Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 38 (2016) 13–27

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Phrase-boundary model for statistical machine translation

Shahram Salami [a,*], Mehrnoush Shamsfard [a], Shahram Khadivi [b]

[a] *Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran*
[b] *Human Language Technology Lab, Amirkabir University of Technology, Tehran, Iran*

Received 17 February 2015; received in revised form 6 September 2015; accepted 12 November 2015
Available online 12 December 2015

Abstract

This paper proposes a new probabilistic synchronous context-free grammar model for statistical machine translation. The model labels nonterminals with classes of boundary words on the target side of aligned phrase pairs. Labeling of the rules is performed with coarse grained and fine grained nonterminals using POS tags and word clusters trained on the target language corpus. Considering the large size of the proposed model due to the diversity of nonterminals, we have also proposed a novel approach for filtered rule extraction based on the alignment pattern of phrase pairs. Using limited patterns of rules, the extraction of hierarchical rules gets restricted from phrase pairs that are decomposable to two aligned subphrases. The proposed filtered rule extraction decreases the model size and the decoding time considerably with no significant impact on the translation quality. Using BLEU as a metric in our experiments, the proposed model achieved a notable improvement rate over the state-of-the-art hierarchical phrase-based model in the translation from Persian, French and Spanish to English language. This is applicable for all languages, even under-resourced ones having no linguistic tools.

Keywords: Statistical machine translation; Hierarchical models; Rules filtering

1. Introduction

The phrase-based model (Zens et al., 2002; Koehn et al., 2003) improved the previous state-of-the-art statistical machine translation approaches using phrases instead of words. The hierarchical phrase-based model (Chiang, 2005) started with the phrase-based model and supported the translation of longer phrases using hierarchical phrases. In that model, hierarchical phrases are induced from the parallel corpus by substituting subphrases with one generic nonterminal. The model suffers from high ambiguity and a very large number of grammar rules. Some solutions have been proposed to fix these problems. One set of solutions decreases the ambiguity of decoding using the context of input or syntactic knowledge which may reduce the model scope to languages that have linguistic tools. Another set of solutions filters grammar rules to reduce the model size without significantly affecting the translation quality.

This paper proposes the phrase-boundary model as a hierarchical model in which both grammar rules and nonterminals are induced from the parallel corpus without using syntactic trees. This model has a higher precision than the hierarchical phrase-based model with one generic nonterminal. The nonterminals of phrase-boundary grammar

[*] Corresponding author. Tel.: +98 21 29904111.
E-mail addresses: sh_salami@sbu.ac.ir (S. Salami), m-shams@sbu.ac.ir (M. Shamsfard), khadivi@aut.ac.ir (S. Khadivi).

http://dx.doi.org/10.1016/j.csl.2015.11.005
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

are defined by boundary word classes on the target side of aligned phrase pairs. In other words, each phrase pair is known by the first and last word of its target side. For example, the French-English phrase pair < idée possible, feasible idea > is known as *JJ-NN* by the concatenation of the boundary POS tags on the English side. Thus, the following rule exists in the phrase-boundary model:

$$JJ\text{-}NN \rightarrow < \text{ idé e possible, feasible idea } > \tag{1}$$

Instead of *JJ-NN* label, the left hand side of this rule is labeled with the generic nonterminal *X* in the hierarchical phrase-based model. Additionally, word classes can be determined by clustering words on the target training corpus. In this case, the words of the target corpus are clustered in an arbitrary number of word classes. Automatic clustering of the words generalizes the usability of the model to the language pairs for which no linguistic tool may be available. Although defining nonterminals based on source word classes is straightforward, the use of target side word classes is proposed. This is because the translation output can be structured based on the target syntax in the rules, while decoding is directed by the input. Target side syntax has also been used in other models such as SAMT (Zollmann and Venugopal, 2006). Although the use of both side word classes is possible, it may lead to model sparseness.

