Soft Syntactic Constraints for Hierarchical Phrased-Based Translation

Yuval Marton and Philip Resnik
Department of Linguistics
and the Laboratory for Computational Linguistics and Information Processing (CLIP)
at the Institute for Advanced Computer Studies (UMIACS)
University of Maryland, College Park, MD 20742-7505, USA
{ymarton, resnik} @t umiacs.umd.edu

Abstract

In adding syntax to statistical MT, there is a tradeoff between taking advantage of linguistic analysis, versus allowing the model to exploit linguistically unmotivated mappings learned from parallel training data. A number of previous efforts have tackled this tradeoff by starting with a commitment to linguistically motivated analyses and then finding appropriate ways to soften that commitment. We present an approach that explores the tradeoff from the other direction, starting with a context-free translation model learned directly from aligned parallel text, and then adding soft constituent-level constraints based on parses of the source language. We obtain substantial improvements in performance for translation from Chinese and Arabic to English.

1 Introduction

The statistical revolution in machine translation, beginning with (Brown et al., 1993) in the early 1990s, replaced an earlier era of detailed language analysis with automatic learning of shallow source-target mappings from large parallel corpora. Over the last several years, however, the pendulum has begun to swing back in the other direction, with researchers exploring a variety of statistical models that take advantage of source- and particularly target-language syntactic analysis (e.g. (Cowan et al., 2006; Zollmann and Venugopal, 2006; Marcu et al., 2006; Galley et al., 2006) and numerous others).

Chiang (2005) distinguishes statistical MT approaches that are "syntactic" in a *formal* sense, going beyond the finite-state underpinnings of phrase-based models, from approaches that are syntactic in a *linguistic* sense, i.e. taking advantage of *a priori* language knowledge in the form of annotations derived from human linguistic analysis or treebanking.[1] The two forms of syntactic modeling are doubly dissociable: current research frameworks include systems that are finite state but informed by linguistic annotation prior to training (e.g., (Koehn and Hoang, 2007; Birch et al., 2007; Hassan et al., 2007)), and also include systems employing context-free models trained on parallel text without benefit of any prior linguistic analysis (e.g. (Chiang, 2005; Chiang, 2007; Wu, 1997)). Over time, however, there has been increasing movement in the direction of systems that are syntactic in both the formal and linguistic senses.

In any such system, there is a natural tension between taking advantage of the linguistic analysis, versus allowing the model to use linguistically unmotivated mappings learned from parallel training data. The tradeoff often involves starting with a system that exploits rich linguistic representations and relaxing some part of it. For example, DeNeefe et al. (2007) begin with a tree-to-string model, using treebank-based target language analysis, and find it useful to modify it in order to accommodate useful "phrasal" chunks that are present in parallel training data but not licensed by linguistically motivated parses of the target language. Similarly, Cowan et al. (2006) focus on using syntactically rich representations of source and target parse trees, but they resort to phrase-based translation for modifiers within

[1] See (Lopez, to appear) for a comprehensive survey.

Proceedings of ACL-08: HLT, pages 1003–1011,
Columbus, Ohio, USA, June 2008.

clauses. Finding the right way to balance linguistic analysis with unconstrained data-driven modeling is clearly a key challenge.

In this paper we address this challenge from a less explored direction. Rather than starting with a system based on linguistically motivated parse trees, we begin with a model that is syntactic only in the formal sense. We then introduce soft constraints that take source-language parses into account to a limited extent. Introducing syntactic constraints in this restricted way allows us to take maximal advantage of what can be learned from parallel training data, while effectively factoring in key aspects of linguistically motivated analysis. As a result, we obtain substantial improvements in performance for both Chinese-English and Arabic-English translation.

In Section 2, we briefly review the Hiero statistical MT framework (Chiang, 2005, 2007), upon which this work builds, and we discuss Chiang's initial effort to incorporate soft source-language constituency constraints for Chinese-English translation. In Section 3, we suggest that an insufficiently fine-grained view of constituency constraints was responsible for Chiang's lack of strong results, and introduce finer grained constraints into the model. Section 4 demonstrates the the value of these constraints via substantial improvements in Chinese-English translation performance, and extends the approach to Arabic-English. Section 5 discusses the results, and Section 6 considers related work. Finally we conclude in Section 7 with a summary and potential directions for future work.

