SPMT: Statistical Machine Translation with Syntactified Target Language Phrases

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight
Language Weaver Inc.
4640 Admiralty Way, Suite 1210
Marina del Rey, CA 90292
{dmarcu,wwang,aechihabi,kknight}@languageweaver.com

Abstract

We introduce SPMT, a new class of statistical Translation Models that use Syntactified target language Phrases. The SPMT models outperform a state of the art phrase-based baseline model by 2.64 Bleu points on the NIST 2003 Chinese-English test corpus and 0.28 points on a human-based quality metric that ranks translations on a scale from 1 to 5.

1 Introduction

During the last four years, various implementations and extentions to phrase-based statistical models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004) have led to significant increases in machine translation accuracy. Although phrase-based models yield high-quality translations for language pairs that exhibit similar word order, they fail to produce grammatical outputs for language pairs that are syntactically divergent. Recent models that exploit syntactic information of the source language (Quirk et al., 2005) have been shown to produce better outputs than phrase-based systems when evaluated on relatively small scale, domain specific corpora. And syntax-inspired formal models (Chiang, 2005), in spite of being trained on significantly less data, have shown promising results when compared on the same test sets with mature phrase-based systems. To our knowledge though, no previous research has demonstrated that a syntax-based statistical translation system could produce better results than a phrase-based system on a large-scale, well-established, open domain translation task. In this paper we present such a system.

Our translation models rely upon and naturally exploit submodels (feature functions) that have been initially developed in phrase-based systems for choosing target translations of source language phrases, and use new, syntax-based translation and target language submodels for assembling target phrases into well-formed, grammatical outputs.

After we introduce our models intuitively, we discuss their formal underpinning and parameter training in Section 2. In Section 3, we present our decoder and, in Section 4, we evaluate our models empirically. In Section 5, we conclude with a brief discussion.

2 SPMT: statistical Machine Translation with Syntactified Phrases

2.1 An intuitive introduction to SPMT

After being exposed to 100M+ words of parallel Chinese-English texts, current phrase-based statistical machine translation learners induce reasonably reliable phrase-based probabilistic dictionaries. For example, our baseline statistical phrase-based system learns that, with high probabilities, the Chinese phrases "ASTRO- -NAUTS", "FRANCE AND RUSSIA" and "COMINGFROM" can be translated into English as "astronauts"/"cosmonauts", "france and russia"/"france and russian" and "coming from"/"from", respectively. [1] Unfortunately, when given as input Chinese sentence 1, our phrase-based system produces the output shown in 2 and not the translation in 3, which correctly orders the phrasal translations into a grammatical sequence. We believe this happens because the distortion/reordering models that are used by state-of-the-art phrase-based systems, which exploit phrase movement and ngram target

[1] To increase readability, in this paper, we represent Chinese words using fully capitalized English glosses and English words using lowercased letters.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 44–52,
Sydney, July 2006.

language models (Och and Ney, 2004; Tillman, 2004), are too weak to help a phrase-based decoder reorder the target phrases into grammatical outputs.

THESE 7PEOPLE INCLUDE COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS . (1)

the 7 people including those from france and the russian cosmonauts . (2)

these 7 people include astronauts coming from france and russia . (3)

One method for increasing the ability of a decoder to reorder target language phrases is that of decorating them with syntactic constituent information. For example, we may make explicit that the Chinese phrase "ASTRO- -NAUTS" may be translated into English as a noun phrase, NP(NNS(astronauts)); that the phrase FRANCE AND RUSSIA may be translated into a complex noun-phrase, NP(NP(NNP(france)) CC(and) NP(NNP(russia))); that the phrase COMINGFROM may be translated into a partially realized verb phrase that is looking for a noun phrase to its right in order to be fully realized, VP(VBG(coming) PP(IN(from) NP:x0)); and that the Chinese particle p-DE, when occurring between a Chinese string that was translated into a verb phrase to its left and another Chinese string that was translated into a noun phrase to its right, VP:x1 p-DE NP:x0, should be translated to nothing, while forcing the reordering of the two constituents, NP(NP:x0, VP:x1). If all these translation rules (labeled r_1 to r_4 in Figure 1) were available to a decoder that derives English parse trees starting from Chinese input strings, this decoder could produce derivations such as that shown in Figure 2. Because our approach uses translation rules with Syntactified target language Phrases (see Figure 1), we call it SPMT.