In the absence of syntactic categories, the hierarchical phrase-based model uses one generic nonterminal, while the granularity of nonterminals in the proposed phrase-boundary model depends on the number of word classes used. Nonterminal diversity increases the precision of the phrase-boundary model. On the other hand, phrases with the same syntactic category may have different nonterminals in the phrase-boundary grammar, which may result in an extra number of non-overlapped hierarchical rules in the model. Larger models consume more time and memory in the training stage and decoding process. According to the fact that hierarchical rules represent the decomposition pattern of phrase pairs, this paper proposes a novel approach to filter out grammar rules based on the alignment pattern of the phrase pairs. This filtering limits the patterns of the hierarchical rules extracted from phrase pairs that are decomposable to two aligned subphrases.

This study examined filtered and non-filtered phrase-boundary models with POS tags and automatic word clustering in language pairs with similar and different word ordering. The experiments showed that the proposed filter reduces model size by more than 70% without significantly impacting translation quality. Furthermore, experimental results demonstrated that the phrase-boundary model is a considerable improvement over the hierarchical phrase-based model using BLEU metric for evaluation. The design and implementation method of the proposed model is presented in this paper. Related work is referenced in Section 2. Section 3 explains the proposed phrase-boundary model. Section 4 introduces the methods for filtered rule extraction. Section 5 shows the results of experiments. Finally, the paper is concluded in Section 6.

2. Related work

Unsupervised grammar induction based on phrase alignment was introduced by Chiang (2005) as hierarchical phrase-based model. This model labeled all nonterminals with one generic label. 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). Discontinuous generation of target words limits pruning of the decoding space with the target language model. Watanabe et al. (2006) generated a target sentence in left-to-right order using hierarchical rules, the target sides of which were in the Greibach Normal Form class. Another work limited decoding space of the hierarchical phrase-based model by avoiding the recursion of hierarchical rules (Huck et al., 2012). It used two different nonterminals on the left and right hand sides of hierarchical rules. Zhou et al. (2008) scored derivations during translation decoding using syntactic knowledge. The hierarchical phrase-based model was augmented with syntactic categories (Zollmann and Venugopal, 2006) and CCG tags (Almaghout et al., 2010) to increase model precision. The precision of the proposed hierarchical model was increased with various nonterminals and without using syntactic categories.

There is a long history of using word classes in statistical machine translation. The alignment template model (Och and Ney, 2004) uses word classes to explicitly define word reordering. In some recent work, the classes of the boundary words in the aligned phrases are used to improve reordering in the hierarchical phrase-based model (Huck et al., 2012) and the phrase-based model (Cherry, 2013). Vilar et al. (2010) induced grammar nonterminals from the training corpus by clustering aligned phrases based on the source and target word classes. They used all words in the phrases, while

Zollmann and Vogel (2011) gave more importance to the boundary words of phrases and showed that translation quality of their model is similar to the SAMT model (Zollmann and Venugopal, 2006).

In comparison with Zollmann and Vogel (2011), the current study also labels nonterminals with classes of phrases boundary words, but differs in both labeling of the rules and the methods for extracting them. Firstly, labeling of the rules with both coarse grained (as their work) and fine grained nonterminals are examined in this work. As the nonterminal naming proposed in this paper differs from their work, the translation qualities of two namings are compared with each other in the experiments. Secondly, filtered rule extraction methods proposed in this paper extract far fewer rules than the basic rule extraction method which is also used in their work. Clearly, extraction of smaller grammar requires less time and memory. Moreover, the experiment shows that the filtered grammar has considerably less decoding time than non-filtered one.

