Monotonic filter for hierarchical translation models

Shahram Salami, Mehrnoush Shamsfard
Faculty of Computer Science and Engineering,
Shahid Beheshti University,
Tehran, Iran
sh_salami@sbu.ac.ir, m-shams@sbu.ac.ir

***Abstract*—The model size and decoding time are known issues in statistical machine translation. Especially, monotonic words order of language pairs makes the size of hierarchical models huge. Considering this fact, the rule extraction method of phrase-boundary model was changed to extract less number of rules. This paper proposes this rule extraction method as a general filter for hierarchical models. Named as monotonic filter, this filter reduces the extracted rules from phrase pairs decomposable to monotonic aligned subphrases. We apply the monotonic filter on the hierarchical phrase-based, SAMT and phrase-boundary models. Our experiments are performed in translations from Persian, German and French to English as the source and target languages with low, medium and high monotonic word order respectively. The reduction amount of the monotonic filter for the model size and decoding time is up to about 70% and 80% respectively, in most cases with no tangible impact on the translation quality.**

Keywords: Statistical machine translation; Hierarchical rules filtering; Filtered rule extraction; Phrase decomposition pattern

I. Introduction

The availability of data allows building larger models to achieve more translation quality in statistical machine translation (SMT), while larger models decrease the efficiency of SMT decoders. On the other side, phrase-based models only use lexical phrases as strings of words but hierarchical models also use hierarchical phrases that include both words and subphrase variables (or nonterminals). Utilizing hierarchical phrases can improve the word order in the translation but, as a side effect, it dramatically increases the size of hierarchical models.

Hierarchical phrases are defined by substitution of subphrases with nonterminals in the aligned phrase pairs. The monotonic word order of aligned phrases results in a larger size of the hierarchical models due to existence of many monotonic aligned subphrases as the placeholders of nonterminals. On the other side, more diversity of nonterminals increases the size of hierarchical models due to the lower convergence of labels for different phrase pairs. Filtering the grammar rules in hierarchical models is necessary to keep the efficiency of decoding in both time and memory usage.

To decrease the size of hierarchical models, some filtering methods have been proposed which can be categorized in two approaches. One approach changes the rule extraction method to prevent the extraction of unnecessary rules based on the word alignments (their patterns or probabilities) or the possibility of minimal derivations. Another approach discards unnecessary rules from the extracted grammar based on the importance of the rules (their number of evidences, patterns or probabilities) or the information redundancy in the model. Generally, the first mentioned approach is preferred because in addition to decoding time, it reduces the training resources required for the grammar extraction. Both approaches can be applied together for more reduction of decoding resources.

Phrase-boundary model was proposed [1] as a hierarchical model which labels rules with the classes of boundary words on the target side phrases. Considering the diversity of nonterminals, the rule extraction method of this model was changed to extract less number of rules from phrase pairs with monotonic word order. We generalize the use of their rule extraction method (in the first mentioned filtering approach) named as *monotonic filter* for the hierarchical translation models.

In this paper, we discuss the monotonic filter and its coverage in general and investigate the efficiency of the monotonic filter with the hierarchical phrase-based [2], variants of SAMT [3] and phrase-boundary [1] models that label rules with one generic nonterminal, categories of target side syntax trees and classes of boundary words on the target side phrases, respectively. The experiments are performed in translations from Persian, German and French to English having low, medium and high amount of monotonic word order respectively. The results show that the monotonic filter reduces the model size and decoding time effectively with no tangible impact on translation quality of these models; the hierarchical phrase-based, a variant of SAMT and phrase-boundary.

The performance of the monotonic filter as a general filtering method is explained in this paper. Related work is referenced in Section II. The monotonic filter is explained in Section III. Section IV discusses the coverage of filtered models. Section V shows the results of experiments. Finally, the paper is concluded in section VI.

978-1-5090-3586-1/16/$31.00

II. Related work

In the filtering approach that eliminates unnecessary rules from the extracted grammar, reference [4] discarded rare hierarchical rules occurred fewer times than a given threshold. They showed that increasing the threshold decreases the translation quality. The best threshold value is dependent on the training corpus. Hierarchical phrase-based model was filtered by discarding hierarchical rules whose source sides appear only in the monotone composed rules [5]. The yield of a monotone composed rule can be obtained by concatenating the yields of the minimal rules. Glue rules in the grammar support serial concatenation of the output phrases. The rules was categorized to different patterns and filtered out those patterns which could be discarded without significant impact on the translation quality [6]. The rules was discarded based on the information redundancy encoded in the translation rules [7].