r_1 :NP(NNS(astronauts)) → ASTRO- -NAUTS
r_2 :NP(NP(NNP(france)) CC(and) NP(NNP(russia))) → FRANCE AND RUSSIA
r_3 :VP(VBG(coming) PP(IN(from) NP:x0)) → COMINGFROM x0
r_4 :NP(NP:x0, VP:x1) → x1 p-DE x0
r_5 :NNP(france) → FRANCE
r_6 :NP(NP(NNP(france)) CC(and) NP:x0) → FRANCE AND x0
r_7 :NNS(astronauts) → ASTRO- -NAUTS
r_8 :NNP(russia) → RUSSIA
r_9 :NP(NNS:x0) → x0
r_{10} :PP(IN:x0 NP:x1) → x0 x1
r_{11} :NP(NP:x0 CC:x1 NP:x2) → x0 x1 x2
r_{12} :NP(NNP:x0) → x0
r_{13} :CC(and) → AND
r_{14} :NP(NP:x0 CC(and) NP:x1) → x0 AND x1
r_{15} :NP(NP:x0 VP(VBG(coming) PP(IN(from) NP:x1))) → x1 COMINGFROM x0

Figure 1: Examples of xRS rules.

2.2 A formal introduction to SPMT

2.2.1 Theoretical foundations

We are interested to model a generative process that explains how English parse trees π and their associated English string yields E, foreign sentences, F, and word-level alignments, A, are produced. We assume that observed (π, F, A) triplets are generated by a stochastic process similar to that used in Data Oriented Parsing models (Bonnema, 2002). For example, if we assume that the generative process has already produced the top NP node in Figure 2, then the corresponding partial English parse tree, foreign/source string, and word-level alignment could be generated by the rule derivation $r_4(r_1, r_3(r_2))$, where each rule is assumed to have some probability.

The extended tree to string transducers introduced by Knight and Graehl (2005) provide a natural framework for expressing the tree to string transformations specific to our SPMT models. The transformation rules we plan to exploit are equivalent to one-state xRS top-down transducers with look ahead, which map subtree patterns to strings. For example, rule r_3 in Figure 1 can be applied only when one is in a state that has a VP as its syntactic constituent and the tree pattern VP(VBG(coming) PP(IN(from) NP)) immediately underneath. The rule application outputs the string "COMINGFROM" as the transducer moves to the state co-indexed by x0; the outputs produced from the new state will be concatenated to the right of the string "COMINGFROM".

Since there are multiple derivations that could lead to the same outcome, the probability of a tuple (π, F, A) is obtained by summing over all derivations $\theta_i \in \Theta$ that are consistent with the tu-

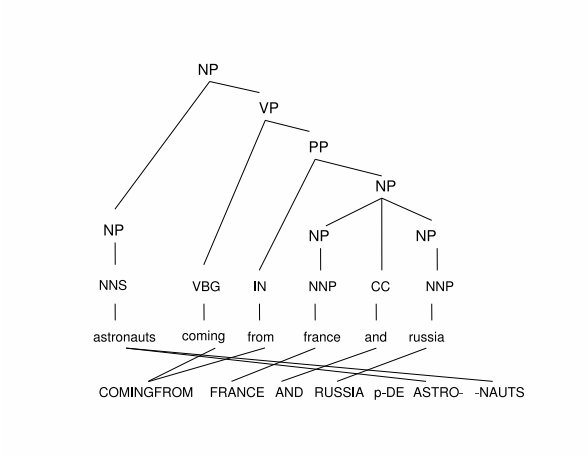

Figure 2: English parse tree derivation of the Chinese string COMINGFROM FRANCE AND RUSSIA p-DE ASTRO- -NAUTS.

ple, $c(\Theta) = (\pi, F, A)$. The probability of each derivation θ_i is given by the product of the probabilities of all the rules $p(r_j)$ in the derivation (see equation 4).

$$Pr(\pi, F, A) = \sum_{\theta_i \in \Theta, c(\Theta)=(\pi,F,A)} \prod_{r_j \in \theta_i} p(r_j) \quad (4)$$

In order to acquire the rules specific to our model and to induce their probabilities, we parse the English side of our corpus with an in-house implementation (Soricut, 2005) of Collins parsing models (Collins, 2003) and we word-align the parallel corpus with the Giza++[2] implementation of the IBM models (Brown et al., 1993). We use the automatically derived ⟨English-parse-tree, English-sentence, Foreign-sentence, Word-level-alignment⟩ tuples in order to induce xRS rules for several models.