There are two approaches to reduce the size of the hierarchical phrase-based model. The first approach filters the extracted grammar rules, and the second one prevents the extraction of too many rules by changing the method of rule extraction. The extracted grammar rules are filtered by discarding monotone rules whose source sides appear only in the monotone rules (He et al., 2009). Iglesias et al. (2009) categorized the rules to different patterns and filtered out those patterns which could be discarded without significantly impacting translation quality. Lee et al. (2012) discarded rules based on the information redundancy encoded in the translation rules. Considering the second approach that changes the extraction of grammar rules, Sankaran et al. (2011) examined the alignment probability of words to restrict the extraction of rules from aligned phrase pairs. Sankaran et al. (2012) extracted a minimum set of translation rules based on which at least one derivation could be constructed for each phrase pair. The proposed filter in this paper changes the method of rule extraction, too. It reduces the extraction of hierarchical rules from phrase pairs that are decomposable to two aligned subphrases.

3. Phrase-boundary model

A new hierarchical model for statistical machine translation is proposed. The model defines weighted rules in a Probabilistic Synchronous Context-Free Grammar (PSCFG) as phrase-boundary grammar. Grammar rules are automatically extracted from aligned sentence pairs of the parallel training corpus. We donate one aligned sentence pair in the parallel corpus by $<f, e, \sim>$, where f is a source sentence, e is a target sentence, and $\sim$ indicates a many-to-many symmetrized word alignment (Och and Ney, 2003) between source and target sentences. The grammar is defined on a set of aligned phrase pairs over word aligned sentences $<f, e, \sim>$ with the conditions that no word is aligned outside the phrase pair and at least one aligned word exists in the phrase pair (Definition 1).

Definition 1. For the aligned sentence pair $<f, e, \sim>$, f_i^j stands for the inclusive source substring from position i to j, and e_m^n stands for the inclusive target substring from position m to n; then, $< f_i^j, e_m^n >$ is an aligned phrase pair of $<f, e, \sim>$ if:

1. $\exists k \in [i, j], \acute{\kappa} \epsilon [m,n] : f_k \sim e_{\acute{\kappa}}$
2. $\forall k, \acute{\kappa}, : f_k \sim e_{\acute{\kappa}}, k \in \left[i, j\right] \leftrightarrow \acute{\kappa} \epsilon [m,n]$

In the conditions above, f_k and $e_{\acute{\kappa}}$ stand for the words in positions k and $\acute{\kappa}$ of the source and target sentences respectively.

The synchronous grammar rules extracted from the aligned phrase pair $< f_i^j, e_m^n >$ have the general form $X_{m,n} \rightarrow <\alpha, \beta, \sim>$. In this form, both α and β are strings of terminals and nonterminals, $\sim$ indicates a one-to-one correspondence between nonterminals in α and β, and $X_{m,n}$ is the left hand side nonterminal defined as other nonterminals by boundary word classes on the target side (Definition 2).

Definition 2. Using a function C that returns the word class of its word argument, 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 defined as:

$$X_{m,n} = \begin{cases} C(e_m) \quad \underline{Glue} \quad C(e_n), & m < n \\ C(e_m), & m = n \end{cases}$$

$$\underline{Glue} = `-' \, (one\, hyphen\, as\, default\, glue)$$

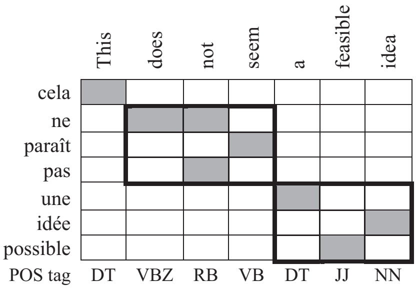

Fig. 1. Alignment between the French-English sentence pair: <cela ne paraît pas une idée possible, this does not seem a feasible idea>.

According to Definition 2, two boundary word classes on the target side are concatenated by one hyphen for phrases with a length greater than one. To achieve more precision, the glue notation could be selected based on the corresponding phrase pair (such as the case examined in the experiments).