2 Hierarchical Phrase-based Translation

2.1 Hiero

Hiero (Chiang, 2005; Chiang, 2007) is a hierarchical phrase-based statistical MT framework that generalizes phrase-based models by permitting phrases with gaps. Formally, Hiero's translation model is a weighted synchronous context-free grammar. Hiero employs a generalization of the standard non-hierarchical phrase extraction approach in order to acquire the synchronous rules of the grammar directly from word-aligned parallel text. Rules have the form $X \rightarrow \langle \bar{e}, \bar{f} \rangle$, where $\bar{e}$ and $\bar{f}$ are phrases containing terminal symbols (words) and possibly co-indexed instances of the nonterminal symbol X.[2] Associated with each rule is a set of translation model features, $\phi_i(\bar{f}, \bar{e})$; for example, one intuitively natural feature of a rule is the phrase translation (log-)probability $\phi(\bar{f}, \bar{e}) = \log p(\bar{e}|\bar{f})$, directly analogous to the corresponding feature in non-hierarchical phrase-based models like Pharaoh (Koehn et al., 2003). In addition to this phrase translation probability feature, Hiero's feature set includes the inverse phrase translation probability $\log p(\bar{f}|\bar{e})$, lexical weights lexwt($\bar{f}|\bar{e}$) and lexwt($\bar{e}|\bar{f}$), which are estimates of translation quality based on word-level correspondences (Koehn et al., 2003), and a rule penalty allowing the model to learn a preference for longer or shorter derivations; see (Chiang, 2007) for details.

These features are combined using a log-linear model, with each synchronous rule contributing

$$\sum_i \lambda_i \phi_i(\bar{f}, \bar{e}) \quad (1)$$

to the total log-probability of a derived hypothesis. Each λ_i is a weight associated with feature ϕ_i, and these weights are typically optimized using minimum error rate training (Och, 2003).

2.2 Soft Syntactic Constraints

When looking at Hiero rules, which are acquired automatically by the model from parallel text, it is easy to find many cases that seem to respect linguistically motivated boundaries. For example,

$$X \rightarrow \langle \text{jingtian } \mathrm{X}_1, \mathrm{X}_1 \text{ this year} \rangle,$$

seems to capture the use of *jingtian/this year* as a temporal modifier when building linguistic constituents such as noun phrases (*the election this year*) or verb phrases (*voted in the primary this year*). However, it is important to observe that nothing in the Hiero framework actually *requires* nonterminal symbols to cover linguistically sensible constituents, and in practice they frequently do not.[3]

[2] This is slightly simplified: Chiang's original formulation of Hiero, which we use, has two nonterminal symbols, X and S. The latter is used only in two special "glue" rules that permit complete trees to be constructed via concatenation of subtrees when there is no better way to combine them.

[3] For example, this rule could just as well be applied with X_1 covering the "phrase" *submitted and* to produce non-constituent substring *submitted and this year* in a hypothesis like *The budget was submitted and this year cuts are likely.*

Chiang (2005) conjectured that there might be value in allowing the Hiero model to favor hypotheses for which the synchronous derivation respects linguistically motivated source-language constituency boundaries, as identified using a parser. He tested this conjecture by adding a soft constraint in the form of a "constituency feature": if a synchronous rule $X \rightarrow \langle \bar{e}, \bar{f} \rangle$ is used in a derivation, and the span of $\bar{f}$ is a constituent in the source-language parse, then a term λ_c is added to the model score in expression (1).[4] Unlike a hard constraint, which would simply prevent the application of rules violating syntactic boundaries, using the feature to introduce a *soft* constraint allows the model to boost the "goodness" for a rule if it is constitent with the source language constituency analysis, and to leave its score unchanged otherwise. The weight λ_c, like all other λ_i, is set via minimum error rate training, and that optimization process determines empirically the extent to which the constituency feature should be trusted.