2.2.2 SPMT Model 1

In our simplest model, we assume that each tuple (π, F, A) in our automatically annotated corpus could be produced by applying a combination of minimally syntactified, lexicalized, phrase-based compatible xRS rules, and minimal/necessary, non-lexicalized xRS rules. We call a rule non-lexicalized whenever it does not have any directly aligned source-to-target words. Rules r_9–r_{12} in Figure 1 are examples of non-lexicalized rules.

Minimally syntactified, lexicalized, phrase-based-compatible xRS rules are extracted via a simple algorithm that finds for each foreign phrase F_i^j, the smallest xRS rule that is consistent with the foreign phrase F_i^j, the English syntactic tree π, and the alignment A. The algorithm finds for each foreign/source phrase span its projected span on the English side and then traverses the English parse tree bottom up until it finds a node that subsumes the projected span. If this node has children that fall outside the projected span, then those children give rise to rules that have variables. For example, if the tuple shown in Figure 2 is in our training corpus, for the foreign/source phrases FRANCE, FRANCE AND, FRANCE AND RUSSIA, and ASTRO- -NAUTS, we extract the minimally syntactified, lexicalized phrase-based-compatible xRS rules r_5, r_6, r_2, and r_7 in Figure 1, respectively. Because, as in phrase-based MT, all our rules have continuous phrases on both the source and target language sides, we call these phrase-based compatible xRS rules.

Since these lexicalized rules are not sufficient to explain an entire (π, F, A) tuple, we also extract the required minimal/necessary, non-lexicalized xRS rules. The minimal non-lexicalized rules that are licensed by the tuple in Figure 2 are labeled r_4, r_9, r_{10}, r_{11} and r_{12} in Figure 1. To obtain the non-lexicalized xRS rules, we compute the set of all minimal rules (lexicalized and non-lexicalized) by applying the algorithm proposed by Galley et al. (2006) and then remove the lexicalized rules. We remove the Galley et al.'s lexicalized rules because they are either already accounted for by the minimally syntactified, lexicalized, phrase-based-compatible xRS rules or they subsume non-continuous source-target phrase pairs.

It is worth mentioning that, in our framework, a rule is defined to be "minimal" with respect to a foreign/source language phrase, i.e., it is the minimal xRS rule that yields that source phrase. In contrast, in the work of Galley et al. (2004; 2006), a rule is defined to be minimal when it is necessary in order to explain a (π, F, A) tuple.

Under SPMT model 1, the tree in Figure 2 can be produced, for example, by the following derivation: $r_4(r_9(r_7), r_3(r_6(r_{12}(r_8))))$.

2.2.3 SPMT Model 1 Composed

We hypothesize that composed rules, i.e., rules that can be decomposed via the application of a sequence of Model 1 rules may improve the performance of an SPMT system. For example, although the minimal Model 1 rules r_{11} and r_{13} are

[2] http://www.fjoch.com/GIZA++.html

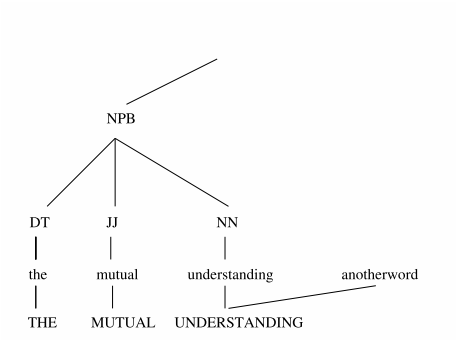

Figure 3: Problematic syntactifications of phrasal translations.

sufficient for building an English NP on top of two NPs separated by the Chinese conjunction AND, the composed rule r_{14} in Figure 1 accomplishes the same result in only one step. We hope that the composed rules could play in SPMT the same role that phrases play in string-based translation models.

To test our hypothesis, we modify our rule extraction algorithm so that for every foreign phrase F_i^j, we extract not only a minimally syntactified, lexicalized xRS rule, but also *one* composed rule. The composed rule is obtained by extracting the rule licensed by the foreign/source phrase, alignment, English parse tree, and the first multi-child ancestor node of the root of the minimal rule. Our intuition is that composed rules that involve the application of more than two minimal rules are not reliable. For example, for the tuple in Figure 2, the composed rule that we extract given the foreign phrases AND and COMINGFROM are respectively labeled as rules r_{14} and r_{15} in Figure 1.

