Two Methods for Extending Hierarchical Rules from the Bilingual Chart Parsing

Martin Čmejrek and **Bowen Zhou**
IBM T. J. Watson Research Center
{martin.cmejrek, zhou}@us.ibm.com

Abstract

This paper studies two methods for training hierarchical MT rules independently of word alignments. Bilingual chart parsing and EM algorithm are used to train bitext correspondences. The first method, rule arithmetic, constructs new rules as combinations of existing and reliable rules used in the bilingual chart, significantly improving the translation accuracy on the German-English and Farsi-English translation task. The second method is proposed to construct additional rules directly from the chart using inside and outside probabilities to determine the span of the rule and its non-terminals. The paper also presents evidence that the rule arithmetic can recover from alignment errors, and that it can learn rules that are difficult to learn from bilingual alignments.

1 Introduction

Hierarchical phrase-based systems for machine translation usually share the same pattern for obtaining rules: using heuristic approaches to extract phrase and rule pairs from word alignments. Although these approaches are very successful in handling local linguistic phenomena, handling longer distance reorderings can be more difficult. To avoid the combinatorial explosion, various restrictions, such as limitations of the phrase length or non-terminal span are used, that sometimes prevent from extracting good rules. Another reason is the deterministic nature of those heuristics that does not easily recover from errors in the word alignment.

In this work, we learn rules for hierarchical phrase based MT systems directly from the parallel data, independently of bilingual word alignments.

Let us have an example of a German-English sentence pair from the Europarl corpus (Koehn, 2005).

(1) GER: die herausforderung besteht darin diese systeme zu den besten der welt zu machen
ENG: the challenge is to make the system the very best

The two pairs of corresponding sequences *diese systeme ... der welt—the system ... best* and *zu machen—to make* are swapped. We believe that the following rule could handle long distance reorderings, still with a reasonably low number of terminals, for example:

(2) $X \rightarrow \langle$besteht darin X_1 zu X_2, is to $X_2 X_1 \rangle$,

There are 127 sentence pairs out of 300K of the training data that contain this pattern, but this rule was not learned using the conventional approach (Chiang, 2007). There are three potential risks: (1) alignment errors (the first *zu* aligned to *to*, or *der welt* (*of the world*) aligned to null); (2) maximum phrase length for extracting rules lower than 11 words; (3) requirement of non-terminals spanning at least 2 words.

The *rule arithmetic* (Cmejrek et al., 2009) constructs the new rule (2) as a combination of good rule usages:

(3) $X \rightarrow \langle$besteht darin, is $\rangle$
$X \rightarrow \langle X_1$ zu X_2, to $X_2 X_1 \rangle$

Coling 2010: Poster Volume, pages 180–188,
Beijing, August 2010

The approach consists of bilingual chart parsing (BCP) of the training data, combining rules found in the chart using a *rule arithmetic* to propose new rules, and using EM to estimate rule probabilities.

In this paper, we study the behavior of the rule arithmetic on two different language pairs: German-English and Farsi-English. We also propose an additional method for constructing new rules directly from the bilingual chart, and compare it with the rule arithmetic.

The paper is structured as follows: In Sec. 1, we explain our main motivation, summarize previous work, and briefly introduce the formalism of hierarchical phrase-based translation. In Sec. 2, we describe the bilingual chart parsing and the EM algorithm. The rule arithmetic is introduced in Sec. 3. The new method for proposing new rules directly from the chart is described in Sec. 4. The experimental setup is described in Sec. 5. Results are thoroughly discussed in Sec. 6. Finally, we conclude in Sec. 7.