Figure 1 illustrates the way the constituency feature worked, treating English as the source language for the sake of readability. In this example, λ_c would be added to the hypothesis score for any rule used in the hypothesis whose source side spanned *the minister*, *a speech*, *yesterday*, *gave a speech yesterday*, or *the minister gave a speech yesterday.* A rule translating, say, *minister gave a* as a unit would receive no such boost.

Chiang tested the constituency feature for Chinese-English translation, and obtained no significant improvement on the test set. The idea then seems essentially to have been abandoned; it does not appear in later discussions (Chiang, 2007).

3 Soft Syntactic Constraints, Revisited

On the face of it, there are any number of possible reasons Chiang's (2005) soft constraint did not work – including, for example, practical issues like the quality of the Chinese parses.[5] However, we focus here on two conceptual issues underlying his use of source language syntactic constituents.

[4]Formally, $\phi_c(\bar{f}, \bar{e})$ is defined as a binary feature, with value 1 if $\bar{f}$ spans a source constituent and 0 otherwise. In the latter case $\lambda_c \phi_c(\bar{f}, \bar{e}) = 0$ and the score in expression (1) is unaffected.

[5]In fact, this turns out not to be the issue; see Section 4.

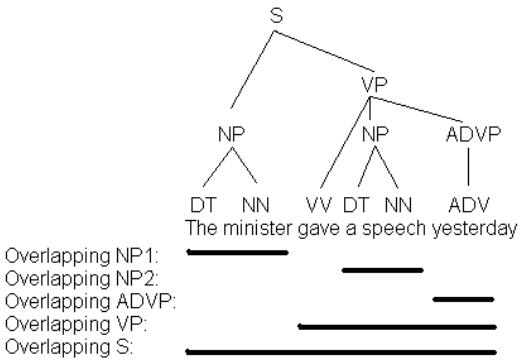

Figure 1: Illustration of Chiang's (2005) syntactic constituency feature, which does not distinguish among constituent types.

First, the constituency feature treats all syntactic constituent types equally, making no distinction among them. For any given language pair, however, there might be some source constituents that tend to map naturally to the target language as units, and others that do not (Fox, 2002; Eisner, 2003). Moreover, a parser may tend to be more accurate for some constituents than for others.

Second, the Chiang (2005) constituency feature gives a rule additional credit when the rule's source side overlaps exactly with a source-side syntactic constituent. Logically, however, it might make sense not just to give a rule $X \rightarrow \langle \bar{e}, \bar{f} \rangle$ extra credit when $\bar{f}$ matches a constituent, but to incur a *cost* when $\bar{f}$ *violates* a constituent boundary. Using the example in Figure 1, we might want to penalize hypotheses containing rules where $\bar{f}$ is *the minister gave a* (and other cases, such as *minister gave*, *minister gave a*, and so forth).[6]

These observations suggest a finer-grained approach to the constituency feature idea, retaining the idea of soft constraints, but applying them using *various* soft-constraint constituency features. Our first observation argues for distinguishing among constituent types (NP, VP, etc.). Our second observation argues for distinguishing the benefit of match-

[6]This accomplishes coverage of the logically complete set of possibilities, which include not only $\bar{f}$ matching a constituent exactly or crossing its boundaries, but also $\bar{f}$ being properly contained within the constituent span, properly containing it, or being outside it entirely. Whenever these latter possibilities occur, $\bar{f}$ will exactly match or cross the boundaries of some other constituent.

ing constituents from the cost of crossing constituent boundaries. We therefore define a space of new features as the cross product

$$\{\text{CP}, \text{IP}, \text{NP}, \text{VP}, \ldots\} \times \{=, +\}.$$

where = and + signify matching and crossing boundaries, respectively. For example, $\phi_{\text{NP=}}$ would denote a binary feature that matches whenever the span of $\bar{f}$ exactly covers an NP in the source-side parse tree, resulting in $\lambda_{\text{NP=}}$ being added to the hypothesis score (expression (1)). Similarly, $\phi_{\text{VP+}}$ would denote a binary feature that matches whenever the span of $\bar{f}$ crosses a VP boundary in the parse tree, resulting in $\lambda_{\text{VP+}}$ being *subtracted from* the hypothesis score.[7] For readability from this point forward, we will omit ϕ from the notation and refer to features such as NP= (which one could read as "NP match"), VP+ (which one could read as "VP crossing"), etc.