The function C: $w \rightarrow \{1, \ldots, c\}$ maps a word to a word class number, and c is a fixed number of word classes. Word classes are defined using POS tags or automatic word clustering on the target training corpus. Automatic word clustering maps a word type from vocabulary to a word class which is not context dependent. In this study, MKCLS tool (Och, 1999) was used, which clusters words to an arbitrary number of word classes. With Definition 2, the grammar contains at most $c^2 + c$ nonterminals for c word classes (c^2 nonterminals for multi-word target phrases and c nonterminals for single-word target phrases). Using automatic word clustering, the phrase-boundary grammar can be defined with coarse grained or fine grained nonterminals, depending on the number of word classes. As an example of fine grained nonterminals, there are at most 2550 nonterminals for 50 word classes.

Following Chiang (2005), grammar has lexical, hierarchical, and glue rules which are induced from the aligned sentences. Lexical rules represent aligned phrase pairs with no nonterminal on the right hand side. Hierarchical rules are defined by substituting subphrases with nonterminals. At most, two nonterminals are allowed on the right hand side of the hierarchical rules. This condition restricts the number of derivations in the decoding of one input span. Glue rules are defined for all grammar nonterminals. These rules support the serial combination of phrases to form the start symbol of the grammar.

To explain the rules and nonterminals of phrase-boundary grammar, consider the alignment between the French sentence and its English translation in Fig. 1. Using POS tags for word classes, the following lexical rules are extracted from phrase pairs highlighted with thickened lines in this figure:

$$VBZ\text{-}VB \rightarrow <\text{ ne paraît pas, does not seem }> \tag{2}$$

$$JJ\text{-}NN \rightarrow <\text{ idée possible, feasible idea }> \tag{3}$$

$$DT\text{-}NN \rightarrow <\text{ une idée possible, a feasible idea }> \tag{4}$$

Hierarchical rules are defined by substituting subphrases with nonterminals. For example, the following hierarchical rules are extracted from the bordered phrase pairs in Fig. 1:

$$VBZ\text{-}VB \rightarrow <\text{ ne } VB^{\sim 1} \text{ pas, does not } VB^{\sim 1} > \tag{5}$$

$$JJ\text{-}NN \rightarrow < NN^{\sim 1} JJ^{\sim 2}, JJ^{\sim 2} NN^{\sim 1} > \tag{6}$$

$$JJ\text{-}NN \rightarrow <\text{ idée } JJ^{\sim 1}, JJ^{\sim 1}\ idea > \tag{7}$$

$$JJ\text{-}NN \rightarrow < NN^{\sim 1} \text{ possible, feasible } NN^{\sim 1} > \tag{8}$$

$$DT\text{-}NN \rightarrow < DT^{\sim 1} JJ-NN^{\sim 2}, DT^{\sim 1} JJ-NN^{\sim 2} > \tag{9}$$

$$DT\text{-}NN \rightarrow <\text{ une } JJ-NN^{\sim 1}, \text{a } JJ-NN^{\sim 1} > \tag{10}$$

$$DT\text{-}NN \rightarrow <\text{ une } NN^{\sim 1} \text{ possible, a feasible } NN^{\sim 1} > \tag{11}$$

As mentioned before, the notation ∼ indicates a one-to-one correspondence between nonterminals on the source and target side.

Above hierarchical rules are supported by other lexical rules such as:

$$DT \rightarrow < \text{une, a} > \tag{12}$$

$$NN \rightarrow < \text{idée, idea} > \tag{13}$$

Glue rules are defined for all grammar nonterminals to support their derivation from the start symbol of grammar. For example, the following glue rules are defined for *JJ-NN* nonterminal (*S* is the start symbol of grammar):

$$S \rightarrow < JJ\text{-}NN^{\sim 1}, JJ\text{-}NN^{\sim 1} > \tag{14}$$

$$S \rightarrow < S^{\sim 1}\ JJ\text{-}NN^{\sim 2}, S^{\sim 1}\ JJ\text{-}NN^{\sim 2} > \tag{15}$$

The decoding process searches among possible derivations of the source sentence by different rules to find the derivation with the lowest cost. The weights of rules are computed in a log-linear model (Och and Ney, 2002) based on the features extracted for each rule. Rule features have a zero value for the glue rules. Based on Chiang (2005), a penalty for using binary glue rules (such as rule 15) was considered in this study. The bigger penalty value gives more preference to the hierarchical rules over the serial combination of phrases with glue rules. In decoding, there is no limit on the length of input which is accepted by one glue rule; therefore, a whole sentence can be accepted by glue rules.

