig. 2). For each word-aligned sentence pair in a partition H_i, the set of eligible derivations (denoted $\mathbf{D}^{-i}$) are those that can be built using only phrase-pairs and productions found in $\mathcal{D}^{-i}$. An essential part of the learning process involves defining the grammar extractor $G(\mathcal{D})$, a function from data to an all-phrase SCFG. We will discuss various extractors in section 4.

Our CV-EM algorithm is an EM instance, guaranteeing convergence and a non-decreasing CV-smoothed data likelihood after each iteration. The running time remains $O(n^6)$, where n is input length, but by considering only derivation spans which do not cross word-alignment points, this runs in reasonable times for relatively large corpora.

INPUT: Word-aligned parallel training data $\mathcal{D}$
Grammar extractor G
The number of parts k to partition $\mathcal{D}$
OUTPUT: SCFG $\mathbf{G}$ with rule probabilities $\hat{p}$

Partition training data $\mathcal{D}$ into parts $H_1, \ldots, H_k$.
For $1 \leq i \leq k$ **do**
  Extract grammar rules set $\mathbf{G}_i = G(H_i)$
Initialise $\mathbf{G} = \cup_i \mathbf{G}_i$, $\hat{p}_0$ uniform
Let $j = 0$
Repeat
  Let $j = j + 1$
  E-step:
    For $1 \leq i \leq k$ **do**
      Calculate expected counts given $\mathbf{G}$, $\hat{p}_{j-1}$,
        for derivations $\mathbf{D}^{-i}$ of H_i
        using rules from $\cup_{k \neq i} G(k)$
  M-step: set $\hat{p}_j$ to ML estimate given
      expected counts
Until convergence

Figure 2: The CV Expectation Maximization algorithm

3.2 Bayesian Aspects of CV-MLE

Beside being an estimator, the CV-MLE learning algorithm has the added value of being a grammar learner focusing on reducing generalisation error, in the sense that probabilities of grammar productions should reflect the frequency with which these productions are expected to be used for translating future data. Additionally, since the CV criterion prohibits for every data point derivations that use rules only extracted from the same data part, such rules are assigned zero probabilities in the final estimate and are effectively excluded from the grammar. In this way, the algorithm 'shapes' the input grammar, concentrating probability mass on productions that are likely to be used with future data.

One view point of CV-MLE is that each partition $\mathcal{D}^{-i}$ and H_i induces a prior probability $Prior(\pi;\, \mathcal{D}^{-i})$ on every parameter assignment π, obtained from $\mathcal{D}^{-i}$. This prior assigns zero probability to all π parameter sets with non-zero probabilities for rules not in $G(\mathcal{D}^{-i})$, and uniformly distributes probability to the rest of the parameter sets. In light of this, the CV-MLE objective can be written as follows:

$$\arg\max_{\pi} \prod_{i} Prior(\pi;\, \mathcal{D}^{-i}) \times P(H_i \mid \pi) \qquad (6)$$

This *data-driven* prior aims to directly favour parameter sets which are expected to better generalise according to the CV criterion, without relying on arbitrary constraints such as limiting the

length of phrase pairs in the right-hand side of grammar rules. Furthermore, other frequently employed priors such as the Dirichlet distribution and the Dirichlet Process promote better generalising rule probability distributions based on externally set hyperparameter values, whose selection is frequently sensitive in terms of language pairs, or even the training corpus itself. In contrast, the CV-MLE prior aims for a data-driven Bayesian model, focusing on getting information from the data, instead of imposing external human knowledge on them (see also (Mackay and Petoy, 1995)).

3.3 Smoothing the Model

One remaining wrinkle in the CV-EM scheme is the treatment of boundary cases. There will often be sentence-pairs in H_i, that cannot be fully derived by the grammar extracted from the rest of the data $\mathcal{D}^{-i}$ either because of (1) 'unknown' words (i.e. not appearing in other parts of the CV partition) or (2) complicated combinations of adjacent word-alignments. We employ external smoothing of the grammar, *prior* to learning.

