Integrating Shallow Syntactic Labels in the Phrase-Boundary Translation Model

SHAHRAM SALAMI and MEHRNOUSH SHAMSFARD, Shahid Beheshti University

Using a novel rule labeling method, this article proposes a hierarchical model for statistical machine translation. The proposed model labels translation rules by matching the boundaries of target side phrases with the shallow syntactic labels including POS tags and chunk labels on the target side of the training corpus. The boundary labels are concatenated if there is no label for the whole target span. Labeling with the classes of boundary words on the target side phrases has been previously proposed as a phrase-boundary model which can be considered as the base form of our model. In the extended model, the labeler uses a POS tag if there is no chunk label in one boundary. Using chunks as phrase labels, the proposed model generalizes the rules to decrease the model sparseness. The sparseness is a more important issue in the language pairs with a lot of differences in the word order because they have less number of aligned phrase pairs for extraction of rules. The extended phrase-boundary model is also applicable for low-resource languages having no syntactic parser. Some experiments are performed with the proposed model, the base phrase-boundary model, and variants of Syntax Augmented Machine Translation (SAMT) in translation from Persian and German to English as source and target languages with different word orders. According to the results, the proposed model improves the translation performance in the quality and decoding time aspects. Using BLEU as our metric, the proposed model has achieved a statistically significant improvement of about 0.5 point over the base phrase-boundary model.

CCS Concepts: • **Computing methodologies → Machine translation**;

Additional Key Words and Phrases: Statistical machine translation, Hierarchical models, POS tag, Chunk label

ACM Reference format:
Shahram Salami and Mehrnoush Shamsfard. 2018. Integrating Shallow Syntactic Labels in the Phrase-Boundary Translation Model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 17, 3, Article 17 (February 2018), 12 pages.
https://doi.org/10.1145/3178460

1 INTRODUCTION

Statistical Machine Translation (SMT) is pervasively used for language pairs for which parallel corpora are available. Using similar sizes of parallel corpora, some language pairs have lower translation qualities than others. One of the main causes for this deficiency is the difference in the word order of source and target languages. Different word order not only makes a challenge for word reordering but also causes less aligned phrase pairs and more sparsity of the model.

Authors' addresses: S. Salami and M. Shamsfard, Faculty of Computer Science and Engineering, Shahid Beheshti University, Daneshjou Boulevard, Shahriari SQ, Chamran HWY, 1983969411 Tehran, Iran; emails: sh_salami@sbu.ac.ir, m-shams@sbu.ac.ir.

https://doi.org/10.1145/3178460

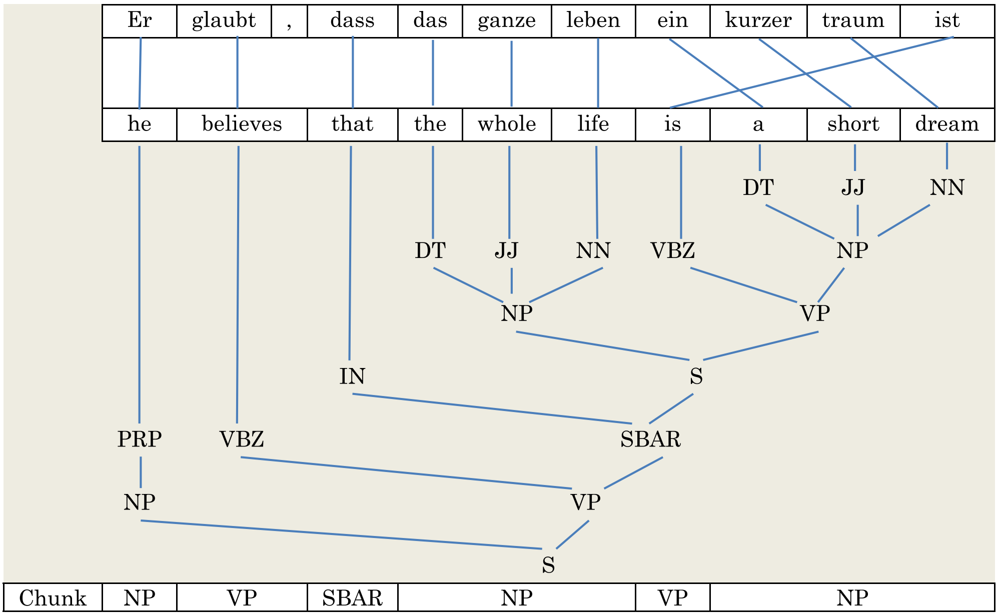
Er glaubt , dass das ganze leben ein kurzer traum ist
he believes that the whole life is a short dream
DT JJ NN
DT JJ NN VBZ NP
NP VP
IN S
PRP VBZ SBAR
NP VP
S
Chunk NP VP SBAR NP VP NP

Fig. 1. Alignment between German and English sentences along with English syntax tree and chunk labels.