4. Filtered rule extraction

The phrase-boundary model, like other hierarchical models, extracts a huge number of hierarchical rules. As mentioned earlier, the base method of rule extraction uniformly extracts hierarchical rules from all aligned phrase pairs by substituting subphrases with nonterminals. To decrease the number of hierarchical rules in the proposed model, the method of rule extraction is changed. This novel approach restricts the patterns of hierarchical rules based on the alignment pattern of phrase pairs. In this paper, the frequent phrase pairs that are decomposable to two aligned subphrases (Definitions 3 and 4) are considered. This filtering is based on the fact that 2-decomposable phrase pairs can be presented by simpler hierarchical rules containing fewer strings of words and nonterminals.

Definition 3. An aligned phrase pair $< f_i^j, e_m^n >$ is **2-decomposable** if:

- $\exists k \in [i, j), \exists a, b, c, d \in [m, n] : < f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$ are aligned phrase pairs.

Definition 4. An aligned phrase pair $< f_i^j, e_m^n >$ is **monotone 2-decomposable** if:

- $\exists k \in [i, j), \exists a, b, c, d \in [m, n], b<c : < f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$ are aligned phrase pairs.

In other words, a 2-decomposable phrase pair has at least two aligned subphrases on the source and target sides with no overlap in their spans that cover the source side phrase. According to the condition $b<c$ in Definition 4, the source and target subphrases have the same order in a monotone 2-decomposable phrase pair. Therefore, a subset of 2-decomposable phrase pairs is monotone 2-decomposable. Considering the conditions of aligned phrase pairs in Definition 1, the words outside the target subphrases of 2-decomposable phrase pairs (e_a^b and e_c^d) are unaligned.

Fig. 2 presents two examples of 2-decomposable phrase pairs (Definition 3) according to the alignment of words in Fig. 1. Phrase pair (a) in this figure is a monotone 2-decomposable phrase pair (Definition 4). Some aligned phrase pairs are not decomposable to distinct subphrases. For example, the phrase pair <ne paraît pas, does not seem> is not 2-decomposable according to the word alignment in Fig. 1. Experiments in this study showed that the majority of

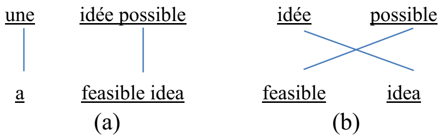

Fig. 2. Examples of (a) monotone and (b) cross alignment. between subphrases of 2-decomposable phrase pairs.

aligned phrase pairs are 2-decomposable. Moreover, in language pairs with similar word ordering, most aligned phrase pairs are monotone 2-decomposable because of the monotonic alignment in these languages.

The proposed filtering restricts patterns of extracted hierarchical rules from 2-decomposable and monotone 2-decomposable phrase pairs. Considering that the rules of the phrase-boundary grammar contain at most two nonterminals, a hierarchical rule can be formed with one of the following patterns on the source or target side:

$$\begin{aligned} rule\_patterns = \{&x_1x_2,\, x_1w_1,\, w_1x_1,\, w_1x_1w_2,\, x_1w_1x_2,\, x_1x_2w_1,\, w_1x_1x_2\ , \\ &w_1x_1x_2w_2,\, x_1w_1x_2w_2,\, w_1x_1w_2x_2,\, w_1x_1w_2x_2w_3\} \end{aligned} \tag{16}$$