1.1 Related work

Many previous works use the EM algorithm to estimate probabilities of translation rules: Wu (1997) uses EM to directly estimate joint word alignment probabilities of Inversion Transduction Grammar (ITG). Marcu and Wong (2002) use EM to estimate joint phrasal translation model (JPTM). Birch et al. (2006) reduce its complexity by using only concepts that match the high-confidence GIZA++ alignments. Similarly, Cherry and Lin (2007) use ITG for pruning. May and Knight (2007) use EM algorithm to train tree-to-string rule probabilities, and use the Viterbi derivations to re-align the training data. Huang and Zhou (2009) use EM to estimate conditional rule probabilities $P(\alpha|\gamma)$ and $P(\gamma|\alpha)$ for Synchronous Context-free Grammar. Others try to overcome the deterministic nature of using bilingual alignments for rule extraction by sampling techniques (Blunsom et al., 2009; DeNero et al., 2008). Galley et al. (2006) define minimal rules for tree-to-string translation, merge them into composed rules (similarly to the rule arithmetic), and train weights by EM. While in their method, word alignments are used to define all rules, rule arithmetic proposes new rules independently of word alignments. Similarly, Liu and Gildea (2009) identify matching long sequences ("big templates") using word alignments and "liberate" matching small subtrees based on chart probabilities. Our method of proposing rules directly from the chart does not use word alignment at all.

1.2 Formally syntax-based models

Our baseline model follows the Chiang's hierarchical model (Chiang, 2007; Chiang, 2005; Zhou et al., 2008) based on Synchronous Context-free Grammar (SCFG). The rules have form

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle, \tag{4}$$

where X is the only non-terminal in the grammar, γ and α are source and target strings with terminals and up to two non-terminals, $\sim$ is the correspondence between the non-terminals. Corresponding non-terminals have to be expanded at the same time.

2 Bilingual chart parsing and EM algorithm

In this section, we briefly overview the algorithm for bilingual chart parsing and EM estimation of SCFG rule features.

Let $\mathbf{e} = e_1^M$ and $\mathbf{f} = f_1^N$ of source and target sentences. For each sentence pair $\mathbf{e}, \mathbf{f}$, the 'E' step of the EM algorithm will use the bilingual chart parser to enumerate all possible derivations Φ, compute inside probabilities $\beta_{ijkl}(X)$ and outside probabilities $\alpha_{ijkl}(X)$, and finally calculate expected counts $c(r)$ how many times each rule r produced the corpus C.

The inside probabilities can be defined recursively and computed dynamically during the chart parsing:

$$\beta_{ijkl} = \sum_{\rho \in t_{ijkl}} P(\rho.r) \prod_{(i'j'k'l') \in \rho.bp} \beta_{i'j'k'l'}, \tag{5}$$

where t_{ijkl} represents the chart cell spanning (e_i^j, f_k^l), and the data structure ρ stores the rule $\rho.r$. If r has non-terminals, then $\rho.bp$ stores backpointers $\rho.bp_1$ and $\rho.bp_2$ to the cells representing their derivations.

The outside probabilities can be computed recursively by iterating the chart in top-down ordering. We start from the root cell $\alpha_{1,M,1,N} := 1$ and propagate the probability mass as

$$\alpha_{\rho.bp_1} += P(\rho.r)\alpha_{ijkl} \quad (6)$$

for rules with one non-terminal, and

$$\alpha_{\rho.bp_1} += P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp_2}, \quad (7)$$
$$\alpha_{\rho.bp_2} += P(\rho.r)\alpha_{ijkl}\beta_{\rho.bp_1}, \quad (8)$$

for rules with two non-terminals. The top-down ordering ensures that each α_{ijkl} accumulates updates from all cells higher in the chart before its own outside probability is used.

The contributions to the rule expected counts are computed as

$$c(\rho.r) += \frac{P(\rho.r)\alpha_{ijkl}\prod_{i=1}^{\rho.n}\beta_{\rho.bp_i}}{\beta_{1,M,1,N}}. \quad (9)$$

Finally, rule probabilities $P(r)$ are obtained by normalizing expected counts in the 'M' step.

To improve the grammar coverage, the ruleset is extended by the following rules providing "backoff" parses and scoring for the SCFG rules:

(10) $\langle X_1, X_1 f\rangle$, $\langle X_1, f X_1\rangle$, $\langle X_1 e, X_1\rangle$, $\langle e X_1, X_1\rangle$,

(11) $\langle X_1 X_2, X_2 X_1\rangle$.