In addition to these individual features, we define three more variants:

- For each constituent type, e.g. NP, we define a feature NP_ that ties the weights of NP= and NP+. If NP= matches a rule, the model score is incremented by $\lambda_{NP\_}$, and if NP+ matches, the model score is decremented by the same quantity.

- For each constituent type, e.g. NP, we define a version of the model, NP2, in which NP= and NP+ are *both* included as features, with separate weights $\lambda_{NP=}$ and λ_{NP+}.

- We define a set of "standard" linguistic labels containing {CP, IP, NP, VP, PP, ADJP, ADVP, QP, LCP, DNP} and excluding other labels such as PRN (parentheses), FRAG (fragment), etc.[8] We define feature XP= as the disjunction of {CP=, IP=, . . ., DNP=}; i.e. its value equals 1 for a rule if the span of $\bar{f}$ exactly covers a constituent having any of the standard labels. The definitions of XP+, XP_, and XP2 are analogous.

- Similarly, since Chiang's original constituency feature can be viewed as a disjunctive "all-labels=" feature, we also defined "all-labels+", "all-labels2", and "all-labels_" analogously.

4 Experiments

We carried out MT experiments for translation from Chinese to English and from Arabic to English, using a descendant of Chiang's Hiero system. Language models were built using the SRI Language Modeling Toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Word-level alignments were obtained using GIZA++ (Och and Ney, 2000). The baseline model in both languages used the feature set described in Section 2; for the Chinese baseline we also included a rule-based number translation feature (Chiang, 2007).

In order to compute syntactic features, we analyzed source sentences using state of the art, tree-bank trained constituency parsers ((Huang et al., 2008) for Chinese, and the Stanford parser v.2007-08-19 for Arabic (Klein and Manning, 2003a; Klein and Manning, 2003b)). In addition to the baseline condition, and baseline plus Chiang's (2005) original constituency feature, experimental conditions augmented the baseline with additional features as described in Section 3.

All models were optimized and tested using the BLEU metric (Papineni et al., 2002) with the NIST-implemented ("shortest") effective reference length, on lowercased, tokenized outputs/references. Statistical significance of difference from the baseline BLEU score was measured by using paired bootstrap re-sampling (Koehn, 2004).[9]

4.1 Chinese-English

For the Chinese-English translation experiments, we trained the translation model on the corpora in Table 1, totalling approximately 2.1 million sentence pairs after GIZA++ filtering for length ratio. Chinese text was segmented using the Stanford segmenter (Tseng et al., 2005).

[7] Formally, $\lambda_{\text{VP+}}$ simply contributes to the sum in expression (1), as with all features in the model, but weight optimization using minimum error rate training should, and does, automatically assign this feature a negative weight.

[8] We map SBAR and S labels in Arabic parses to CP and IP, respectively, consistent with the Chinese parses. We map Chinese DP labels to NP. DNP and LCP appear only in Chinese. We ran no ADJP experiment in Chinese, because this label virtually aways spans only one token in the Chinese parses.

[9] Whenever we use the word "significant", we mean "statistically significant" (at $p < .05$ unless specified otherwise).

LDC ID	Description
LDC2002E18	Xinhua Ch/Eng Par News V1 beta
LDC2003E07	Ch/En Treebank Par Corpus
LDC2005T10	Ch/En News Mag Par Txt (Sinorama)
LDC2003E14	FBIS Multilanguage Txts
LDC2005T06	Ch News Translation Txt Pt 1
LDC2004T08	HK Par Text (only HKNews)

Table 1: Training corpora for Chinese-English translation

We trained a 5-gram language model using the English (target) side of the training set, pruning 4-gram and 5-gram singletons. For minimum error rate training and development we used the NIST MTeval MT03 set.