Under the SPMT composed model 1, the tree in Figure 2 can be produced, for example, by the following derivation: $r_{15}(r_9(r_7), r_{14}(r_{12}(r_5), r_{12}(r_8)))$.

2.2.4 SPMT Model 2

In many instances, the tuples (π, F, A) in our training corpus exhibit alignment patterns that can be easily handled within a phrase-based SMT framework, but that become problematic in the SPMT models discussed until now.

Consider, for example, the (π, F, A) tuple fragment in Figure 3. When using a phrase-based translation model, one can easily extract the phrase pair (THE MUTUAL; the mutual) and use it during the phrase-based model estimation phrase and in decoding. However, within the xRS transducer framework that we use, it is impossible to extract an equivalent syntactified phrase translation rule that subsumes the same phrase pair because valid xRS translation rules cannot be multi-headed. When faced with this constraint, one has several options:

- One can label such phrase pairs as non-syntactifiable and ignore them. Unfortunately, this is a lossy choice. On our parallel English-Chinese corpus, we have found that approximately 28% of the foreign/source phrases are non-syntactifiable by this definition.
- One can also traverse the parse tree upwards until one reaches a node that is xRS valid, i.e., a node that subsumes the entire English span induced by a foreign/source phrase and the corresponding word-level alignment. This choice is also inappropriate because phrase pairs that are usually available to phrase-based translation systems are then expanded and made available in the SPTM models only in larger applicability contexts.
- A third option is to create xRS compatible translation rules that overcome this constraint.

Our SPMT Model 2 adopts the third option by rewriting on the fly the English parse tree for each foreign/source phrase and alignment that lead to non-syntactifiable phrase pairs. The rewriting process adds new rules to those that can be created under the SPMT model 1 constraints. The process creates one xRS rule that is headed by a pseudo, non-syntactic nonterminal symbol that subsumes the target phrase and corresponding multi-headed syntactic structure; and one sibling xRS rule that explains how the non-syntactic nonterminal symbol can be combined with other genuine nonterminals in order to obtain genuine parse trees. In this view, the foreign/source phrase THE MUTUAL and corresponding alignment in Figure 3 licenses the rules ⋆NPB⋆_NN(DT(the) JJ(mutual)) → THE MUTUAL and NPB(⋆NPB⋆_NN:x0 NN:x1) → x0 x1 even though the foreign word UNDERSTANDING is aligned to an English word outside the NPB consituent. The name of the non-syntactic nonterminal reflects the intuition that the English phrase "the mutual" corresponds to a partially realized NPB that needs an NN to its right in order to be fully realized.

Our hope is that the rules headed by pseudo nonterminals could make available to an SPMT system all the rules that are typically available to a phrase-based system; and that the sibling rules could provide a sufficiently robust generalization layer for integrating pseudo, partially realized constituents into the overall decoding process.

2.2.5 SPMT Model 2 Composed

The SPMT composed model 2 uses all rule types described in the previous models.

2.3 Estimating rule probabilities

For each model, we extract all rule instances that are licensed by a symmetrized Giza-aligned parallel corpus and the constraints we put on the model. We condition on the root node of each rule and use the rule counts $f(r)$ and a basic maximum likelihood estimator to assign to each rule type a conditional probability (see equation 5).

$$p(r|root(r)) = \frac{f(r)}{\sum_{r':root(r')=root(r)} f(r')} \quad (5)$$

It is unlikely that this joint probability model can be discriminative enough to distinguish between good and bad translations. We are not too concerned though because, in practice, we decode using a larger set of submodels (feature functions).

Given the way all our lexicalized xRS rules have been created, one can safely strip out the syntactic information and end up with phrase-to-phrase translation rules. For example, in string-to-string world, rule r_5 in Figure 1 can be rewritten as "france → FRANCE"; and rule r_6 can be rewritten as "france and → FRANCE AND". When one analyzes the lexicalized xRS rules in this manner, it is easy to associate with them any of the submodel probability distributions that have been proven useful in statistical phrase-based MT. The non-lexicalized rules are assigned probability distributions under these submodels as well by simply assuming a NULL phrase for any missing lexicalized source or target phrase.