Our solution is to extend the SCFG extracted from $\mathcal{D}^{-i}$ with new emission productions deriving the 'unknown' phrase-pairs (i.e., found in H_i but not in $\mathcal{D}^{-i}$). Crucially, the probabilities of these productions are drawn from a fixed smoothing distribution, i.e., they remain constant throughout estimation. Our smoothing distribution of phrase-pairs for all pre-terminals considers source-target phrase lengths drawn from a Poisson distribution with unit mean, drawing subsequently the words of each of the phrases uniformly from the vocabulary of each language, similar to (Blunsom et al., 2009).

$$p_{smooth}(f/e) = \frac{p_{poisson}(|f|;1)\ p_{poisson}(|e|;1)}{V_f^{|f|}\ V_e^{|e|}}$$

Since the smoothing distribution puts stronger preference on shorter phrase-pairs and avoids competing with the 'known' phrase-pairs, it leads the learner to prefer using as little as possible such smoothing rules, covering only the phrase-pairs required to complete full derivations.

4 Parameter Spaces and Grammar Extractors

A *Grammar Extractor* (GE) plays a major role in our probabilistic SCFG learning pipeline. A GE is a function from a word-aligned parallel corpus to a probabilistic SCFG model. Together with the constraints that render a proper probabilistic SCFG[1], this defines the parameter space.

The extractors used in this paper create SCFGs productions of two different kinds: (a) hierarchical synchronous productions that define the space of possible derivations up to the level of the SCFG pre-terminals, and (2) the phrase-pair emission rules that expand the pre-terminals to phrase-pairs of varying lengths. Given the word-alignments, the set of phrase-pairs extracted is the set of *all* translational equivalents (without length upper-bound) under the word-alignment as defined in (Och and Ney, 2004; Koehn et al., 2003).

Below we focus on the two grammar extractors employed in our experiments. We start out from the most generic, BITG-like formulation, and aim at incremental refinement of the hierarchical productions in order to capture relevant, content-based phrase-pair reordering preferences in the training data.

Single non-terminal SCFG This is a phrase-based binary SCFG grammar employing a single non-terminal X covering each extracted phrase-pair. The other productions consist of monotone and switching expansions of phrase-pair spans covered by X. Finally, the whole sentence-pair is considered to be covered by X. We will call this 'plain SCFG' extractor. See Fig. 1.

Lexicalised Reordering SCFG One weakness of the plain SCFG is that the reordering decisions in the derivations are made without reference to lexical content of the phrases; this is because all phrase-pairs are covered by the same non-terminal. As a refinement, we propose a grammar extractor that aims at modelling the reordering behaviour of phrase-pairs by taking their content into account. This time, the X non-terminal is reserved for phrase-pairs and spans which will take part in monotonic productions only. Two fresh non-terminals, XSL and XSR, are used for covering phrase-pairs that participate in order switching with other, adjacent phrase-pairs. The non-terminal XSL covers phrase-pairs which appear first in the source language order, and the latter those which follow them. The grammar rules produced by this GE, dubbed 'switch grammar', are listed in Fig. 3.

[1]The sum of productions that have the same left-hand label must be one.

Start $S \rightarrow X_{\boxed{1}} \,/ X_{\boxed{1}}$

Monotone Expansion

$X \rightarrow X_{\boxed{1}}\, X_{\boxed{2}} \,/ X_{\boxed{1}}\, X_{\boxed{2}}$
$XSL \rightarrow X_{\boxed{1}}\, X_{\boxed{2}} \,/\, X_{\boxed{1}}\, X_{\boxed{2}}$
$XSR \rightarrow X_{\boxed{1}}\, X_{\boxed{2}} \,/ X_{\boxed{1}}\, X_{\boxed{2}}$