In the filtering approach in which the extraction of too many rules is prevented by changing the method of rule extraction, the Alignment probability of the words was examined to restrict the extraction of rules to aligned phrase pairs with the higher probability [8]. A minimum set of translation rules was extracted on which at least one derivation could be constructed for each phrase pair [9]. The monotonic filter was proposed for phrase-boundary model based on the alignment pattern of phrase pairs [1]. This filter cuts down the patterns of hierarchical rules extracted from phrase pairs which are decomposable to monotonic aligned subphrases. Decomposition pattern of phrase pairs is also used to label hierarchical rules without linguistic resources [10].

Although, the base idea of the monotonic filter [1] is similar to the [5], the monotonic filter extracts rules from all phrase pairs. Thus, the monotonic filtered model is dependent on the glue rules less than their filtered model to encourage more use of hierarchical rules. In addition, the monotonic filter reduces the required resources for the grammar extraction, too.

The translation table of phrase-based models contains all phrase pairs found in the parallel corpus. Although, phrase-based models are far smaller than hierarchical ones, their translation table is pruned to discard noisy translations and to decrease the decoding time. Pruning of unlikely phrase pairs was suggested by significance testing of phrase pair co-occurrence in the parallel corpus [11]. Phrase pairs were discarded which are derivable using smaller ones with similar probabilities [12].

III. Monotonic filter

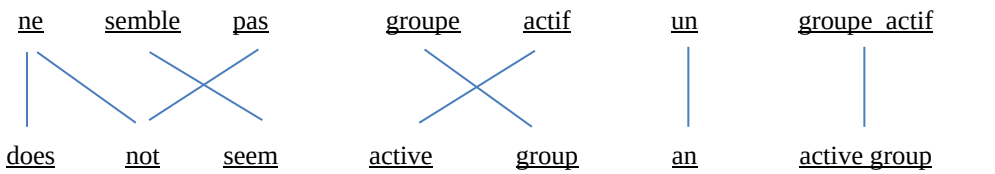
The monotonic filter is proposed for hierarchical translation models that use weighted rules in the Probabilistic Synchronous Context-Free Grammar (PSCFG). Synchronous grammar rules have the general form $X \rightarrow \langle \alpha, \beta, \sim \rangle$. In this form, X is the left hand side nonterminal, α and β are strings of terminals and nonterminals, and $\sim$ indicates a one-to-one correspondence between nonterminals in α and β. Using at most two nonterminals on the right hand side of the rules, source or target part of the rules (α or β) can be formed with one of the following patterns (where w_i is a string of words and x_i is a nonterminal):

all-patterns =
$\{w_1, x_1 x_2, x_1 w_1, w_1 x_1, w_1 x_1 w_2, x_1 w_1 x_2, x_1 x_2 w_1, w_1 x_1 x_2, w_1 x_1 x_2 w_2, x_1 w_1 x_2 w_2, w_1 x_1 w_2 x_2, w_1 x_1 w_2 x_2 w_3\}$ (1)

As hierarchical rules are defined by substitution of aligned subphrases with nonterminals, word alignment restricts the applicable patterns for one aligned phrase pair. For example, source side of phrase pair (a) in Fig. 1 can be only formulated as w_1 and $w_1 x_1 w_2$ (<*ne semble pas*> and <*ne* x_1 *pas*>). This phrase pair is not decomposable to aligned subphrases, as it has only one aligned subphrase. Note that contiguous spans of the source and target words with at least one aligned word constitute an aligned phrase pair containing no word aligned outside it.