In comparison to phrase-based models, hierarchical models support longer phrase pairs and explicitly explain word reordering using hierarchical phrases. As a basic hierarchical translation system with no use of linguistic information, Chiang et al. (2005) proposed the Hiero that extracts grammar rules from aligned phrases using one generic label. To increase the precision of labels using syntactic information, a known solution was proposed as syntax augmented machine translation (SAMT). SAMT labels the rules by matching target side of aligned phrases with the nodes of syntax trees on the target side of the training corpus. Phrases with no corresponding sub-tree on the target side syntax trees are named with one default label. As there are many aligned phrase pairs with no corresponding node in the syntax tree, some partial matches have been proposed in different variants of SAMT. For example, in a variant of SAMT (Zollmann and Venugopal 2006), the following annotation is used:

- 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, respectively.
- $N_1 + N_2$: target side phrase spans two adjacent syntactic categories.

Using the above annotation, the following rules are extracted from the German-English sentence pair in Figure 1 (Notation "," separates source and target part of the rules and notation "~" indicates a one-to-one correspondence between the source and target part nonterminals):

$$\text{X} \rightarrow \text{<dass das ganze, that the whole>}, \tag{1}$$

$$\text{SBAR/VP} \rightarrow \text{<dass das ganze leben, that the whole life>}, \tag{2}$$

$$\text{NN} \rightarrow \text{<leben, life>}, \tag{3}$$

$$\text{NP} \rightarrow \text{<das ganze leben, the whole life>}, \tag{4}$$

$$\text{S} \rightarrow \text{<}\text{NP}^{\sim 1}\text{NP}^{\sim 2}\ \text{ist, NP}^{\sim 1}\text{is NP}^{\sim 2}\text{>}\,. \tag{5}$$

The corresponding aligned phrase pair of Rules 3–5 are matched with the syntactic categories on the target side syntax tree. Rule 2 presents a “SBAR” missing a “VP” on the right as a partial syntactic category. With no corresponding syntactic or partial syntactic category, Rule 1 is labeled with the default label “X.” Although partial syntactic categories (such as Rule 2) increase the coverage of SAMT labeling, Moses SAMT4 (Koehn et al. 2007) with more partial syntactic categories covered at most 50% of aligned phrase pairs in Chinese-English translation (Almaghout et al. 2010). On the other hand, using more labels for partial syntactic categories increases the sparseness of the model.

SAMT describes the whole content of target side phrases but using a simpler annotation; phrase-boundary model (Salami et al. 2016) only describes their boundaries. This hierarchical model labels rules with the classes of boundary words on the target side phrases. The word classes are uniformly concatenated with one hyphen for phrases longer than one word. Word clustering or POS tags on the target side of the training corpus are used to define word classes. For example, using POS tags (on the leaves of the syntax tree in Figure 1) as word classes, the following rules are extracted similar to Rules 1–5:

$$\text{IN-JJ} \rightarrow \text{<dass das ganze, that the whole>}, \tag{6}$$

$$\text{IN-NN} \rightarrow \text{<dass das ganze leben, that the whole life>}, \tag{7}$$

$$\text{NN} \rightarrow \text{<leben, life>}, \tag{8}$$

$$\text{DT-NN} \rightarrow \text{<das ganze leben, the whole life>}, \tag{9}$$

$$\text{DT-NN} \rightarrow \text{<DT-NN}^{\sim 1}\text{DT-NN}^{\sim 2}\text{ist, DT-NN}^{\sim 1}\text{ is DT-NN}^{\sim 2}\text{>}\,. \tag{10}$$

In comparison to SAMT, labeling of phrase-boundary model covers all phrases with no use of any default label, but only uses word classes as the leaves of syntax trees. Using POS tags and chunk labels as shallow syntactic labels, this article generalizes the labeling of phrase-boundary model to utilize phrase labels along with word tags. Chunk labels are results of shallow syntactic parsing (or chunking) that identify the sentence constituents (such as noun and verb phrases in Figure 1) uniquely, but their internal structure or their roles are not specified in the sentence.