Rules (10) enable insertions and deletions, while rule (11) allows for aligning swapped constituents in addition to the standard glue rule.

3 Proposing new rules with rule arithmetic

The main idea of this work is to propose new rules independently of the bilingual word alignments. We parse each sentence pair using the baseline ruleset extended by the new rule types (10) and (11). Then we select the *most promising* rule usages and combine each two of them using the *rule arithmetic* to propose new rules. We put the new rules into a temporary pool, and parse and compute probabilities and expected counts again, this time we use rules from the baseline and from the temporary pool. Finally, we dump expected counts for proposed rules, and empty the temporary pool. This way we can try to propose many rules for each sentence pair, and to filter them later using accumulated expected counts from the EM.

The term *most promising* is purposefully vague — to cover all possible approaches to filtering rule usages. In our implementation, we are limited by space and time, and we have to prune the number of rules that we can combine. We use expected counts as the main scoring criterion. When computing the contributions to expected counts from particular rule usages as described by (9), we remember the n-best contributors, and use them as candidates after the expected counts for the given sentence pair have been estimated.

The *rule arithmetic* combines existing rules using *addition* operation to create new rules. The idea is shown in Example 12.

(12) Addition

$\langle 5,13,5,11,13,13\rangle$ $\langle 4,10,6,10,5,5\rangle$ $X \to \langle X_1$ zu X_2, to X_2 $X_1\rangle$
$\langle 5,11,6,11,0,0\rangle$ $\langle 6,10,7,10,0,0\rangle$ $X \to \langle$diese X_1, the $X_1\rangle$

1:	...	4	5	6	...	11	12	13	3	4	5	6	7	...	10
2:	...	0	-1	-1	...	-1	zu	-2	0	to	-2	-1	-1	...	-1
3:	...	0	diese	-3	...	-3	0	0	0	0	0	the	-3	...	-3
4:	...	0	diese	-3	...	-3	zu	-2	0	to	-2	the	-3	...	-3

5: $\langle 5,13,6,11,13,13\rangle$ $\langle 4,10,7,10,5,5\rangle$ $X \to \langle$diese X_1 zu X_2, to X_2 the $X_1\rangle$

First, create span projections for both source and target sides of both rules. Use symbol 0 for all unspanned positions, copy terminal symbols as they are, and use symbols -1, -2, -3, and -4 to transcribe X_1 and X_2 from the first rule, and X_1 and X_2 from the second rule. Repeat the non-terminal symbol on all spanned positions. In Example 12 line 1 shows the positions in the sentence, lines 2 and 3 show the rule span projections of the two rules.

Second, merge source span projections (line 4), record mappings of non-terminal symbols. We require that merged projections are *continuous*. We allow substituting non-terminal symbols by terminals, but we require that the whole span of the non-terminal is fully replaced. In other words, shortenings of non-terminal spans are not allowed.

Third, collect new rule. The merged rule usages (lines 5) are generalized into rules, so that they are not limited to the particular span for which they were originally proposed.

The rule arithmetic can combine all types of rules – phrase pairs, abstract rules, glues, swaps, insertions and deletions. However, we require that

at least one of the rules is either a phrase pair or an abstract rule.

4 Proposing directly from chart

One of the issues observed while proposing new rules with the rule arithmetic is the selection of the best candidates. The number of all candidates that can be combined depends on the length of the sentence pair and on the number of competing parsing hypotheses. Using a fixed size of the n-best can constitute a risk of selecting bad candidates from shorter sentences. On the other hand, the spans of the best candidates extracted from long sentences can be far from each other, so that most combinations are not valid rules (e.g., the combination of two discontinuous phrasal rules is not defined).

In our new approach we propose new rules directly from the bilingual chart, relying on the inside and outside probabilities computed after the parsing of the sentence pair. The method has two steps. In the first step we identify best matching parallel sequences; in the second step we propose "holes" for non-terminals.