Table 2 presents our results. We first evaluated translation performance using the NIST MT06 (nist-text) set. Like Chiang (2005), we find that the original, undifferentiated constituency feature (Chiang-05) introduces a negligible, statistically insignificant improvement over the baseline. However, we find that several of the finer-grained constraints (IP=, VP=, VP+, QP+, and NP=) achieve statistically significant improvements over baseline (up to .74 BLEU), and the latter three also improve significantly on the undifferentiated constituency feature. By combining multiple finer-grained syntactic features, we obtain significant improvements of up to 1.65 BLEU points (NP_, VP2, IP2, all-labels_, and XP+).

We also obtained further gains using combinations of features that had performed well; e.g., condition IP2.VP2.NP_ augments the baseline features with IP2 and VP2 (i.e. IP=, IP+, VP= and VP+), and NP_ (tying weights of NP= and NP+; see Section 3). Since component features in those combinations were informed by individual-feature performance on the test set, we tested the best performing conditions from MT06 on a new test set, NIST MT08. NP= and VP+ yielded significant improvements of up to 1.53 BLEU. Combination conditions replicated the pattern of results from MT06, including the same increasing order of gains, with improvements up to 1.11 BLEU.

4.2 Arabic-English

For Arabic-English translation, we used the training corpora in Table 3, approximately 100,000 sen-

Chinese	**MT06**	**MT08**
Baseline	.2624	.2064
Chiang-05	.2634	.2065
PP=	.2607	
DNP+	.2621	
CP+	.2622	
AP+	.2633	
AP=	.2634	
DNP=	.2640	
IP+	.2643	
PP+	.2644	
LCP=	.2649	
LCP+	.2654	
CP=	.2657	
NP+	.2662	
QP=	.2674^+	.2071
IP=	**.2680*+**	.2061
VP=	.2683*	.2072
VP+	**.2693**++**	**.2109*+**
QP+	**.2694**++**	.2091
NP=	**.2698**++**	**.2217**++**
Multiple / conflated features:		
QP2	.2614	
NP2	.2621	
XP=	.2630	
XP2	.2633	
all-labels+	.2633	
VP_	.2637	
QP_	.2641	
NP.VP.IP=.QP.VP+	.2646	
IP_	.2647	
IP2+VP2	.2649	
all-labels2	.2673*-	.2070
NP_	**.2690**++**	.2101^+
IP2.VP2.NP_	**.2699**++**	**.2105*+**
VP2	**.2722**++**	**.2123**++**
all-labels_	**.2731**++**	**.2125*++**
IP2	**.2750****++**	**.2132**+**
XP+	**.2789**++**	**.2175**++**

Table 2: Chinese-English results. *: Significantly better than baseline ($p < .05$). **: Significantly better than baseline ($p < .01$). ^: Almost significantly better than baseline ($p < .075$). +: Significantly better than Chiang-05 ($p < .05$). ++: Significantly better than Chiang-05 ($p < .01$). -: Almost significantly better than Chiang-05 ($p < .075$).

LDC ID	Description
LDC2004T17	Ar News Trans Txt Pt 1
LDC2004T18	Ar/En Par News Pt 1
LDC2005E46	Ar/En Treebank En Translation
LDC2004E72	eTIRR Ar/En News Txt

Table 3: Training corpora for Arabic-English translation

tence pairs after GIZA++ length-ratio filtering. We trained a trigram language model using the English side of this training set, plus the English Gigaword v2 AFP and Gigaword v1 Xinhua corpora. Development and minimum error rate training were done using the NIST MT02 set.

Table 4 presents our results. We first tested on on the NIST MT03 and MT06 (nist-text) sets. On MT03, the original, undifferentiated constituency feature did not improve over baseline. Two individual finer-grained features (PP+ and AdvP=) yielded statistically significant gains up to .42 BLEU points, and feature combinations AP2, XP2 and all-labels2 yielded significant gains up to 1.03 BLEU points. XP2 and all-labels2 also improved significantly on the undifferentiated constituency feature, by .72 and 1.11 BLEU points, respectively.