Switching Expansion

$X \rightarrow XSL_{\boxed{1}}\, XSR_{\boxed{2}} \,/ XSR_{\boxed{2}}\, XSL_{\boxed{1}}$
$XSL \rightarrow XSL_{\boxed{1}}\, XSR_{\boxed{2}} \,/ XSR_{\boxed{2}}\, XSL_{\boxed{1}}$
$XSR \rightarrow XSL_{\boxed{1}}\, XSR_{\boxed{2}} \,/ XSR_{\boxed{2}}\, XSL_{\boxed{1}}$

Phrase-Pair Emission

$X \rightarrow e/f$
$XSL \rightarrow e \,/\, f$
$XSR \rightarrow e \,/\, f$

Figure 3: Lexicalised-Reordering SCFG

The reordering information captured by the switch grammar is in a sense orthogonal to that of Hiero-like systems utilising rules such as those listed in section 2. Hiero rules encode hierarchical reordering patterns based on surrounding context. In contrast, the switch grammar models the reordering preferences of the phrase-pairs themselves, similarly to the monotone-swap-discontinuous reordering models of Phrase-based SMT models (Koehn et al., 2003). Furthermore, it strives to match pairs of such preferences, combining together phrase-pairs with compatible reordering preferences.

5 Experiments

In this section we proceed to integrate our estimates within an SCFG-based decoder. We subsequently evaluate our performance in relation to a state-of-the-art Hiero baseline on a French to English translation task.

5.1 Decoding

The joint model of bilingual string derivations provided by the learned SCFG grammar can be used for translation given a input source sentence, since $\arg\max_e p(e|f) = \arg\max_e p(e, f)$. We use our learned stochastic SCFG grammar with the decoding component of the Joshua SCFG toolkit (Li et al., 2009). The full translation model interpolates log-linearly the probability of a grammar derivation together with the language model probability of the target string. The model is further smoothed, similarly to phrase-based models and the Hiero system, with smoothing features ϕ_i such as the lexical translation scores of the phrase-pairs involved and rule usage penalties. As usual with statistical translation, we aim for retrieving the target sentence e corresponding to the most probable derivation $D \stackrel{*}{\Rightarrow} (f, e)$ with rules r, with:

$$p(D) \propto p(e)^{\lambda_{lm}} p_{scfg}(e, f)^{\lambda_{scfg}} \prod_i \prod_{r \in D} \phi_i(r)^{\lambda_i}$$

The interpolation weights are tuned using Minimum Error Rate Training (Och, 2003).

5.2 Results

We test empirically the learner's output grammars for translating from French to English, using $k = 5$ for the Cross Validation data partitioning. The training material is a GIZA++ word-aligned corpus of 200K sentence-pairs from the Europarl corpus (Koehn, 2005), with our development and test parallel corpora of 2K sentence-pairs stemming from the same source. Training the grammar parameters until convergence demands around 6 hours on an 8-core 2.26 GHz Intel Xeon system. Decoding employs a 4-gram language model, trained on English Europarl data of 19.5M words, smoothed using modified Kneser-Ney discounting (Chen and Goodman, 1998), and lexical translation smoothing features based on the GIZA++ alignments.

In a sense, the real baseline to which we might compare against should be a system employing the MLE estimate for the grammar extracted from the whole training corpus. However, as we have already discussed, this assigns zero probability to all sentence-pairs outside of the training data and is subsequently bound to perform extremely poorly, as decoding would then completely rely on the smoothing features. Instead, we opt to compare against a hierarchical translation baseline provided by the Joshua toolkit, trained and tuned on the same data as our learning algorithm. The grammar used by the baseline is much richer than the ones learned by our algorithm, also employing rules which translate with context, as shown in section 2. Nevertheless, since it is not clear how the reordering rules probabilities of a grammar similar to the ones we use could be trained heuristically, we choose to relate the performance of our learned stochastic SCFG grammars to the particular, state-of-the-art in SCFG-based translation, system.