Phrase pairs with monotonic word order have many aligned subphrases (with overlap) as placeholders of nonterminals in different patterns (1). Thus, many hierarchical rules are extractable from these phrase pairs. On the other side, simple patterns (such as $x_1 w_1$, $w_1 x_1$) are applicable for monotonic aligned phrase pairs. For example, source side of phase pair (c) in Fig. 1 can be represented as <x_1 *groupe actif*> or <*un* x_1>. According to this fact, the monotonic filter was proposed for filtered rule extraction with the phrase-boundary model based on the alignment pattern of phrase pairs [1]. This filter confines the pattern of rules to simple patterns for those extracted from phrase pairs which are decomposable to two monotonic aligned subphrases.

The monotonic filter accepts the rules which are candidate for extraction from the aligned phrase pair $\langle f_i^j, e_m^n \rangle$ (where f_i^j and e_m^n stand for the inclusive source and target substrings from position i to j and position m to n respectively) with the following conditions:



(a) Non-decomposable (b) 2-decomposable (c) Monotone 2-decomposable

Fig. 1. Different alignment patterns between French and English phrases

978-1-5090-3586-1/16/$31.00 ©2016 IEEE

1. Candidate rules consistent with one of the following patterns (a subset of (1)) in the source part:

   $Boundary_2 = \{w_1, x_1\, w_1, w_1\, x_1, x_1\, w_1\, x_2\}$ (2)

2. Other candidate rules where $<f_i^j, e_m^n>$ is not *monotone 2-decomposable*.

Phrase pair $<f_i^j, e_m^n>$ is monotone 2-decomposable if $\exists\, k \in [i,j), \exists\, \acute{k} \in [m,n) : <f_i^k, e_m^{\acute{k}}>$ and $<f_{k+1}^j, e_{\acute{k}+1}^n>$ are aligned phrase pairs. (3)

To say more, monotone 2-decomposable phrase pairs can be constituted by monotonic concatenation of aligned subphrases. For example, phrase pair (c) in Fig. 1 is monotone 2-decomposable. This phrase pair is composed of monotonic aligned subphrases <*un, a*> and <*groupe actif, active group*>. $Boundary_2$ filtering pattern (2) represents rules with at most two nonterminals in the source part boundaries. In other words, extracting all rules from other phrase pairs, the monotonic filter confines source part position of nonterminals to the boundaries in the hierarchical rules extracted from monotone 2-decomposable phrase pairs (excluding other patterns such as $w_1\, x_1\, w_2$). It seems that the patterns with one nonterminal ($x_1\, w_1$, $w_1\, x_1$) are sufficient to represent the decomposition of monotone 2-decomposable phrase pairs but the experiments showed [1] that adding pattern $x_1\, w_1\, x_2$ is useful to achieve more translation quality. This subject is more discussed in the next section.

For example, the following rules are candidate for extraction from phrase pair (c) in Fig. 1 in the hierarchical phrase-based model:

$X \rightarrow$ <un groupe actif, an active group> (4)

$X \rightarrow <\text{un } X^{\sim 1}, \text{an } X^{\sim 1}>$ (5)

$X \rightarrow <X^{\sim 1} \text{ groupe actif}, X^{\sim 1} \text{ active group}>$ (6)

$X \rightarrow <X^{\sim 1} \text{ groupe } X^{\sim 2}, X^{\sim 1}\, X^{\sim 2} \text{ group}>$ (7)

$X \rightarrow <\text{un } X^{\sim 1} \text{ actif, an active } X^{\sim 1}>$ (8)

As phrase pair (c) in Fig. 1 is monotone 2-decomposable, Rule 8 is not accepted by the monotonic filter due to its source side pattern ($w_1\, x_1\, w_2$) but the following rule is accepted for non-decomposable phrase pair (a) in that figure:

$X \rightarrow <\text{ne } X^{\sim 1} \text{ pas, does not } X^{\sim 1}>$ (9)

The monotonic filter reduces the model size considerably. According to this fact, most of the aligned phrase pairs are monotone 2-decomposable. Giving more preference to the rules with one of the $Boundary_2$ patterns in the source part, the filtered model achieves more translation quality. A penalty feature is added to the filtered grammar to penalize non preferred rules. This feature named *pattern penalty* is defined as following:

- Pattern penalty feature has the value of 0 if the rule is consistent with one of the $Boundary_2$ patterns in the source part; otherwise, it has the value of 1 to penalize other rules. (10)

Applying monotonic filter requires adding the pattern penalty feature to the filtered grammar. For example, this feature has the value of 0 for Rules 4 to 7 but it has the value of 1 for Rule 9.