For MT06, Chiang's original feature improved the baseline significantly — this is a new result using his feature, since he did not experiment with Arabic — as did our our IP=, PP=, and VP= conditions. Adding individual features PP+ and AdvP= yielded significant improvements up to 1.4 BLEU points over baseline, and in fact the improvement for individual feature AdvP= over Chiang's undifferentiated constituency feature approaches significance ($p < .075$).

More important, several conditions combining features achieved statistically significant improvements over baseline of up 1.94 BLEU points: XP2, IP2, IP, VP=.PP+.AdvP=, AP2, PP+.AdvP=, and AdvP2. Of these, AdvP2 is also a significant improvement over the undifferentiated constituency feature (Chiang-05), with $p < .01$. As we did for Chinese, we tested the best-performing models on a new test set, NIST MT08. Consistent patterns reappeared: improvements over the baseline up to 1.69 BLEU ($p < .01$), with AdvP2 again in the lead (also outperforming the undifferentiated constituency feature, $p < .05$).

Arabic	**MT03**	**MT06**	**MT08**
Baseline	.4795	.3571	.3571
Chiang-05	.4787	**.3679****	**.3678****
VP+	.4802	.3481	
AP+	.4856	.3495	
IP+	.4818	.3516	
CP=	.4815	.3523	
NP=	.4847	.3537	
NP+	.4800	.3548	
AP=	.4797	.3569	
AdvP+	.4852	.3572	
CP+	.4758	.3578	
IP=	.4811	**.3636****	**.3647****
PP=	.4801	**.3651****	**.3662****
VP=	.4803	**.3655****	**.3694****
PP+	**.4837****	**.3707****	**.3700****
AdvP=	**.4823****	**.3711**-**	**.3717****
Multiple / conflated features:			
XP+	.4771	.3522	
all-labels2	**.4898**+**	.3536	.3572
all-labels_	.4828	.3548	
VP2	.4826	.3552	
NP2	.4832	.3561	
AdvP.VP.PP.IP=	.4826	.3571	
VP_	.4825	.3604	
all-labels+	.4825	.3600	
XP2	**.4859****+	.3605^	**.3613****
IP2	.4793	**.3611***	.3593
IP_	.4791	**.3635***	**.3648****
XP=	.4808	**.3659****	**.3704**+**
VP=.PP+.AdvP=	**.4833****	**.3677****	**.3718****
AP2	**.4840****	**.3692****	**.3719****
PP+.AdvP=	.4777	**.3708****	**.3680****
AdvP2	.4803	**.3765**++**	**.3740**+**

Table 4: Arabic-English Experiments. Results are sorted by MT06 BLEU score. *: Better than baseline ($p < .05$). **: Better than baseline ($p < .01$). +: Better than Chiang-05 ($p < .05$). ++: Better than Chiang-05 ($p < .01$). -: Almost significantly better than Chiang-05 ($p < .075$)

5 Discussion

The results in Section 4 demonstrate, to our knowledge for the first time, that significant and sometimes substantial gains over baseline can be obtained by incorporating soft syntactic constraints into Hiero's translation model. Within language, we also see considerable consistency across multiple test sets, in terms of which constraints tend to help most.

Furthermore, our results provide some insight into why the original approach may have failed to yield a positive outcome. For Chinese, we found that when we defined finer-grained versions of the exact-match features, there was value for some constituency types in biasing the model to favor matching the source language parse. Moreover, we found that there was significant value in allowing the model to be sensitive to violations (crossing boundaries) of source parses. These results confirm that parser quality was not the limitation in the original work (or at least not the only limitation), since in our experiments the parser was held constant.