Our extended phrase-boundary model labels the rules by matching the boundaries of target side phrases with the shallow syntactic labels on the target side of the training corpus. In other words, we extend the phrase-boundary model for labeling with the target side chunks. Hierarchical translation models (such as the proposed one) promise a better word reordering than phrase-based models. On the other hand, the language pairs with a lot of differences in the word order are subject to the sparseness of the model because they have a lesser number of aligned phrase pairs for extraction of rules. Utilizing chunks as phrase labels, our extended model generalizes the rules of the base phrase-boundary model and can decrease the sparseness.

Using the BLEU as our metric, we compare the translation results of the extended phrase-boundary model with its base form, Hiero, and variants of SAMT. The experiments are performed in Persian-English and German-English translations as language pairs with different word orders. Hierarchical translation models other than Hiero have a long decoding time. To tackle this challenge, we filtered out the grammar rules of translation models. According to the results, the extended phrase-boundary model achieved a translation performance better than the selected baselines in the quality or decoding time aspects.

The design and implementation method of the proposed model is presented in this article. Some related work is referenced in Section 2. Section 3 explains the proposed model. Section 4 shows the results of experiments. Finally, the article is concluded in Section 5.

2 RELATED WORK

The hierarchical phrase-based model (Chiang 2005) is introduced for unsupervised grammar induction based on phrase alignment with one generic label. To improve the translation quality, one approach changes labeling of the rules and another approach finds better derivations in this model by the methods such as scoring the derivations.

In the second approach, to generate a target sentence in left-to-right order, Watanabe et al. (2006) use hierarchical rules, the target sides of which were in the Greibach Normal Form class. Note that discontinuous generation of target words limits pruning of the decoding space with the language model. Zhou et al. (2008) scored the derivations during translation decoding using syntactic information. For better rule selection in the decoding process of the hierarchical phrase-based model, the context of input was used in forms such as POS tags (He et al. 2008) and CCG (Combinatory Categorial Grammar) tags (Haque et al. 2010). To indicate matching with syntax trees, some features were added as soft syntactic constraints to the hierarchical phrase-based model (Marton and Resnik 2008; Venugopal et al. 2009; Huang et al. 2010). Chiang (2010) proposed a model that uses some features for both syntax trees of the source and target side languages. Using the classes of phrase-boundary words in the decoding process, the word reordering is improved in the hierarchical phrase-based model (Huck et al. 2012) and the phrase-based model (Cherry 2013) but we use them to label the rules.

In the approach that changes labeling of the rules, using another nonterminal on the left-hand side of hierarchical rules, Huck et al. (2012) avoided the recursion of hierarchical rules to limit decoding space of the hierarchical phrase-based model. A decomposition pattern of phrase pairs was used to label hierarchical rules without linguistic resources (Wenniger and Sima'an 2015). To label the rules, Li et al. (2012) used a POS tag of head words. Using the labels on the target side of the training corpus, the hierarchical phrase-based model was augmented with the syntactic categories as SAMT (Zollmann and Venugopal 2006) and CCG tags (Almaghout et al. 2010). Mino et al. (2014) clustered syntax labels of SAMT to reduce the number of rules. Zollmann and Vogel (2011) labeled rules with the word classes and showed that the translation quality of their model is similar to a variant of SAMT. To name nonterminals, they concatenate classes of boundary words with the glue notations and use a special notation for target side phrases of length 2. Labeling with the classes of phrase-boundary words, Salami et al. (2016) proposed the phrase-boundary model with filtering methods in the grammar extraction step which decreases both model size and decoding time. They uniformly concatenate classes of boundary words with one hyphen and showed that their model has a better performance than the model of Zollmann and Vogel (2011). The hierarchical model proposed in this article generalizes the boundary labeling to the shallow syntactic labels including POS tags and chunk labels to achieve better translation quality. Like some of other mentioned models in this section, our model uses syntactic information but we use shallow syntactic information which is attainable for most languages including low-resource ones.

There are some filtering methods examined in this article. In general, two filtering approaches were proposed for hierarchical translation models. One filtering approach eliminates unnecessary rules from the extracted grammar but another approach prevents extraction of too many rules by changing the method of rule extraction. Generally, the latter one is preferred because it also reduces the training resources.

In the filtering approach that eliminates unnecessary rules, a known method (Zollmann et al. 2008) discards rare hierarchical rules occurring fewer times than a given threshold. As there is a tradeoff between threshold and translation quality, the best threshold value depends on the training corpus. A hierarchical phrase-based model was filtered by discarding hierarchical rules in which their source sides appear only in the monotone composed rules (He et al. 2009). Iglesias et al.

(2009) categorized the rules into different patterns and filtered out those patterns which could be discarded without a significant impact on the translation quality. The rules were discarded based on information redundancy encoded in the translation rules (Lee et al. 2012).