4.1 Identifying best matching sequences

To identify the best matching sequences, we score all sequences (e_i^j, f_k^l) by a scoring function:

$$score_{ijkl} = \frac{\alpha_{ijkl}\beta_{ijkl}}{\beta_{1,M,1,N}} Lex(i,j,k,l), \quad (13)$$

where the lexical score is defined as:

$$Lex(i,j,k,l) = \sum_{j'=1}^{N} \prod_{i'=0}^{M} t(f_{j'}|e_{i'})\delta_{ijkli'j'} \quad (14)$$

The t is the lexical probability from the word-to-word translation table, and $\delta_{ijkli'j'}$ is defined as δ_{ins} if $i' \in \langle i,j \rangle$ and $j' \in \langle k,l \rangle$, and as δ_{out} if $i' \notin \langle i,j \rangle$ and $j' \notin \langle k,l \rangle$, and as 0 elsewhere. The purpose of this function is to score only the pairs of words that are both either from within the sequence or from outside the sequence. Usually $0 \leq \delta_{out} \leq \delta_{ins}$ to put more weight on words within the parallel sequence.

The scoring function is a combination of expected counts contribution of a sequence (e_i^j, f_k^l) estimated from the chart with the IBM Model 1 lexical score.

Since only the sequences spanned by filled chart cells can have non-zero expected counts, we can select the n-best matching sequences relatively efficiently.

4.2 Proposing non-terminal positions

Similar approach can be used to propose best positions for non-terminals. We score every combination of non-terminal positions. The expected counts can be estimated using Eq. 9. Since we are proposing new rules, the probability $P(r)$ used in that equation is not defined. Again, we can use Model 1 score instead, and use the following scoring function:

$$s_{ijkl}(bp_1, bp_2) = \frac{Lex(i,j,k,l,bp_1,bp_2)\alpha_{ijkl}\beta_{bp_1}\beta_{bp_2}}{\beta_{1,M,1,N}}, \quad (15)$$

$Lex(i,j,k,l,bp_1,bp_2)$ is defined as in Eq. 14. This time using $0 \leq \delta_{out} \leq \delta_{NT1} = \delta_{NT2} \leq \delta_{term}$, restricting the IBM Model 1 to score only word pairs that both belong either to the terminals of the proposed rule, or to the sequences spanned by the same non-terminal, or outside of the rule span. The scoring function for rules with one non-terminal is just a special case of 15.

Again, the candidates can be scored efficiently, taking into account only those combinations of non-terminal spans that correspond to filled cells in the chart.

The proposed method is again independent of bilingual alignment, but at the same time utilizes the information obtained from the bilingual chart parsing.

5 Experiments

We carried out experiments on two language pairs, German-English and Farsi-English.

The **German-English** data is a subset (297k sentence pairs) of the Europarl (Koehn, 2005) corpus. Since we are focused on speech-to-speech translation, the punctuation was removed, and the text was lowercased. The dev set and test set contain each 1k sentence pairs with one reference.

The word alignments were trained by GIZA++ toolkit (Och and Ney, 2000). Phrase pairs were

extracted using grow-diag-final (Koehn et al., 2007). The baseline ruleset was obtained as in (Chiang, 2007). The maximum phrase length for rule extraction was set to 10, the minimum required non-terminal span was 2.

Additional rules for insertion, deletion, and swap were added to improve the parsability of the data, and to help EM training and rule arithmetic. However, these rules are not used by the decoder, since they would degrade the performance.

New rules were proposed after the first iteration of EM[1], either by rule arithmetic or directly from the chart.

Only non-terminal rules proposed by the rule arithmetic from at least two different sentence pairs and ranked (by expected counts $c(r)$) in the top 100k were used. Figure 4 presents a sample of the new rules.

New rules were also proposed directly from the chart, using the approach in Sec. 4. 5% of best matching parallel sequences, and 5 best scoring rules were selected from each parallel sequence. Non-terminal rules from the 200k-best rank were added to the model. Figure 5 presents a sample of the new rules.

Finally, one more iteration of EM was used to adjust the probabilities of the new and baseline rules. These probabilities were used as features in the decoding.