In the experiments described in this paper, we use the following submodels (feature functions):

Syntax-based-like submodels:

- $p_{root}(r_i)$ is the root normalized conditional probability of all the rules in a model.
- $p_{cfg}(r_i)$ is the CFG-like probability of the non-lexicalized rules in the model. The lexicalized rules have by definition $p_{cfg} = 1$.
- *is_lexicalized*(r_i) is an indicator feature function that has value 1 for lexicalized rules, and value 0 otherwise.
- *is_composed*(r_i) is an indicator feature function that has value 1 for composed rules.
- *is_lowcount*(r_i) is an indicator feature function that has value 1 for the rules that occur less than 3 times in the training corpus.

Phrase-based-like submodels:

- *lex_pef*(r_i) is the direct phrase-based conditional probability computed over the foreign/source and target phrases subsumed by a rule.
- *lex_pfe*(r_i) is the inverse phrase-based conditional probability computed over the source and target phrases subsumed by a rule.
- *m1*(r_i) is the IBM model 1 probability computed over the bags of words that occur on the source and target sides of a rule.
- *m1inv*(r_i) is the IBM model 1 inverse probability computed over the bags of words that occur on the source and target sides of a rule.
- *lm*(e) is the language model probability of the target translation under an ngram language model.
- *wp*(e) is a word penalty model designed to favor longer translations.

All these models are combined log-linearly during decoding. The weights of the models are computed automatically using a variant of the Maximum Bleu training procedure proposed by Och (2003).

The phrase-based-like submodels have been proved useful in phrase-based approaches to SMT (Och and Ney, 2004). The first two syntax-based submodels implement a "fused" translation and lexical grounded distortion model (p_{root}) and a syntax-based distortion model (p_{cfg}). The indicator submodels are used to determine the extent to which our system prefers lexicalized vs. non-lexicalized rules; simple vs. composed rules; and high vs. low count rules.

3 Decoding

3.1 Decoding with one SPMT model

We decode with each of our SPMT models using a straightforward, bottom-up, CKY-style decoder that builds English syntactic constituents on the top of Chinese sentences. The decoder uses a binarized representation of the rules, which is obtained via a syncronous binarization procedure (Zhang et al., 2006). The CKY-style decoder computes the probability of English syntactic constituents in a bottom up fashion, by log-linearly interpolating all the submodel scores described in Section 2.3.

The decoder is capable of producing nbest derivations and nbest lists (Knight and Graehl, 2005), which are used for Maximum Bleu training (Och, 2003). When decoding the test corpus, the decoder returns the translation that has the most probable derivation; in other words, the sum operator in equation 4 is replaced with an argmax.

3.2 Decoding with multiple SPMT models

Combining multiple MT outputs to increase performance is, in general, a difficult task (Matusov et al., 2006) when significantly different engines compete for producing the best outputs. In our case, combining multiple MT outputs is much simpler because the submodel probabilities across the four models described here are mostly identifical, with the exception of the root normalized and CFG-like submodels which are scaled differently – since Model 2 composed has, for example, more rules than Model 1, the root normalized and CFG-like submodels have smaller probabilities for identical rules in Model 2 composed than in Model 1. We compare these two probabilities across the submodels and we scale all model probabilities to be compatible with those of Model 2 composed.

With this scaling procedure into place, we produce 6,000 non-unique nbest lists for all sentences in our development corpus, using all SPMT submodels. We concatenate the lists and we learn a new combination of weights that maximizes the Bleu score of the combined nbest list using the same development corpus we used for tuning the individual systems (Och, 2003). We use the new weights in order to rerank the nbest outputs on the test corpus.

4 Experiments

4.1 Automatic evaluation of the models

We evaluate our models on a Chinese to English machine translation task. We use the same training corpus, 138.7M words of parallel Chinese-English data released by LDC, in order to train several statistical-based MT systems:

- PBMT, a strong state of the art phrase-based system that implements the alignment template model (Och and Ney, 2004); this is the system ISI has used in the 2004 and 2005 NIST evaluations.
- four SPMT systems (M1, M1C, M2, M2C) that implement each of the models discussed in this paper;
- a SPMT system, *Comb*, that combines the outputs of all SPMT models using the procedure described in Section 3.2.

In all systems, we use a rule extraction algorithm that limits the size of the foreign/source phrases to four words. For all systems, we use a Kneser-Ney (1995) smoothed trigram language model trained on 2.3 billion words of English. As development data for the SPMT systems, we used the sentences in the 2002 NIST development corpus that are shorter than 20 words; we made this choice in order to finish all experiments in time for this submission. The PBMT system used all sentences in the 2002 NIST corpus for development. As test data, we used the 2003 NIST test set.