In this set, the notations w and x denote one string of words and one nonterminal, respectively. In the top-down view of decoding, a hierarchical rule the right hand side of which explains subphrases of the phrase on its left hand side represents the decomposition pattern of one phrase pair to the subphrases. Regardless of the target side pattern of rules, the rules can be filtered according to the fact that hierarchical rules with two parts ($x_1\ x_2$, $x_1\ w_1$, $w_1\ x_1$) on the source side can present the decomposition of 2-decomposable phrase pairs. Of course, these patterns are more ambiguous than patterns with more parts. Using patterns of rules with few strings of words and nonterminals on the source sides, two different filters that restrict the extraction of hierarchical rules from 2-decomposable and monotone 2-decomposable phrase pairs are made. The patterns used in these filters contain at least one of the mentioned patterns with two parts. As explained in the following, the proposed filters do not discard the extracted rules, but change the rule extraction method to prevent the extraction of too many rules. In addition to a considerable reduction in model size, this approach reduces the amount of required resources for training the model.

Definition 5. A hierarchical rule with two nonterminals is **non-lexical** if the rule contains no words on the source side (source side pattern: $x_1\ x_2$).

4.1. Non-lexical filter for 2-decomposable phrase pairs

Non-lexical rules (Definition 5) can represent the decomposition of 2-decomposable phrase pairs. This paper proposes a filter for phrase-boundary grammar that only extracts non-lexical rules from 2-decomposable phrase pairs. As a formal definition, the proposed filter accepts the candidate rules with the following conditions:

1. All lexical rules (with no nonterminal).
2. All non-lexical rules.
3. Hierarchical rules with corresponding phrase pairs that are not 2-decomposable.

Rules 6 and 9 in Section 3 are examples of non-lexical rules with both cross and monotone alignments between subphrases. Non-lexical rules are purely defined by word classes on the source side. Although defining grammar with non-lexical rules decreases the model size considerably, it increases the model ambiguity. Increasing the granularity of nonterminals increases the precision of the filtered model. Indeed, non-lexical filtering and fine grained nonterminals are inseparable. The granularity of nonterminals in the phrase-boundary model is increased by more word classes in the model training. With automatic word clustering, fine grained nonterminals can be labeled with the proper number of word classes.

To explain the proposed filtering, consider again rules 2 to 11 in section 3 for the aligned sentences in Fig. 1. The non-lexical filter accepts lexical rules 2, 3, and 4 and the non-lexical rules 6 and 9. Rules 5, 7, 8, 10, and 11 are not non-lexical. As the corresponding phrase pair of rule 5 is not 2-decomposable, it is accepted, but other rules which correspond to the 2-decomposable phrase pairs are rejected.

4.2. Monotonic filter for monotone 2-decomposable phrase pairs

The monotonic word alignment in phrase pairs results in many monotone-aligned subphrases with overlap in their spans. More aligned subphrases (as placeholders for nonterminals) increase the number of hierarchical rules in the model. A filter is proposed for phrase-boundary grammar that restricts the position or number of nonterminals in the extraction of hierarchical rules from monotone 2-decomposable phrase pairs (Definition 4). Table 1 introduces

Table 1
Filtering patterns for extraction of rules from monotone 2-decomposable phrase pairs (Notations w and x denote one string of words and one nonterminal.).

Filtering pattern	Nonterminals		Source side patterns of the rules
	Number	Source side position	
boundary$_1$	1	Boundaries	$x_1\ w_1,\ w_1\ x_1$
boundary$_2$	Max. 2	Boundaries	$x_1\ w_1,\ w_1\ x_1,\ x_1\ w_1\ x_2$
floating$_1$	1	Free	$x_1\ w_1,\ w_1\ x_1,\ w_1\ x_1\ w_2$

the examined filtering patterns (source side patterns of accepted rules) which can present the decomposition of 2-decomposable phrase pairs.