In the filtering approach that changes the method of rule extraction, the alignment probability of words was examined to restrict the extraction of rules to aligned phrase pairs with the higher probability (Sankaran et al. 2011). A minimum set of translation rules was extracted on which at least one derivation could be constructed for each phrase pair (Sankaran et al. 2012). Based on the alignment pattern of phrase pairs, the monotonic filter was proposed for hierarchical translation models (Salami and Shamsfard 2016). This filter cuts down the patterns of hierarchical rules extracted from phrase pairs which are decomposable to monotonic aligned subphrases.

3 THE MODEL

The proposed model defines weighted rules in a Probabilistic Synchronous Context-Free Grammar extracted from a set of aligned phrase pairs $<f_i^j, e_m^n>$ where f_i^j and e_m^n stand for inclusive source and target substrings from position i to j and position m to n, respectively. Following Chiang (2005), the grammar rules are defined as lexical, hierarchical, and glue rules. Lexical rules represent aligned phrase pairs with no nonterminal on the right-hand side. Hierarchical rules are defined by at most two substitutions of subphrases with nonterminals. Glue rules are defined for all grammar nonterminals for serial concatenation of output phrases.

A phrase-boundary model in the base form uniformly concatenates the classes of boundary words on the target side phrases with one hyphen to label nonterminals (Salami et al. 2016). Using POS tags as the words classes, the aligned phrase pair $<f_i^j, e_m^n>$ is labeled as

$$X_{m,n} = \begin{cases} POS\,(e_m) & if\ m = n \\ POS(e_m)\text{ - }POS(e_n) & else \end{cases}. \tag{11}$$

Rules 6–10 are samples of rules in the base phrase-boundary model. In addition to POS tags, the extended phrase-boundary model uses chunk labels on the boundaries of the target side phrases to label the extracted rules. Considering priority of chunk labels, the proposed model names nonterminals by the matched label for the whole target span or by concatenation of the boundary labels on the target side phrases. If a chunk spans the entire phrase pair, it is used as a label. Otherwise, the chunks inside the target side phrase that start from the left side or end at the right side of target side phrases are used to make labels. As a final fallback, POS tags are used if no chunk exists in one phrase boundary or for an entire phrase of length 1. The naming convention is formally shown in Definition 1.

Definition 1. The corresponding nonterminal of the aligned phrase pair $<f_i^j, e_m^n>$ in which the target side starts with e_m and ends with e_n is labeled as

$$X_{m,n} = \begin{cases} Chunk(e_m^n) & if\, e_m^n\ matches\ with\ a\ chunk \\ POS(e_m) & else\ if\ m\ = n \\ Left\text{-}Label\text{ - }Right\text{-}Label & else \end{cases}$$

$$Left\text{-}Label = \begin{cases} Chunk\ (e_m^a)\colon m \leq\ a < n & if\, e_m^a\ matches\ with\ a\ chunk \\ POS\ (e_m) & else \end{cases}$$

$$Right\text{-}Label = \begin{cases} Chunk\ (e_b^n) : m < b \leq n & if\, e_b^n\ matches\ with\ a\ chunk \\ POS\ (e_n) & else \end{cases}$$

Same as the base phrase-boundary model, the length of 12 words is the default limit for aligned phrase pairs. The labeling method in Definition 1 is straightforward because POS tags exist in the boundaries of all phrase pairs. One hyphen is used as glue notation to concatenate left and right labels. The function *Chunk* (e_x^y) returns the chunk label which starts in the position x and ends in the position y. With the priority of existing chunk labels, left and right boundary labels on the target side phrase are concatenated by one hyphen where there is no label for the whole target span. Definition 1 assumes that *Left-Label* and *Right-Label* are functions (with a single output value) and "a" is smaller than "b" because there is no overlap between chunk labels.

For example, Rules 12–16 extracted from the sentence pair in Figure 1 are samples of the rules in the extended phrase-boundary model. These rules represent a generalization of Rules 6–10 using chunk labels (SBAR and NP).

$$\text{SBAR-JJ} \rightarrow \text{<dass das ganze, that the whole>}, \tag{12}$$

$$\text{SBAR-NP} \rightarrow \text{<dass das ganze leben, that the whole life>}, \tag{13}$$

$$\text{NN} \rightarrow \text{<leben, life>}, \tag{14}$$

$$\text{NP} \rightarrow \text{<das ganze leben, the whole life>}, \tag{15}$$

$$\text{NP-NP} \rightarrow \text{<}\text{NP}^{\sim 1}\text{NP}^{\sim 2}\text{ ist, NP}^{\sim 1}\text{ is NP}^{\sim 2}\text{>}\,. \tag{16}$$

According to Definition 1, POS tags are used if there is no chunk label in the phrase boundary or for whole target span of length 1. Rules 14 and 15 are labeled with *POS* (e_m) and *Chunk* (e_m^n), respectively, but "*Left-Label - Right-Label*" is used for labeling the left-hand side of other rules. Rule 12 is only matched with a chunk label on the left side of the target side phrase, but Rules 13 and 16 are matched with the chunk labels on the both left and right sides of the target side phrase. Note that Rule 1 (which is similar to Rule 12) is labeled with the default label in SAMT.