Table 1 shows the number of string-to-string or tree-to-string rules extracted by each system and the performance on both the subset of sentences in the test corpus that were shorter than 20 words and the entire test corpus. The performance is measured using the Bleu metric (Papineni et al., 2002) on lowercased, tokenized outputs/references.

The results show that the SPMT models clearly outperform the phrase-based systems – the 95% confidence intervals computed via bootstrap resampling in all cases are around 1 Bleu point. The results also show that the simple system combination procedure that we have employed is effective in our setting. The improvement on the development corpus transfers to the test setting as well.

A visual inspection of the outputs shows significant differences between the outputs of the four models. The models that use composed rules prefer to produce outputs by using mostly lexicalized

System	# of rules (in millions)	Bleu score on Dev (4 refs) < 20 words	Bleu score on Test (4 refs) < 20 words	Bleu score on Test (4 refs)
PBMT	125.8	34.56	34.83	31.46
SPMT-M1	34.2	37.60	38.18	33.15
SPMT-M1C	75.7	37.30	38.10	32.39
SPMT-M2	70.4	37.77	38.74	33.39
SPMT-M2C	111.1	37.48	38.59	33.16
SPMT-Comb	111.1	39.44	39.56	34.10

Table 1: Automatic evaluation results.

rules; in contrast, the simple M1 and M2 models produce outputs in which content is translated primarily using lexicalized rules and reorderings and word insertions are explained primarily by the non-lexical rules. It appears that the two strategies are complementary, succeeding and failing in different instances. We believe that this complementarity and the overcoming of some of the search errors in our decoder during the model rescoring phase explain the success of the system combination experiments.

We suspect that our decoder still makes many search errors. In spite of this, the SPTM outputs are still significantly better than the PBMT outputs.

4.2 Human-based evaluation of the models

We also tested whether the Bleu score improvements translate into improvements that can be perceived by humans. To this end, we randomly selected 138 sentences of less than 20 words from our development corpus; we expected the translation quality of sentences of this size to be easier to assess than that of sentences that are very long.

We prepared a web-based evaluation interface that showed for each input sentence:

- the Chinese input;
- three English reference translations;
- the output of seven "MT systems".

The evaluated "MT systems" were the six systems shown in Table 1 and one of the reference translations. The reference translation presented as automatically produced output was selected from the set of four reference translations provided by NIST so as to be representative of human translation quality. More precisely, we chose the second best reference translation in the NIST corpus according to its Bleu score against the other three reference translations. The seven outputs were randomly shuffled and presented to three English speakers for assessment.

The judges who participated in our experiment were instructed to carefully read the three reference translations and seven machine translation outputs, and assign a score between 1 and 5 to each translation output on the basis of its quality. Human judges were told that the translation quality assessment should take into consideration both the grammatical fluency of the outputs and their translation adequacy. Table 2 shows the average scores obtained by each system according to each judge. For convenience, the table also shows the Bleu scores of all systems (including the human translations) on three reference translations.

The results in Table 2 show that the human judges are remarkably consistent in preferring the syntax-based outputs over the phrase-based outputs. On a 1 to 5 quality scale, the difference between the phrase-based and syntax-based systems was, on average, between 0.2 and 0.3 points. All differences between the phrase-based baseline and the syntax-based outputs were statistically significant. For example, when comparing the phrase-based baseline against the combined system, the improvement in human scores was significant at $P = 4.04e^{-6}(t = 4.67, df = 413)$.

The results also show that the LDC reference translations are far from being perfect. Although we selected from the four references the second best according to the Bleu metric, this human reference was judged to be at a quality level of only 4.67 on a scale from 1 to 5. Most of the translation errors were fluency errors. Although the human outputs had most of the time the right meaning, the syntax was sometimes incorrect.

In order to give readers a flavor of the types of re-orderings enabled by the SPMT models, we present in Table 3, several translation outputs produced by the phrase-based baseline and the com-

System	Bleu score on Dev (3 refs) < 20 words	Judge 1	Judge 2	Judge 3	Judge avg
PBMT	31.00	3.00	3.34	2.95	3.10
SPMT-M1	33.79	3.28	3.49	3.04	3.27
SPMT-M1C	33.66	3.23	3.43	3.26	3.31
SPMT-M2	34.05	3.24	3.45	3.10	3.26
SPMT-M2C	33.42	3.24	3.48	3.13	3.28
SPMT-Combined	35.33	3.31	3.59	3.25	3.38
Human Ref	40.84	4.64	4.62	4.75	4.67

Table 2: Human-based evaluation results.

bined SPMT system. The outputs were selected to reflect both positive and negative effects of large-scale re-orderings.