Looking at combinations of new features, some "double-feature" combinations (VP2, IP2) achieved large gains, although note that more is not necessarily better: combinations of more features did not yield better scores, and some did not yield any gain at all. No conflated feature reached significance, but it is not the case that all conflated features are worse than their same-constituent "double-feature" counterparts. We found no simple correlation between finer-grained feature scores (and/or boundary condition type) and combination or conflation scores. Since some combinations seem to cancel individual contributions, we can conclude that the higher the number of participant features (of the kinds described here), the more likely a cancellation effect is; therefore, a "double-feature" combination is more likely to yield higher gains than a combination containing more features.

We also investigated whether non-canonical linguistic constituency labels such as PRN, FRAG, UCP and VSB introduce "noise", by means of the XP features — the XP= feature is, in fact, simply the undifferentiated constituency feature, but sensitive only to "standard" XPs. Although performance of XP=, XP2 and all-labels+ were similar to that of the undifferentiated constituency feature, XP+ achieved the highest gain. Intuitively, this seems plausible: the feature says, at least for Chinese, that a translation hypothesis should incur a penalty if it is translating a substring as a unit when that substring is not a canonical source constituent.

Having obtained positive results with Chinese, we explored the extent to which the approach might improve translation using a very different source language. The approach on Arabic-English translation yielded large BLEU gains over baseline, as well as significant improvements over the undifferentiated constituency feature. Comparing the two sets of experiments, we see that there are definitely language-specific variations in the value of syntactic constraints; for example, AdvP, the top performer in Arabic, cannot possibly perform well for Chinese, since in our parses the AdvP constituents rarely include more than a single word. At the same time, some IP and VP variants seem to do generally well in both languages. This makes sense, since — at least for these language pairs and perhaps more generally — clauses and verb phrases seem to correspond often on the source and target side. We found it more surprising that no NP variant yielded much gain in Arabic; this question will be taken up in future work.

6 Related Work

Space limitations preclude a thorough review of work attempting to navigate the tradeoff between using language analyzers and exploiting unconstrained data-driven modeling, although the recent literature is full of variety and promising approaches. We limit ourselves here to several approaches that seem most closely related.

Among approaches using parser-based syntactic models, several researchers have attempted to reduce the strictness of syntactic constraints in order to better exploit shallow correspondences in parallel training data. Our introduction has already briefly noted Cowan et al. (2006), who relax parse-tree-based alignment to permit alignment of non-constituent subphrases on the source side, and translate modifiers using a separate phrase-based model, and DeNeefe et al. (2007), who modify syntax-based extraction and binarize trees (following (Wang et al., 2007b)) to improve phrasal cov-

erage. Similarly, Marcu et al. (2006) relax their syntax-based system by rewriting target-side parse trees on the fly in order to avoid the loss of "non-syntactifiable" phrase pairs. Setiawan *et al.* (2007) employ a "function-word centered syntax-based approach", with synchronous CFG and extended ITG models for reordering phrases, and relax syntactic constraints by only using a small number function words (approximated by high-frequency words) to guide the phrase-order inversion. Zollman and Venugopal (2006) start with a target language parser and use it to provide constraints on the extraction of hierarchical phrase pairs. Unlike Hiero, their translation model uses a full range of named nonterminal symbols in the synchronous grammar. As an alternative way to relax strict parser-based constituency requirements, they explore the use of phrases spanning generalized, categorial-style constituents in the parse tree, e.g. type NP/NN denotes a phrase like *the great* that lacks only a head noun (say, *wall*) in order to comprise an NP.

In addition, various researchers have explored the use of hard linguistic constraints on the source side, e.g. via "chunking" noun phrases and translating them separately (Owczarzak et al., 2006), or by performing hard reorderings of source parse trees in order to more closely approximate target-language word order (Wang et al., 2007a; Collins et al., 2005).

Finally, another soft-constraint approach that can also be viewed as coming from the data-driven side, adding syntax, is taken by Riezler and Maxwell (2006). They use LFG dependency trees on both source and target sides, and relax syntactic constraints by adding a "fragment grammar" for unparsable chunks. They decode using Pharaoh, augmented with their own log-linear features (such as $p(e_{snippet}|f_{snippet})$ and its converse), side by side to "traditional" lexical weights. Riezler and Maxwell (2006) do not achieve higher BLEU scores, but do score better according to human grammaticality judgments for in-coverage cases.