Like SAMT and the base phrase-boundary model, the extended phrase-boundary model uses labels on the target side of the training corpus. In this manner, while decoding is directed by the input, the translation output can be structured based on the target side syntactic labels. Like the base phrase-boundary model, the proposed model only uses the boundary labels on the target side phrases. Naming nonterminals with all labels included in the target side phrase increases the model sparseness. In comparison to the base phrase-boundary model, the proposed model uses chunk labels in addition to POS tags. Utilizing chunk labels generalizes the rules using phrase labels instead of word classes. In comparison to SAMT, the base phrase-boundary model and the proposed one use less linguistic information and label all phrases uniformly.

4 EXPERIMENTS

Series of experiments are performed in Persian-English and German-English translations. Persian and English have different word orders. 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. Persian-English translation is trained on Mizan corpus (Kashefi 2018) with 1M sentences. The 1k+1K sentences of this corpus are selected for the development and test sets. German-English translation is trained on the Europarl-V7 corpus (Koehn 2005) with 2M sentences. The last 500 sentences of the WMT 2012 translation task are used for the development set, and the first 1K sentences of the WMT 2013 translation task are used for the test set. Statistics of the used corpora are presented in Table 1.

We compare our results with three baselines: Hiero, the base phrase-boundary model that labels rules using POS tags on the target side phrases and variants of SAMT (Zollmann and Venugopal 2006) that use syntax trees on the target side of the training corpus for labeling the rules. The models are trained and evaluated with Joshua toolkit (Li et al. 2009). The words are aligned in both directions of translation by GIZA++ (Och and Ney 2000) and the results are symmetrized (Och

Table 1. Statistics of the Training Corpora

Training corpora	No. of words	No. ofword types
Persian-English	15M+13M	135K+98K
German-English	39M+41M	333k+104k

and Ney 2003). The language models (LMs) are trained on the target side of the training corpus by the Berkeley LM tool (Pauls and Klein 2011) with Kneser-Ney smoothing. We use 3-gram LM for Persian-English translation and 4-gram LM for German-English translation (with larger training corpus). The scaling factors of models are trained by Minimum Error Rate Training (Och 2003). The BLEU-4 (Papineni et al. 2002) is used as the evaluation metric in our experiments.

A new version of Thrax 2.0 (Post et al. 2013)—the grammar extraction tool of Joshua toolkit—is developed to support the proposed model. Extraction of the proposed grammar needs word POS tags and chunk labels. These labels are defined by the SENNA tool (Collobert et al. 2011) on the English side of the parallel training corpus. The target side syntax trees for SAMT are generated by the Stanford parser (Klein and Manning 2003).

4.1 Configurations

The baseline models are configured using default settings. These settings limit source and target parts of lexical rules to the length of 10 words and hierarchical rules to the length of 5, including a maximum of two nonterminals. 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 relative frequencies of phrase pairs: computed in both source-to-target and target-to-source directions.
- Negative log of lexical weights of phrase pairs (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 translation by the fixed value of 1 to encourage using rules with a longer right-hand side.

The length of rules and the features selected for the proposed model are the same as those of the baseline models.

4.2 Results of Non-filtered Models

The grammar extraction tool, Thrax, supports a variant of SAMT as defined by Zollmann and Venugopal (2006). In addition, this tool has an option to extend partial syntactic categories with double-plus nonterminals (the annotation N_1+N_2+N_3). In the experiments, we examine the following variants of SAMT and the phrase-boundary model:

- SAMT/single: SAMT with X, N_1, $N_2 \backslash N_1$, N_1/N_2, and N_1+N_2 labels (X is the default label for non-syntactic phrases and N_i is a syntactic category).
- SAMT/double: SAMT/single plus the annotation N_1+N_2+N_3.
- Boundary/base: the base phrase-boundary model labeled with POS tags (as of Equation (11)).
- Boundary/CHK: the extended phrase-boundary model labeled with POS tags and chunk labels.