IV. Filtered model coverage

As mentioned, the monotonic filter confines the source part of the rules accepted for monotone 2-decomposable phrase pairs to the $Boundary_2$ filtering pattern (2). The two part rules ($x_1\, w_1$ and $w_1\, x_1$) in the $Boundary_2$ pattern are expected to support constitution of monotone 2-decomposable phrase pairs but the number of tokens on the right hand side of hierarchical rules is conventionally limited to five (including nonterminals). Therefore, a long phrase pair may not be presented as $w_1\, x_1$ or $x_1\, w_1$ due to the length of string w_1.

On the other side, a monotone 2-decomposable phrase pair $<f_i^j, e_m^n>$ can be decomposed to monotone aligned subphrases $<f_i^k, e_m^{\acute{k}}>$ and $<f_{k+1}^j, e_{\acute{k}+1}^n>$ (3). The grammar of hierarchical models (such as those examined in this paper) contains binary glue rules (like rule 11) that support serial concatenation of these subphrases (*S* is the start symbol of grammar):

$$S \rightarrow <S^{\sim 1}\, X^{\sim 2}, S^{\sim 1}\, X^{\sim 2}> \quad (11)$$

Phrase pairs $<f_i^k, e_m^{\acute{k}}>$ and $<f_{k+1}^j, e_{\acute{k}+1}^n>$ can be derived from *S* and *X* respectively. Monotonic filter is only applied on *monotone* 2-decomposable phrase pairs which have no need for reordering of their subphrases. For example, this filter accepts all candidate rules for phrase pair (b) in Fig. 1. This phrase pair is decomposable to aligned subphrases but it needs reordering of subphrases. Although, glue rules are sufficient to compose monotone 2-decomposable phrase pairs, hierarchical rules support linguistic dependencies better than glue rules. The pattern with two nonterminals ($x_1\, w_1\, x_2$) in the $Boundary_2$ patterns increases the chance of deriving long phrase pairs with hierarchical rules.

V. Experiments

Series of translations are performed from Persian, German and French to English. Although, Persian is almost a free word order language, its formal sentence structure is SOV, which differs from the SVO structure of English. German and English in many cases have different word orders, too. For example, in German, infinitive verbs are generally placed after their respective objects. French and English have similar word orders and most of the word reordering is local.

978-1-5090-3586-1/16/$31.00 ©2016 IEEE

Table1. Statistics of the training corpora

Training corpora	No. of words	No. of word types
Persian-English	15M+13M	135K+98K
German-English	20M+21M	222K+76K
French-English	22M+20M	91K + 73K

Persian-English translation is trained on the Mizan corpus [13] with 1M sentences. The 1K+1K sentences of this corpus are selected for the development and test sets. German-English and French-English translations are trained on the first 1M sentences of Europarl-V7 corpus [14]. The last 500 sentences of WMT 2012 translation task are used for the development set, and the first 1K sentences of WMT 2013 translation task are used for the test set. The statistics of the used corpora are presented in Table 1.

The experiments are performed with the hierarchical phrase-based [2], variants of SAMT [3] and phrase-boundary [1] models which are configured with the default settings of these models. The models are trained and evaluated with Joshua toolkit [15]. The words are aligned in both directions of translation by GIZA++ [16] and the results are symmetrized [17]. The trigram language models on the target side of the training corpus are built by Berkeley LM tool [18] with Kneser-Ney smoothing. The scaling factors of the models are trained by Minimum Error Rate Training [19]. The BLEU-4 [20] is the metric of evaluations.

Hierarchical phrase-based model needs no linguistic tool for training but SAMT and phrase-boundary models label rules using target side syntax trees and word classes respectively. The target side syntax trees for SAMT are generated by Stanford parser [21]. The target side POS tags for phrase-boundary model are defined by SENNA tool [22] on the English side of the parallel training corpus.

In the following, we present the configurations of the experiments, the experiment results for the hierarchical phrase-based (HPB), SAMT and phrase-boundary models once without filtering and once with filtering. Finally, distribution of rules in the filtered models is analyzed in this section.