The performance of rule arithmetic was also verified on **Farsi-English** translation. The training corpus contains conversational spoken data from the DARPA TransTac program extended by movie subtitles and online dictionaries downloaded from the web (297k sentence pairs). The punctuation was removed, and the text was lowercased. The dev set is 1,420 sentence pairs held out from the training data, with one reference. The test set provided by NIST contains 470 sentences with 4 references. The sentences are about 30% longer and more difficult.

The training pipeline was the same as for the German-English experiments. 122k new non-terminal rules were proposed using the rule arithmetic.

[1]Since our initial experiments did not show any significant gain from proposing rules after additional (lengthy) iterations of EM.

The feature weights were tuned on the dev set for each translation model separately. The translation quality was measured automatically by BLEU score (Papineni et al., 2001).

6 Discussion of results

The BLEU score results are shown in the Table 3. The cumulative gain of rule arithmetic and EM (RA + EM-i0) is 1 BLEU point for German-English translation and 2 BLEU points for Farsi-English. The cumulative gain of rules proposed from the chart (DC + EM-i0) is 0.2 BLEU points for German-English. For comparison of effects of various components of our method, we also show scores after the first five iterations of EM (EM-i0–EM-i4) without adding any new rules, just using EM-trained probabilities as feature weights, and also scores for new rules added into the baseline without adjusting their costs by EM (RA).

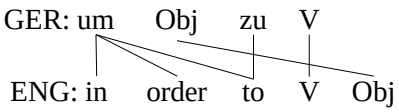
The qualities of proposed rules are discussed in this section.

6.1 German-English rules from rule arithmetic

The Figure 4 presents a sample of new rules proposed during this experiment. The table is divided into three parts, presenting rules from the top, middle, and bottom of the 100K list. The quality of the rules is high even in the middle part of the table, the tail part is worse.

We were surprised by seeing short rules consisting of frequent words. For example $\langle$um X_1, in order $X_1\rangle$. When looking into word-level alignments, we realized that these rules following the pattern 16 prevent the baseline approach from extracting the rule.

(16)

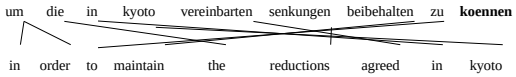


Similarly many other rules match the pattern of beginning of a subordinated clause, such as *that is why*, or insertions, such as *of course*, which both have to be strictly followed by VSO construction in German, in contrast to the SVO word order in English.

We also studied the cases of rule arithmetic correcting for systematic word alignment errors. For

example the new rule $\langle X_1$ zu koennen, to $X_1\rangle$ was learned from the sentence

(17)



The English translation often uses a different modality, thus the modal verb *koennen* is always aligned with null. Since unaligned words are usually not allowed at the edges of sub-phrases generalized into non-terminals (Chiang, 2007), this rule cannot be learned by the baseline.

We observe that many new proposed rules correspond to patterns with a non-terminal spanning one word. For example $\langle$um X_1 zu X_2, to X_2 $X_1\rangle$ corresponds to the same pattern 16, where X_2 spans one verb. The line *baseline min1* in the Table 3 shows 0.3 BLEU improvement of a model trained without the minimum non-terminal span requirement. However, this improvement comes at a cost of more than four times increased model size, as shown in Table 2. We observe that using the minimum span requirement while learning from bitext alignments combined with rule arithmetic that can learn the most reliable rules spanning one word yields better performance in speed, memory, and precision.

We can also study the new rules quantitatively. We want to know how the rules proposed by the rule arithmetic are used in decoding. We traced the translation of the 1,000 test set sentences to mark the rules that were used to generate the best scoring hypotheses.

The stats are presented in the Table 1. The chance that a new rules will be used in the test set decoding (0.86%) is more than 7 times higher than that of all rules (0.12%). Encouraging evidence is that while the rule arithmetic rules constitute only 1.87% of total rules, they present 9.17% of rules used in the decoding.