Table 2. Translation Results for Non-Filtered Models (Rules in Millions and Average Decoding-Time-Per-Sentence in Seconds)

	Persian-English				German-English			
Model	Rules	Label types	BLEU	Time	Rules	Label types	BLEU	Time
Hiero	11	1	11.75	0.2	201	1	19.65	0.6
SAMT/single	17	3,866	11.91	14.0	229	4,453	19.87	20.0
SAMT/double	21	24,441	12.41	16.1	376	46,017	19.89	28.0
Boundary/base	29	1,555	12.27	21.2	821	1,755	19.20	14.9
Boundary/CHK	28	1,918	12.63	18.2	808	2,252	20.62	16.0

Table 2 presents the performance of the Hiero and above variants of SAMT and phrase-boundary in model size (number of rules), number of label types, translation quality, and decoding time. The results in the following tables present the number of rules in millions, the resulting case-insensitive BLEU score, and the average decoding-time-per-sentence in seconds.

Considering the number of label types in Table 2, using more partial syntactic categories, SAMT/double has a label set much larger than SAMT/single. In addition, the label set of SAMT/single is larger than the variants of phrase-boundary. Note that SAMT describes whole content of a phrase with a node name in the syntax tree, a range of partial syntactic categories, or the default label. To decrease the search space and the model size of SAMT, the grammar extraction tool only allows the default label on the left-hand side of lexical rules. Regarding this fact, variants of SAMT with a larger set of label types has a smaller model size than variants of phrase-boundary. On the other hand, using a simpler annotation, phrase-boundary describes a phrase with a chunk label, a POS tag, or describes its boundaries with a uniform N_i-N_j notation. Although utilizing chunk labels increases the number of label types, boundary/CHK has a model size slightly smaller than boundary/base due to its better label convergence.

Generally, utilizing more label types increases the model sparseness. The sparseness issue is more important for the language pairs with a lot of differences in the word order (such as examined ones) due to the lesser number of aligned phrase pairs. In fact, language pairs with more monotonic word order have a larger model size because monotonic word order increases the number of aligned phrase pairs and extracted rules. The model size of Persian-English translation is far smaller than that of German-English translation. Although we use a smaller training corpus for Persian-English translation, the smaller size is mostly related to the difference between the word order of Persian and English languages.

According to the results, SAMT/double with a longer decoding time has a better translation quality than SAMT/single for translation from Persian to English. The proposed phrase-boundary model (Boundary/CHK) has a better translation quality in German-English translation. This model also achieved a slight improvement over other models in Persian-English translation. In comparison to SAMT, the proposed model with a smaller label set has a better translation quality for the examined language pairs with non-monotonic word orders. Hiero with a lower translation quality has a much shorter decoding time than other models. Considering this fact, in the next section, we filtered out the grammar rules of SAMT and phrase-boundary to reduce their decoding time.

Variants of phrase-boundary and SAMT use information on the target side of the training corpus to label the rules. Although English as the target language of our experiments has a rich set of linguistic tools, phrase-boundary is also applicable for low-resource languages without a well-known syntactic parser. So, using the Hazm[1] toolkit, we define POS tags and chunks of Persian

[1] http://www.sobhe.ir/hazm/.

Table 3. Translation Results for English to Persian Translation

Model	Rules	Label types	BLEU	Time
Hiero	11	1	10.98	0.2
Boundary/base	28	512	11.86	13.9
Boundary/CHK	28	770	12.20	22.2

language to perform additional experiments on English to Persian translation. Table 3 represents the results of these experiments with Hiero and variants of the phrase-boundary model.

According to the experiments, the proposed model significantly improves the translation quality of English to Persian translation over the Hiero as a model with a generic label. Due to a better label convergence, Boundary/CHK with a larger set of labels has a model size similar to Boundary/base. The convergence of labels is important to avoid more sparseness in the low-resource language pairs (such as Persian-English) which have a limited set of training corpora.

4.3 Results of Filtered Models

Hierarchical models with diversity of nonterminal labels have a long decoding time. Using two known filtering methods, we filter the translation models of SAMT and phrase-boundary to reduce their decoding space. Although reducing the decoding space may affect the translation quality in sparse models, it increases the translation quality in some cases (as represented in the following results).

We use the monotonic filter (Salami and Shamsfard 2016) proposed for filtering the rules in the grammar extraction step based on the alignment pattern of phrase pairs. This filter remarkably reduces both training and decoding resources of the examined models. The monotonic filter accepts the rules which are candidates for extraction with the following conditions:

(1) Candidate rules consistent with one of the following patterns on the source side (where w_i is one string of words and x_i is one nonterminal):

$$\text{Patterns} = \{w_1, x_1w_1, w_1x_1, x_1w_1x_2\} \tag{17}$$

(2) Other rules candidate for extraction from those phrase pairs that are not *monotone 2-decomposable* (as defined in the following paragraph).