The *boundary*$_1$ filtering pattern restricts the number of nonterminals to one and their source side position to the boundaries, the *boundary*$_2$ filtering patterns restrict the position of nonterminals to the source boundaries of rules, and the *floating*$_1$ filtering pattern restricts the number of nonterminals to one. In the rule pattern set (16), patterns other than those selected in Table 1 have two nonterminals, at least one of which is not located in the source boundaries. The coverage of the proposed filtering patterns has the following relationship:

$$\begin{aligned} &\mathit{boundary}_1 \subset \mathit{boundary}_2 \\ &\mathit{boundary}_1 \subset \mathit{floating}_1 \end{aligned} \tag{17}$$

A filtering pattern with more coverage of rules results in a larger model size. According to the experiments *boundary*$_2$ filtering pattern results in a larger model size and a higher translation quality. As a formal definition, the proposed filter accepts the candidate rules with the following conditions:

1. All lexical rules (with no nonterminal).
2. All hierarchical rules consistent with the selected filtering pattern (Table 1).
3. Hierarchical rules with corresponding phrase pairs that are not monotone 2-decomposable.

To explain the proposed filtering, consider rules 2 to 11 in section 3 for the aligned sentences in Fig. 1. For example, the *boundary*$_1$ filtering pattern accepts lexical rules 2, 3, and 4 and hierarchical rules 7, 8, and 10. Non-lexical rules 6 and 9 with no words on the right hand side are not included in the candidate rules of the model because of their low precision. Source patterns of rules 5 and 11 are not in the *boundary*$_1$ filtering pattern. Since the corresponding phrase pair of rule 5 is not monotone 2-decomposable, it is accepted, but rule 11 which corresponds to the monotone 2-decomposable phrase pair is rejected.

4.3. Features of filtered model

The results of the current study showed that the proposed filtering decreased the model size by more than 70%. This demonstrates that generally, most aligned phrase pairs are 2-decomposable. Furthermore, aligned phrase pairs are monotone 2-decomposable in languages that have similar word ordering. Thus, more frequent use of lexical rules and the extracted hierarchical rules from 2-decomposable phrase pairs in the translation of sentences can be expected. In addition to conventional features, a penalty feature is proposed to penalize the rules that are not similar to the patterns of expected rules. The following feature has been added to the models that are filtered by different patterns:

- Pattern penalty: this feature has a value of 0 if the rule is lexical or consistent with the filtering pattern; otherwise, it has a value of 1. The filtering pattern is either non-lexical (Definition 5) or one of the monotonic filtering patterns (Table 1).

For example, in the non-lexical filtered model, this feature penalizes the rules that contain both words and nonterminals on the source side. Of course, the weight of this proposed feature should be computed together with the weights of the model's other features.

4.4. *Coverage of filtered grammar*

Only extraction of hierarchical rules from 2-decomposable or monotone 2-decomposable phrase pairs is restricted by the proposed filters. Considering the two parts of these phrase pairs, rules with two parts ($x_1\ x_2$, $w_1\ x_1$ or $x_1\ w_1$) on the source side are included in the filtering patterns. In the following, two lemmas are demonstrated to prove the coverage of the filtered grammar for the aligned phrase pairs in the training corpus. The proofs show how the phrase pairs are derived in the filtered grammar.

Lemma 1. *If the aligned phrase pair* $< f_i^j, e_m^n >$ *in the training corpus is derivable from the non-filtered grammar then it is also derivable from the non-lexical filtered one.*

Proof: Derivation can be done with the same set of rules in the non-filtered grammar, if $< f_i^j, e_m^n >$ is matched by the right hand side of one lexical rule or if it is not 2-decomposable. Otherwise, according to Definition 3, phrase pair $< f_i^j, e_m^n >$ can be obtained by composition of phrase pairs $< f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$ with one non-lexical rule in the filtered grammar as:

$$X_{m,n} \rightarrow < X^{\sim 1}_{a,b} \quad X^{\sim 2}_{c,d}, \ldots > \tag{18}$$

Considering rule 18, derivability must be proved for the subphrases $< f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$. As aligned subphrases in the training corpus are derivable in the non-filtered grammar, so recursively, both are derivable in the filtered one. ■

As mentioned, the best translation result of non-lexical filtered model is obtained from fine grained nonterminals that increase the precision of non-lexical rules, while decrease the coverage of other hierarchical rules. Although this lemma is not affected by different granularities of nonterminals, fine grained nonterminals may decrease the coverage of non-lexical filtered grammar for the phrase pairs other than the ones observed in training.