The Figure 1 lists the most frequently used new rules in the decoding. We can see many rules with 2 non-terminals that model complex verb forms ($\langle$wird X_1 haben,will have $X_1\rangle$), reordering in clauses ($\langle$um X_1 zu gewaehrleisten, to ensure $X_1\rangle$), or reordering of verbs from the second position in German to SVO in English ($\langle$heute X_1 wir X_2, today we X_1 $X_2\rangle$).

	RA Ger.	DC Ger.	RA Farsi
Sentences translated	1,000	1,000	417
|ALL| (all rules)	5.359,751	5.459,751	8.532,691
|NEW| (new rules)	100,000	200,000	121,784
$\frac{\text{\|NEW\|}}{\text{\|ALL\|}}$	1.87%	3.66%	1.43%
|hits ALL|	10,122	7,256	2,521
|glue|	2,910	271	267
|hits ALL unique|	6.303	6,433	2,058
$\frac{\text{\|hits ALL unique\|}}{\text{\|ALL\|}}$	0.12%	0.12%	0.02
|hits NEW|	928	1,541	125
|hits NEW unique|	858	1,504	110
$\frac{\text{\|hits NEW unique\|}}{\text{\|NEW\|}}$	0.86%	0.75 %	0.09
$\frac{\text{\|hits NEW\|}}{\text{\|hits ALL\|}}$	9.17%	21.23%	4.96%
|terminals from NEW|	4,385	7,825	407
$\frac{\text{\|terminals from NEW\|}}{\text{\|hits NEW\|}}$	4,73	5.08	3.26

Table 1: Rule hits for 1,000 test set.

Model	#phrases	#rules
Ger-Eng baseline	8.5M	5.3M
Ger-Eng baseline min1	8.5M	23.M

Table 2: Model sizes.

We also studied the correlation between the rank of the proposed rules (ranked by expected counts) and the hit rate during the decoding. The Figure 2 measures the hit rate for each of 1,000 best ranking rules, and should be read as follows: the rules ranking 0 to 999 were used 70 times, the hit rate decreases as the rank grows so that there were no hits for rules ranking 90k and more. The rank is a good indicator of the usefulness of new rules.

We hypothesize that the new rules are capable of combining partial solutions to form hypotheses with better word order, or better complex verb forms so that these hypotheses are better scored and are parts of the winning solutions more often.

6.2 German-English rules proposed directly from the chart

We also studied why the rules proposed directly from the bilingual chart yield smaller improvement than the rule arithmetic. The number of new rules used in the decoding (1,541) is even higher than that of the rule arithmetic, and it constitutes 21.23% of all cases. The two experiments were

#hits	Ger	Eng
5	X_1 stellt X_2 dar	X_1 is X_2
3	X_1 sowohl X_2 als auch	X_1 both X_2 and
3	X_1 ist es X_2	it is X_2 X_1
3	X_1 die X_2 ist	X_1 which is X_2
2	wird X_1 haben	will have X_1
2	wir X_1 damit X_2	we X_1 so that X_2
2	was X_1 hat X_2	what X_1 has X_2
2	was X_1 betrifft so	as regards X_1
2	und X_1 muessen wir X_2	and X_1 we must X_2
2	um X_1 zu gewaehrleisten	to ensure X_1
2	um X_1 zu X_2	to X_2 X_1
2	sowohl X_1 als auch	both X_1 and
2	sie X_1 auch X_2	they also X_1 X_2
2	in erster linie X_1	X_1 in the first instance
2	in X_1 an	in X_1
2	ich X_1 meine	i X_1
2	heute X_1 wir X_2	today we X_1 X_2
2	herr praesident X_1 und herren	mr president X_1 and gentlemen
2	gleich X_1	X_1 a moment
2	es muss X_1 werden	it must be X_1

Figure 1: Examples of the most frequently h rules during the decoding.

tuned separately, so that they used different glu rule weights. That is why we observe the differ ence in the number of glues (and the number c total rules) in the Table 1. We do not observe an significant correlation between the rank of the rul and the hit rate. The Figure 3 shows that the first 10k-ranked rules are hit several times, and then the hit rate stays flat.