One phrase pair is *monotone 2-decomposable* if it can be broken into monotonic aligned subphrases. For example, Equation (18) presents a monotone 2-decomposable phrase pair in Figure 1.

$$\text{<ein kurzer traum, a short dream>} = \text{<ein, a>} + \text{<kurzer traum, short dream>}. \tag{18}$$

To preserve the translation quality, a penalty feature is added to the monotonic filtered grammar. This feature named as *pattern penalty* is defined as follows:

— *Pattern penalty* feature has the value of 0 if the source part of the rule is consistent with one of the filtering patterns (Set 17); otherwise, it has the value of 1 to penalize other rules.

The pattern penalty feature is added to the grammar rules together with other features (Section 4.1) for the monotonic filtered models.

We also use a known post-filtering of extracted grammar introduced by Zollmann et al. (2008). They discarded rare hierarchical rules occurring fewer times than a given threshold and showed that 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

Table 4. Translation Results for Filtered Models

		Persian-English			German-English		
Model	Filter	Rules	BLEU	Time	Rules	BLEU	Time
SAMT/double	*Monotonic*	14	11.50	4.2	158	19.94	6.8
	MRC_2	2	09.03	4.7	38	20.12	11.7
	MRC_3	-	-	-	19	19.76	7.1
	Monotonic+MRC_2	-	-	-	18	19.03	1.8
Boundary/base	*Monotonic*	16	11.95	7.3	256	19.11	4.8
	MRC_2	4	9.07	9.7	102	19.50	6.2
	Monotonic+MRC_2	-	-	-	46	20.01	1.3
Boundary/CHK	*Monotonic*	16	12.50	8.0	248	20.30	5.9
	Monotonic+MRC_2	-	-	-	45	20.46	1.8

Table 5. The Best BLEU Scores of Filtered Phrase-Boundary Models Also Computed with 95% Confidence Interval (BLEU*)

	Persian-English			German-English		
	Monotonic			*Monotonic + MRC_2*		
Model	BLEU	BLEU*	p-value	BLEU	BLEU*	p-value
Boundary/base	11.95	12.11	-	20.01	19.84	-
Boundary/CHK	12.50	12.69	0.041	20.46	20.27	0.048

n (MRC_n in Table 3). Table 4 presents the results of applying this filter and monotonic filter on examined models.

Regarding the results of Persian-English translation, MRC_n filtering causes a drop in the translation quality because of the general rarity of rules, but the monotonic filter has a suitable performance with variants of phrase-boundary. Because of the sparseness of the model, the monotonic filter in Persian-English translation and the combined filter (*Monotonic+MRC_2*) in German-English translation affect the translation quality of SAMT/double with a large set of label types (Table 2). Using a smaller set of label types than SAMT, variants of phrase-boundary better tolerate the sparsity made by the applied filters. Considering the results of filtered models, the proposed phrase-boundary model has a translation quality better than our baselines. This model has a decoding time shorter than SAMT/double, too.

The best results of two variants of the phrase-boundary model are presented in Table 5. To study the statistical significance of the improvements, the BLEU scores in this table are also computed with 95% confidence interval (BLEU*) using a bootstrap resampling method (Koehn 2004).

The BLEU* and p-value in this table are computed between the proposed model and the base phrase-boundary model by paired bootstrap resampling script in the Moses decoder (Koehn et al. 2007). According to the results, improvement of the extended phrase-boundary model over the base phrase-boundary model is statistically significant with "$p\text{-value} < 0.05$."

5 CONCLUSION

The proposed phrase-boundary model in this article names nonterminals with shallow syntactic labels on the boundaries of target side phrases in the training corpus. This model in the base form just uses POS tags as classes of boundary words but our extended model in addition uses chunk labels. The proposed model is more applicable than SAMT because the shallow syntactic parser (for the chunk labeling) is more attainable than the syntactic parser for most languages, including low-resource ones.

Considering BLEU score and decoding time, we compared the results of the extended phrase-boundary model with the Hiero, the base form of phrase-boundary model, and variants of SAMT for translation from Persian and German to English as language pairs with different word orders. The extended phrase-boundary model achieved a performance better than the baselines especially in the filtered models. Variants of SAMT use a wide range of syntactic and partial syntactic categories (along with the default label) to describe the content of phrases while variants of phrase-boundary uniformly describe boundaries of target side phrases. Generally, using more label types increases the model sparseness. Due to the lesser number of aligned phrase pairs for extraction of rules, sparseness is an important issue in the language pairs with non-monotonic word order (such as the examined ones). Using a label set smaller than SAMT, the proposed model achieved a better translation quality than our baselines. Considering the model sparseness, filtering of rules impacted the quality of Persian-English translation with SAMT.