Table 1 presents the translation performance results of our systems and the baseline. On first

System	Lexical Smoothing	BLEU
joshua-baseline	No	27.79
plain scfg	No	28.04
switch scfg	No	28.48
joshua-baseline	Yes	29.96
plain scfg	Yes	29.75
switch scfg	Yes	29.88

Table 1: Empirical results, with and without additional lexical translation smoothing features during decoding

observation, it is evident that our learning algorithm outputs stochastic SCFGs which manage to generalise, avoiding the degenerate behaviour of plain MLE training for these models. Given the notoriety of the estimation process, this is noteworthy on its own. Having a learning algorithm at hand which realises in a reasonable extent the potential of each stochastic grammar design (as implemented in the relevant grammar extractors), we can now compare between the two grammar extractors used in our experiments. The results table highlights the importance of conditioning the reordering process on lexical grounds. The plain grammar with the single phrase-pair non-terminal cannot accomplish this and achieves a lower BLEU score. On the other hand, the switch SCFG allows such conditioning. The learner takes advantage of this feature to output a grammar which performs better in taking reordering decisions, something that is reflected in both the actual translations as well as the BLEU score achieved.

Furthermore, our results highlight the importance of the smoothing decoding features. The unsmoothed baseline system itself scores considerably less when employing solely the heuristic translation score. Our unsmoothed switch grammar decoding setup improves on the baseline by a considerable difference of 0.7 BLEU. Subsequently, when adding the smoothing lexical translation features, both systems record a significant increase in performance, reaching comparable levels of performance.

The degenerate behaviour of MLE for SCFGs can be greatly limited by constraining ourselves to grammars employing *minimal* phrase-pairs ; phrase-pairs which cannot be further broken down into smaller ones according to the word-alignment. One could argue that it is enough to perform plain MLE with such minimal phrase-pair SCFGs, instead of using our more elaborate learning algorithm with phrase-pairs of all lengths. To investigate this, for our final experiment we used a plain MLE estimate of the switch grammar to translate, limiting the grammar's phrase-pair emission rules to only those which involve minimal phrase-pairs. The very low score of 17.82 BLEU (without lexical smoothing) not only highlights the performance gains of using longer phrase-pairs in hierarchical translation models, but most importantly provides a strong incentive to address the overfitting behaviour of MLE estimators for such models, instead of avoiding it.

6 Related work

Most learning of phrase-based models, e.g., (Marcu and Wong, 2002; DeNero et al., 2006; Mylonakis and Sima'an, 2008), works without hierarchical components (i.e., not based on the explicit learning of an SCFG/BITG). These learning problems pose other kinds of learning challenges than the ones posed by explicit learning of SCFGs. Chiang's original work (Chiang, 2007) is also related. Yet, the learning problem is not expressed in terms of an explicit objective function because surface heuristic counts are used. It has been very difficult to match the performance of Chiang's model without use of these heuristic counts.

A somewhat related work, (Blunsom et al., 2008b), attempts learning new non-terminal labels for synchronous productions in order to improve translation. This work differs substantially from our work because it employs a heuristic estimate for the phrase pair probabilities, thereby concentrating on a different learning problem: that of refining the grammar symbols. Our approach might also benefit from such a refinement but we do not attempt this problem here. In contrast, (Blunsom et al., 2008a) works with the expanded phrase pair set of (Chiang, 2005), formulating an exponential model and concentrating on marginalising out the latent segmentation variables. Again, the learning problem is rather different from ours. Similarly, the work in (Zhang et al., 2008) reports on a multi-stage model, *without* a latent segmentation variable, but with a strong prior preferring sparse estimates embedded in a Variational Bayes (VB) estimator. This work concentrates the efforts on pruning both the space of phrase pairs and the space of (ITG) analyses.

To the best of our knowledge, this work is the first to attempt learning probabilistic phrase-based BITGs as translation models in a setting where both a phrase segmentation component and a hierarchical reordering component are assumed latent variables. Like this work, (Mylonakis and Sima'an, 2008; DeNero et al., 2008) also employ an all-phrases model. Our paper shows that it is possible to train such huge grammars under iterative schemes like CV-EM, without need for sampling or pruning. At the surface of it, our CV-EM estimator is also a kind of Bayesian learner, but in reality it is a more specific form of regularisation, similar to smoothing techniques used in language modelling (Chen and Goodman, 1998; Mackay and Petoy, 1995).