We offer an explanation based on our observations of rules used for the decoding. The rules proposed directly from the chart contain a big portion of content words. These rules do not capture any important differences between the structures of the two languages that could not be handled by phrasal rules as well. For example, the rule ⟨die neuen vorschriften sollen X_1,the new rules are X_1⟩ is correct, but a combination of a baseline phrasal rule and glue will produce the same result.

We also see many rules with non-terminals spanning one word. For example, the sequence

(18) die europaeische kommission—the european commission

will produce the rule

(19) ⟨die X_1 kommission, the X_1 commission⟩.

Although the sequence and the rule are high scored by 13 and 15, we intuitively feel that gen-

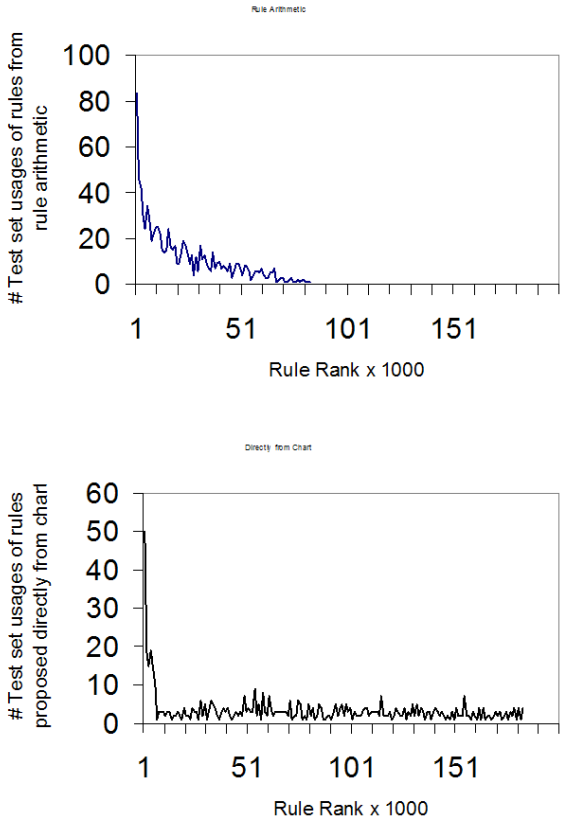


Figure 3: Usage of new rules (DC).

eralizing the word *european* is not very helpful in this context.

The rule arithmetic could propose the rule 19 as

(20) ⟨die X_1, the X_1⟩ + ⟨kommission, commission⟩,

but since the candidates for combination are selected as rules with the highest expected counts (Sec. 3), the rules 20 will most likely loose to the phrase pair 18 and will not be selected.

To conclude our comparison, we observe that both methods produce reliable rules that are often reused in decoding. Nevertheless, since the rule arithmetic combines the most successful rules from each parallel parse, the resulting rules enable structural transformations that could not be handled by baseline rules.

	German-English		Farsi-English	
Model	dev set	test set	dev set	test set
baseline	23.9	25.4	41.1	38.2
RA + EM-i0	24.8	**26.4**	41.8	**40.2**
DC + EM-i0	24.6	25.6		
EM-i0	24.4	26.1	40.8	39.1
EM-i1	24.4	25.8	41.3	38.5
EM-i2	24.4	25.9	41.4	38.2
EM-i3	24.4	26.0	41.3	39.3
EM-i4	24.4	26.0	41.6	39.6
RA	24.4	26.1	40.7	38.4
baseline min1	24.0	25.7		

Table 3: BLEU scores

um X_1	in order X_1
natuerlich X_1	of course X_1
deshalb X_1	this is why X_1
X_1 zu koennen	to X_1
X_1 ist	it is X_1
nach der tagesordnung folgt die X_1	the next item is the X_1
herr X_1 herr kommissar X_2	mr X_1 commissioner X_2
die X_1 der X_2	X_1 the X_2
im gegenteil X_1	on the contrary X_1
nach der tagesordnung folgt X_1	the next item is X_1
X_1 die X_2	the X_1 the X_2
die X_1 die	the X_1
ausserdem X_1	in addition X_1
daher X_1	that is why X_1
wir X_1 nicht X_2	we X_1 not X_2
die X_1 der X_2	the X_2 X_1
deshalb X_1	for this reason X_1
um X_1 zu X_2	to X_2 X_1
X_1 nicht X_2 werden	X_1 not be X_2