A. Configurations

The default settings of the hierarchical models limit lexical rules (rules having no nonterminal on the right hand side) to the length of 10 words and hierarchical rules to the length of 5 tokens, including a maximum of two nonterminals. The extraction of the hierarchical rules is limited to the span of 10 words for hierarchical phrase-based model and 12 words for others. Abstract rules (rules having no word) are not used in the grammars. The following default features for the grammar rules are used:

- Negative log of phrase probabilities: computed in both source-to-target and target-to-source directions.
- Negative log of lexical weights (Koehn et al., 2003): computed in both source-to-target and target-to-source directions.
- Rarity penalty: penalizes rare rules by the value of *exp* (1 – *Rule Frequency*).
- Phrase penalty: penalizes each rule in the translation by the fixed value of 1 to encourage using rules with a longer right hand side.

The pattern penalty feature (10) is added to the filtered models together with other features.

The phrase-boundary model is trained with target side POS tags as word classes. We examine a variant of SAMT [3] which labels rules with the following annotations:

- X: target side phrase does not correspond to a span in the parse tree.
- N_1: target side phrase corresponds to a syntactic category N_1.
- $N_2\backslash N_1$ or N_1/N_2: a partial syntactic category N_1 missing a N_2 to the left or right.
- N_1+N_2: target side phrase spans two adjacent syntactic categories.

Thrax [23] – the grammar extraction tool of Joshua – has an option that allows double-plus nonterminals (the annotation N1+N2+N3). We refer to this variant as SAMT/double in the experiments.

B. Results of non-filtered models

Table 2 presents the results of non-filtered models including the number of rules in millions, the resulting case-insensitive BLEU score and the average decoding-time-per-sentence in seconds.

As we expected, the models extracted for the language pairs with more monotonic word order have a larger size. Persian-English translation with different word order has the smallest model size and French-English translation with similar word order has the largest one. In addition, Persian has a rich morphology which leads to a weak word alignment and consequently to less number of rules.

A translation model with more diversity of nonterminals results in a larger model size. In the examined models, hierarchical phrase-based model with one generic nonterminal has the smallest size. SAMT and phrase-boundary models with the translation quality better than hierarchical phrase-based model have long decoding time without filtering.

According to the results, SAMT/double with a longer decoding time has a better translation quality than SAMT in translation from Persian and German to English. More difference in the word order of these language pairs needs more partial matching of phrases with the target side syntax trees provided by SAMT/double.

C. Results of filtered models

Table 3 presents the results of the models filtered by the monotonic filter. Comparing the results in Table 2 with Table 3, the monotonic filter remarkably reduces the model size and decoding time with no significant impact on the translation quality of the filtered models but SAMT/double.

978-1-5090-3586-1/16/$31.00 ©2016 IEEE

Table2. Translation results for the non-filtered hierarchical phrase-based (HPB), SAMT and phrase-boundary models (Rules in millions and Time in seconds)

	Persian-English			German-English			French-English		
Model	Rules	BLEU	Time	Rules	BLEU	Time	Rules	BLEU	Time
HPB	11	11.75	0.29	101	18.53	0.13	176	27.07	0.02
SAMT	17	11.91	14.0	114	18.75	10.7	232	28.21	24.6
SAMT/double	21	12.41	16.1	186	19.16	13.7	302	27.68	26.0
Phrase-boundary	29	12.27	21.2	411	18.91	7.9	916	28.66	13.8