7 Conclusion

When hierarchical phrase-based translation was introduced by Chiang (2005), it represented a new and successful way to incorporate syntax into statistical MT, allowing the model to exploit non-local dependencies and lexically sensitive reordering without requiring linguistically motivated parsing of either the source or target language. An approach to incorporating parser-based constituents in the model was explored briefly, treating syntactic constituency as a soft constraint, with negative results.

In this paper, we returned to the idea of linguistically motivated soft constraints, and we demonstrated that they can, in fact, lead to substantial improvements in translation performance when integrated into the Hiero framework. We accomplished this using constraints that not only distinguish among constituent types, but which also distinguish between the benefit of matching the source parse bracketing, versus the cost of using phrases that cross relevant bracketing boundaries. We demonstrated improvements for Chinese-English translation, and succeed in obtaining substantial gains for Arabic-English translation, as well.

Our results contribute to a growing body of work on combining monolingually based, linguistically motivated syntactic analysis with translation models that are closely tied to observable parallel training data. Consistent with other researchers, we find that "syntactic constituency" may be too coarse a notion by itself; rather, there is value in taking a finer-grained approach, and in allowing the model to decide how far to trust each element of the syntactic analysis as part of the system's optimization process.

Acknowledgments

This work was supported in part by DARPA prime agreement HR0011-06-2-0001. The authors would like to thank David Chiang and Adam Lopez for making their source code available; the Stanford Parser team and Mary Harper for making their parsers available; David Chiang, Amy Weinberg, and CLIP Laboratory colleagues, particularly Chris Dyer, Adam Lopez, and Smaranda Muresan, for discussion and invaluable assistance.

References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2007. CCG supertags in factored statistical machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation 2007*.

P.F. Brown, S.A.D. Pietra, V.J.D. Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–313.

S. F. Chen and J. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Tech. Report TR-10-98, Comp. Sci. Group, Harvard U.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL-05*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL-05*.

Brooke Cowan, Ivona Kucerova, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proc. EMNLP*.

S DeNeefe, K. Knight, W. Wang, and D. Marcu. 2007. What can syntax-based MT learn from phrase-based MT? In *Proceedings of EMNLP-CoNLL*.

J. Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *ACL Companion Vol.*

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proc. EMNLP 2002*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING/ACL-06*.

H. Hassan, K. Sima'an, and A. Way. 2007. Integrating supertags into phrase-based statistical machine translation. In *Proc. ACL-07*, pages 288–295.

Zhongqiang Huang, Denis Filimonov, and Mary Harper. 2008. Accuracy enhancements for mandarin parsing. Tech. report, University of Maryland.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proceedings of ACL-03*, pages 423–430.

Dan Klein and Christopher D. Manning. 2003b. Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*, 15(NIPS 2002):3–10.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proc. EMNLP+CoNLL*, pages 868–876, Prague.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 127–133.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*.

Adam Lopez. (to appear). Statistical machine translation. *ACM Computing Surveys*. Earlier version: A Survey of Statistical Machine Translation. U. of Maryland, UMIACS tech. report 2006-47. Apr 2007.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proc. EMNLP*, pages 44–52.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447. GIZA++.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 160–167.

K. Owczarzak, B. Mellebeek, D. Groves, J. Van Genabith, and A. Way. 2006. Wrapper syntax for example-based machine translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 148–155.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpusbased comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of the Human Language Technology Conference (ACL'2002)*, pages 124–127, San Diego, CA.

Stefan Riezler and John Maxwell. 2006. Grammatical machine translation. In *Proc. HLT-NAACL*, New York, NY.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. Ordering phrases with function words. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 712–719.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

Chao Wang, Michael Collins, and Phillip Koehn. 2007a. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP*.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007b. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proc. EMNLP+CoNLL 2007*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the SMT Workshop, HLT-NAACL*.