7 Discussion and Future Research

Phrase-based stochastic SCFGs provide a rich formalism to express translation phenomena, which has been shown to offer competitive performance in practice. Since learning SCFGs for machine translation has proven notoriously difficult, most successful SCFG models for SMT rely on rules extracted from word-alignment patterns and heuristically computed rule scores, with the impact and the limits imposed by these choices yet unknown.

Some of the reasons behind the challenges of SCFG learning can be traced back to the introduction of latent variables at different, competing levels: word and phrase-alignment as well as hierarchical reordering structure, with larger phrase-pairs reducing the need for extensive reordering structure and vice versa. While imposing priors such as the often used Dirichlet distribution or the Dirichlet Process provides a method to overcome these pitfalls, we believe that the data-driven regularisation employed in this work provides an effective alternative to them, focusing more on the data instead of importing generic external human knowledge.

We believe that this work makes a significant step towards learning synchronous grammars for SMT. This is an objective not only worthy because of promises of increased performance, but, most importantly, also by increasing the depth of our understanding on SCFGs as vehicles of latent translation structures. Our usage of the induced grammars directly for translation, instead of an intermediate task such as phrase-alignment, aims exactly at this.

While the latent structures that we explored in this paper were relatively simple in comparison with Hiero-like SCFGs, they take a different, content-driven approach on learning reordering preferences than the context-driven approach of Hiero. We believe that these approaches are not merely orthogonal, but could also prove complementary. Taking advantage of the possible synergies between content and context-driven reordering learning is an appealing direction of future research. This is particularly promising for other language pairs, such as Chinese to English, where Hiero-like grammars have been shown to perform particularly well.

Acknowledgments: Both authors are supported by a VIDI grant (nr. 639.022.604) from The Netherlands Organization for Scientific Research (NWO).

References

P. Blunsom, T. Cohn, and M. Osborne. 2008a. A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 200–208. Association for Computational Linguistics.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008b. Bayesian synchronous grammar induction. In *Advances in Neural Information Processing Systems 21*, Vancouver, Canada, December.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics*, Singapore, August. Association for Computational Linguistics.

R. Bod, R. Scha, and K. Sima'an, editors. 2003. *Data Oriented Parsing*. CSLI Publications, Stanford University, Stanford, California, USA.

S. Chen and J. Goodman. 1998. *An empirical study of smoothing techniques for language modeling*. Technical Report TR-10-98, Harvard University, August.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33.

J. DeNero, D. Gillick, J. Zhang, and D. Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 31–38, New York City. Association for Computational Linguistics.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 314–323, Honolulu, Hawaii, October. Association for Computational Linguistics.

T. Hastie, R. Tibshirani, and J. H. Friedman. 2001. *The Elements of Statistical Learning*. Springer.

Tom Heskes. 1998. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10:1425–1433.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003*.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

David J. C. Mackay and Linda C. Bauman Petoy. 1995. A hierarchical dirichlet language model. *Natural Language Engineering*, 1:1–19.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of Empirical methods in natural language processing*, pages 133–139. Association for Computational Linguistics.

Markos Mylonakis and Khalil Sima'an. 2008. Phrase translation probabilities with itg priors and smoothing as learning objective. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 630–639, Honolulu, USA, October.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

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

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

H. Zhang, Ch. Quirk, R. C. Moore, and D. Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proceedings of ACL-08: HLT*, pages 97–105, Columbus, Ohio, June. Association for Computational Linguistics.

A. Zollmann and K. Sima'an. 2006. An efficient and consistent estimator for data-oriented parsing. *Journal of Automata, Languages and Combinatorics (JALC)*, 10 (2005) Number 2/3:367–388.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City, June. Association for Computational Linguistics.