Table3. Translation results for the filtered models

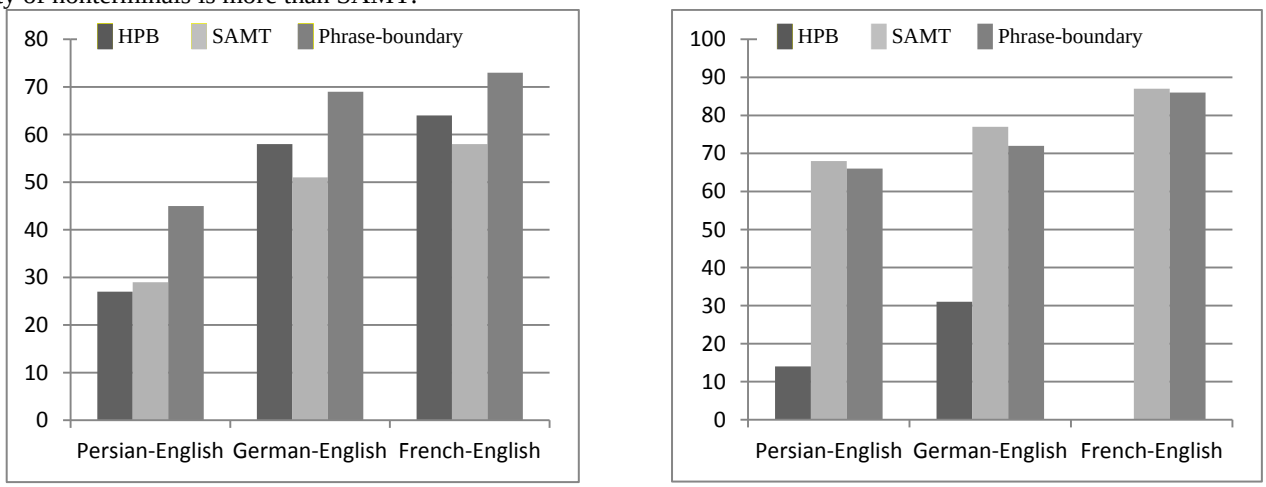
		Persian-English			German-English			French-English		
Model	Filter	Rules	BLEU	Time	Rules	BLEU	Time	Rules	BLEU	Time
HPB	*monotonic*	8	11.49	0.25	42	18.23	0.09	63	26.67	0.14
SAMT	*monotonic*	12	11.73	4.5	56	18.50	2.5	98	27.93	3.1
SAMT/double	*monotonic*	14	11.50	4.2	79	18.41	2.5	-	-	-
SAMT/double	*MRC2*	2	09.03	4.7	18	18.96	4.7	-	-	-
Phrase-boundary	*monotonic*	16	11.95	7.3	128	18.86	2.2	249	28.63	2.0

Although, SAMT/double has better quality than SAMT in translations from Persian and German to English, its quality is affected by the monotonic filter. Thus, SAMT/double is also filtered by a known post-filtering of extracted grammar [4] which discards rare hierarchical rules occurred fewer times than a given threshold. As shown in [4], increasing this threshold decreases the translation quality. We enable this filter in the grammar extraction tool – Thrax – by setting the parameter "Min-Rule-Count" to the threshold value of 2 (*MRC2* in Table 3). *MRC2* filter has a good performance in German-English translation but it causes a drop in the quality of Persian-English translation. Generally, filtering of rules has impact on SAMT/double more than SAMT because its diversity of nonterminals is more than SAMT.

Diagrams (a) and (b) in Fig. 2 present the reduction percentage of the model size and decoding time respectively by the monotonic filter for the examined translation models. The reduction percentage is higher for language pairs with more monotonic word order. French-English translation has the most reduction in our experiments. Moreover, SAMT and phrase-boundary models have the most reduction in decoding time and model size respectively.

D. Distribution of rules

We analyzed the distribution of rules in the monotonic filtered models based on their right hand side. Table 4 presents the percentage of rules in two categories; the rules consistent with



(a) Model size reduction percentage (b) Decoding time reduction percentage

Fig. 2. Reduction percentages for monotonic filtered models

Table4. Distribution of rules by percentage based on their Right Hand Side (RHS) for monotonic filtered models

		Persian-English	German-English	French-English
Model	RHS	Rules %	Rules %	Rules %
HPB	$Boundary_2$	88	93	97
	Others	12	7	3
SAMT	$Boundary_2$	84	93	98
	Others	16	7	2
Phrase-boundary	$Boundary_2$	94	94	98
	Others	6	6	2

978-1-5090-3586-1/16/$31.00 ©2016 IEEE

Boundary$_2$ filtering pattern (2) and other rules.