5 Discussion

The SPMT models are similar to the models proposed by Chiang (2005) and Galley et al. (2006). If we analyze these three models in terms of expressive power, the Galley et al. (2006) model is more expressive than the SPMT models, which in turn, are more expressive than Chiang's model. The xRS formalism utilized by Galley et al. (2006) allows for the use of translation rules that have multi-level target tree annotations and discontinuous source language phrases. The SPMT models are less general: they use translation rules that have multi-level target tree annotations but require that the source language phrases are continuous. The Syncronous Grammar formalism utilized by Chiang is stricter than SPMT since it allows only for single-level target tree annotations.

The parameters of the SPMT models presented in this paper are easier to estimate than those of Galley et al's (2006) and can easily exploit and expand on previous research in phrase-based machine translation. Also, the SPMT models yield significantly fewer rules that the model of Galley et al. In contrast with the model proposed by Chiang, the SPMT models introduced in this paper are fully grounded in syntax; this makes them good candidates for exploring the impact that syntax-based language models could have on translation performance.

From a machine translation perspective, the SPMT translation model family we have proposed in this paper is promising. To our knowledge, we are the first to report results that show that a syntax-based system can produce results that are better than those produced by a strong phrase-based system in experimental conditions similar to those used in large-scale, well-established independent evaluations, such as those carried out annually by NIST.

Although the number of syntax-based rules used by our models is smaller than the number of phrase-based rules used in our state-of-the-art baseline system, the SPMT models produce outputs of higher quality. This feature is encouraging because it shows that the syntactified translation rules learned in the SPMT models can generalize better than the phrase-based rules.

We were also pleased to see that the Bleu score improvements going from the phrase- to the syntax-based models, as well as the Bleu improvements going from the simple syntax-based models to the combined models system are fully consistent with the human qualitative judgments in our subjective evaluations. This correlation suggests that we can continue to use the Bleu metric to further improve our models and systems.

Acknowledgements. This research was partially supported by the National Institute of Standards and Technology's Advanced Technology Program Award 70NANB4H3050 to Language Weaver Inc.

References

R. Bonnema. 2002. Probability models for DOP. In *Data-Oriented Parsing*. CSLI publications.

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

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June.

System	Output
PBMT	fujian is china 's coastal areas most rapid development of foreign trade of the region .
SPMT-Combined	china 's coastal areas of fujian is one of the areas of the most rapid development of foreign trade and economic cooperation .
PBMT	investment in macao has become the largest foreign investors .
SPMT-Combined	the chinese - funded enterprises have become the largest foreign investor in macao.
PBMT	they are now two people were unaccounted for .
SPMT-Combined	currently , both of them remain unaccounted for .
PBMT	there was no further statement .
SPMT-Combined	the statement did not explain further .

Table 3: Sample translations.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 29(4):589–637, December.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL'2004: Main Proceedings*, pages 273–280, Boston, Massachusetts, USA, May 2 - May 7.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inferences and training of context-rich syntax translation models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'2006)*, Sydney, Australia, July.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, volume 1, pages 181–184.

Kevin Knight and Jonathan Graehl. 2005. An overview of probabilistic tree transducers for natural language processing. In *Proc. of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'2005)*, pages 1–25. Springer Verlag.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'2003)*, Edmonton, Canada, May 27–June 1.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2002)*, pages 133–139, Philadelphia, PA, July 6-7.

Evgeny Matusov, Nicola Ueffi ng, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL'2006)*, Trento, Italy.

Franz Joseph Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), December.

Franz Joseph Och. 2003. Minimum error training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'2003)*, pages 160–167, Sapooro, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpus-based 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, March 24-27.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'2005)*, pages 271–279, Ann Arbor, Michigan, June.

Radu Soricut. 2005. A reimplementation of Collins's parsing models.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7.

Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. 2006. Syncronous binarization for machine translation. In *Proceding of the Human Language Technology and North American Chapter of the Association for Computational Linguistics (HLT-NAACL'2006)*, New York, June.