Figure 4: Sample rules (RA).

ausserdem X_1 wir	we X_1 also
die X_1 des kommissars	the commissioner 's X_1
den X_1 ratsvorsitz	the X_1 presidency
ich hoffe dass X_1	i would hope that X_1
X_1 ist zu X_2 geworden	X_1 has become X_2
die X_1 des vereinigten koenigreichs	the uk X_1
X_1 maij weggen X_2	X_1 maij weggen X_2
X_1 wir auf X_2 sind	X_1 we are on X_2
ich frage mich X_1	i wonder X_1

Figure 5: Sample rules (DC).

6.3 Farsi-English rules from the rule arithmetic

Although we have only limited resources to qualitatively analyze the Farsi-English experiments, we noticed that there are two major groups of new rules.

The first group corresponds to the fact that Farsi does not have definite article and allows pro-drop. We observe many new rules that could not be learned from word alignments, since some definite articles or pronouns in English were aligned to null (and unaligned words are not allowed at the edges of phrases). However, if the chart contains an insertion (of the determiner or pronoun) with a high expected count, the rule arithmetic may propose new rule by combining it with other rules.

The second group contains rules that help word reordering. We observe rules moving verbs from the S PP O V in Farsi into SVO in English as well as rules reordering wh-clauses.

Most of the rules traced during the test set decoding belong to the second group. Figure 1 shows that the number of new rules hit during the decoding is smaller compared to the German-English experiments. On the other hand, the rules have smaller number of terminals so that we assume that the positive effect of these rules comes from the reordering of non-terminals.

7 Conclusion

In this work, we studied two new methods for learning hierarchical MT rules: the rule arithmetic and proposing directly from the parse forest. We discussed systematic patterns where the rule arithmetic outperforms alignment-based approaches and verified its significant improvement on two different language pairs (German-English and Farsi-English). We also hypothesized why the second method – proposing rules directly from the chart – improves the baseline less than the rule arithmetic.

Acknowledgment

This work is partially supported by the DARPA TRANSTAC program under the contract number NBCH2030007. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Birch, Alexandra, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proceedings on WSMT'06*, pages 154–157.

Blunsom, Phil, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A gibbs sampler for phrasal synchronous grammar induction. In *ACL '09*, pages 782–790.

Cherry, Colin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *NAACL-HLT'07/SSST'07*.

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

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

Cmejrek, Martin, Bowen Zhou, and Bing Xiang. 2009. Enriching SCFG rules directly from efficient bilingual chart parsing. In *IWSLT'09*, pages 136–143.

DeNero, John, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a bayesian translation model. In *EMNLP '08*, pages 314–323.

Galley, Michel, 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 *Proc. of ACL*, pages 961–968.

Huang, Songfang and Bowen Zhou. 2009. An EM algorithm for SCFG in formal syntax-based translation. In *Proc. IEEE ICASSP'09*, pages 4813–4816.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Liu, Ding and Daniel Gildea. 2009. Bayesian learning of phrasal tree-to-string templates. In *EMNLP '09*, pages 1308–1317.

Marcu, Daniel and W Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP'02*.

May, Jonathan and Kevin Knight. 2007. Syntactic realignment models for machine translation. In *Proceedings of EMNLP-CoNLL'07*, pages 360–368.

Och, F. J. and H. Ney. 2000. Improved statistical alignment models. In *Proc. of ACL*, pages 440–447, Hong Kong, China, October.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176, IBM T. J. Watson Research Center.

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

Zhou, Bowen, Bing Xiang, Xiaodan Zhu, and Yuqing Gao. 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. In *Proceedings of the ACL'08: HLT SSST-2*, pages 19–27.