According to the distributions, the percentage of rules which are not consistent with *Boundary*$_2$ filtering pattern (*Others*) is small in different models. Thus, the majority of aligned phrase pairs are monotone 2-decomposable. Note that only the rules consistent with *Boundary*$_2$ pattern are accepted for monotone 2-decomposable phrase pairs. As the rules not corresponding to the monotone 2-decomposable phrase pairs are not filtered by the monotonic filter, the pattern penalty feature (10) decreases the effect of noisy data in these rules.

VI. Conclusion

This paper proposed the use of monotonic filter for hierarchical models of SMT as a general filtering method based on the alignment pattern of phrase pairs. This filter reduces the number of hierarchical rules extracted from phrase pairs which are decomposable to monotonic aligned subphrases.

We examined the monotonic filter for the language pairs with different amounts of monotonic word order and with various models (the hierarchical phrase-based, SAMT and phrase-boundary models). According to the results, the more monotonic word order the language pairs have, the more reduction in model size and decoding time by the monotonic filter occurs. The results showed that this filter effectively reduces the size and decoding time of translation models. Although, the translation quality of a variant of SAMT is affected by filtering, monotonic filter has no tangible impact on other translation models.

The rule extraction methods in SAMT and phrase-boundary models are similar to the hierarchical phrase-based model (e.g. using at most two nonterminals on the right hand side of the rules or using glue rules for monotone concatenation of the output phrases). In fact, our experiments showed that usability of the monotonic filter is not dependent on the rule labeling method. This filter is applicable for the hierarchical models having rule extraction method similar to the hierarchical phrase-based model.

References

[1] S. Salami, M. Shamsfard, and S. Khadivi, "Phrase-boundary model for statistical machine translation," *Comput. Speech Lang.*, vol. 38, pp. 13–27, 2016.

[2] D. Chiang, "Hierarchical phrase-based translation," *Comput. Linguist.*, vol. 33, no. 2, pp. 201–228, 2007.

[3] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," in *Proceedings of the Workshop on Statistical Machine Translation*, 2006, pp. 138–141.

[4] A. Zollmann, A. Venugopal, F. Och, and J. Ponte, "A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 2008, pp. 1145–1152.

[5] Z. He, Y. Meng, and H. Yu, "Discarding monotone composed rule for hierarchical phrase-based statistical machine translation," in *Proceedings of the 3rd International Universal Communication Symposium*, 2009, pp. 25–29.

[6] G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne, "Rule filtering by pattern for efficient hierarchical translation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 380–388.

[7] S.-W. Lee, D. Zhang, M. Li, M. Zhou, and H.-C. Rim, "Translation model size reduction for hierarchical phrase-based statistical machine translation," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, pp. 291–295.

[8] B. Sankaran, G. Haffari, and A. Sarkar, "Bayesian extraction of minimal scfg rules for hierarchical phrase-based translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 2011, pp. 533–541.

[9] B. Sankaran, G. Haffari, and A. Sarkar, "Compact rule extraction for hierarchical phrase-based translation," in *The 10th biennial conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA. Association for Computational Linguistics*, 2012.

[10] G. M. de Buy Wenniger and K. Sima'an, "Labeling hierarchical phrase-based models without linguistic resources," *Mach. Transl.*, pp. 1–41, 2016.

[11] J. H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving translation quality by discarding most of the phrasetable," 2007.

[12] W. Ling, J. Graça, I. Trancoso, and A. Black, "Entropy-based pruning for phrase-based machine translation," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 962–971.

[13] S. C. of ICT, "Mizan English-Persian Parallel Corpus," 2013.

[14] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, 2005, vol. 5, pp. 79–86.

[15] Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, S. Khudanpur, L. Schwartz, W. N. G. Thornton, J. Weese, and O. F. Zaidan, "Joshua: An open source toolkit for parsing-based machine translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 135–139.

[16] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000, pp. 440–447.

[17] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, 2003.

[18] A. Pauls and D. Klein, "Faster and smaller n-gram language models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011, pp. 258–267.

[19] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 160–167.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002, pp. 311–318.

[21] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 2003, pp. 423–430.

[22] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[23] M. Post, J. Ganitkevitch, L. Orland, J. Weese, Y. Cao, and C. Callison-Burch, "Joshua 5.0: Sparser, better, faster, server," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 2013, pp. 206–212.

978-1-5090-3586-1/16/$31.00 ©2016 IEEE