This article proposed a framework to utilize shallow linguistic labels for the rule labeling. In the future, we extend the phrase-boundary model using semantic role labels as shallow semantic labels. This extension can decrease the semantic errors in the translation.

REFERENCES

H. Almaghout, J. Jiang, and A. Way. 2010. CCG augmented hierarchical phrase based machine-translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation.* 211–218.

C. Cherry. 2013. Improved reordering for phrase-based translation using sparse features. In *Proceedings of HLT-NAACL.* 22–31.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.* 263–270.

D. Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* 1443–1452.

D. Chiang et al. 2005. The Hiero machine translation system: Extensions, evaluation, and analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.* 779–786.

R. Collobert et al. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 12 (2011), 2493–2537.

R. Haque et al. 2010. Supertags as source language context in hierarchical phrase-based SMT. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas.*

Z. He, Q. Liu, and S. Lin. 2008. Improving statistical machine translation using lexicalized rule selection. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1. 321–328.

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

Z. Huang, M. Čmejrek, and B. Zhou. 2010. Soft syntactic constraints for hierarchical phrase-based translation using latent syntactic distributions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* 138–147.

M. Huck et al. 2012. Discriminative reordering extensions for hierarchical phrase-based machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation.* 313–320.

G. Iglesias et al. 2009. Rule filtering by pattern for efficient hierarchical translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* 380–388.

O. Kashefi. 2018. MIZAN: A large persian-english parallel corpus. *CoRR* abs/1801.02107. Available at: http://arxiv.org/abs/1801.02107.

D. Klein and C. D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Volume 1.* 423–430.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit.* 79–86.

P. Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions.* 177–180.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP.* 388–395.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Vol. 1. 48–54.

S.-W. Lee et al. 2012. 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,* Vol. 2. 291–295.

J. Li et al. 2012. Head-driven hierarchical phrase-based translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers,* Vol. 2. 33–37.

Z. Li et al. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the 4th Workshop on Statistical Machine Translation.* 135–139.

Y. Marton and P. Resnik. 2008. Soft syntactic constraints for hierarchical phras-based translation. In *Proceedings of ACL.* 1003–1011.

H. Mino, T. Watanabe, and E. Sumita. 2014. Syntax-augmented machine translation using syntax-label clustering. In *Proceedings of EMNLP.* 165–171.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Vol. 1. 160–167.

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

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics.* 440–447.

K. Papineni et al. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* 311–318.

A. Pauls and D. Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 258–267.

M. Post et al. 2013. Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the 8th Workshop on Statistical Machine Translation.* 206–212.

S. Salami and M. Shamsfard. 2016. Monotonic filter for hierarchical translation models. In *Proceedings of the 2016 6th International Conference on Computer and Knowledge Engineering (ICCKE'16).* 19–24.

S. Salami, M. Shamsfard, and S. Khadivi. 2016. Phrase-boundary model for statistical machine translation. *Computer Speech & Language* 38 (2016), 13–27. Available at http://www.sciencedirect.com/science/article/pii/S0885230815001096.

B. Sankaran, G. Haffari, and A. Sarkar. 2011. Bayesian extraction of minimal SCFG rules for hierarchical phrase-based translation. In *Proceedings of the 6th Workshop on Statistical Machine Translation.* 533–541.

B. Sankaran, G. Haffari, and A. Sarkar. 2012. Compact rule extraction for hierarchical phrase-based translation. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas (AMTA'12), Association for Computational Linguistics.*

A. Venugopal et al. 2009. Preference grammars: Softening syntactic constraints to improve statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* 236–244.

T. Watanabe, H. Tsukada, and H. Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics.* 777–784.

G. Maillette de Buy Wenniger and K. Sima'an. 2015. Labeling hierarchical phrase-based models without linguistic resources. *Machine Translation* 29, 3–4 (2015), 225–265.

B. Zhou et al. 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. In *Proceedings of the 2nd Workshop on Syntax and Structure in Statistical Translation.* 19–27.

A. Zollmann et al. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of the 22nd International Conference on Computational Linguistics,* Vol. 1. 1145–1152.

A. Zollmann and A. Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation.* 138–141.

A. Zollmann and S. Vogel. 2011. A word-class approach to labeling PSCFG rules for machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 1–11.

Received January 2017; revised August 2017; accepted December 2017