Lemma 2. If the aligned phrase pair $< f_i^j, e_m^n >$ in the training corpus is derivable from the non-filtered grammar then it is also derivable from the monotonic filtered one.

Proof: Derivation can be done with the same set of rules in the non-filtered grammar if $< f_i^j, e_m^n >$ is matched by the right hand side of one lexical rule or if it is not monotone 2-decomposable. Otherwise, according to Definition 4, phrase pair $< f_i^j, e_m^n >$ can be obtained by monotone composition of phrase pairs $< f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$. Regardless of other hierarchical rules, there is at least the following binary glue rule in the grammar that supports serial combination of these phrase pairs:

$$S \rightarrow < S^{\sim 1} X^{\sim 2}_{c,d}, S^{\sim 1} X^{\sim 2}_{c,d} > \tag{19}$$

Considering rule 19, derivability must be proved for the subphrases $< f_i^k, e_a^b >$ and $< f_{k+1}^j, e_c^d >$. As aligned subphrases in the training corpus are derivable in the non-filtered grammar, so recursively, both are derivable in the filtered one. ■

The number of tokens in the right hand side of hierarchical rules is conventionally limited to five tokens (including nonterminals). Therefore, a long phrase pair may not be presented as $w_1\ x_1$ or $x_1\ w_1$ due to length of string w_1, while subphrases of a monotone 2-decomposable phrase pair can be respectively concatenated by one binary glue rule (such as rule 15). This is why monotonic filter is only proposed for monotone 2-decomposable phrase pairs with no need for reordering of subphrases. Of course, hierarchical rules support linguistic dependencies better than glue rules. The $boundary_2$ filtering pattern with two nonterminals has more chance to keep long phrase pairs.

5. Experiments

The performance of filtered and non-filtered phrase-boundary models in translating language pairs with similar and different word orders was studied. A set of experiments was performed for Persian-English, French-English, and Spanish-English translations. Whereas the word order in the European languages is similar, Persian has a different word order. Although Persian is almost a free word order language, its formal sentence structure is SOV, which differs from

Table 2
Statistics of the parallel training corpora.

Training corpora	No. of words	No. of word types
Persian-English	15M + 13M	135K + 98K
French-English	22M + 20M	91K + 73K
Spanish-English	21M + 21M	112K + 74K

the SVO structure of English. Persian language is head-initial in noun phrases, while English is head-final. Our Persian-English translation system is trained on the Mizan corpus (Supreme Council of ICT, 2013). The 1k + 1K sentences of this corpus were selected for the development and test sets. French-English and Spanish-English translation systems is trained on the first 1M sentences of Europarl-V7 corpus (Koehn, 2005). The last 500 sentences of WMT 2012 translation task were used for the development set, and the first 1K sentences of WMT 2013[1] translation task were used for the test set. The statistics of the used corpora are presented in Table 2.

The baseline of the experiments in the current study was the hierarchical phrase-based model (Chiang, 2007), which was configured with the default settings of this model. The models were trained and evaluated with Joshua toolkit (Li et al., 2009). The words were aligned in both directions of translation by GIZA++ (Och and Ney, 2000), and the results were symmetrized. The trigram language models on the target side corpora were built with Berkeley LM tool (Pauls and Klein, 2011) with Kneser–Ney smoothing. The scaling factors of the models were trained by Minimum Error Rate Training (Och, 2003). The BLEU-4 (Papineni et al., 2002) was the metric of the evaluations.

A new version of Thrax 2.0 (Post et al., 2013) – the grammar extraction tool of Joshua toolkit – was developed to support the phrase-boundary grammar and proposed filtering in this paper. The extraction of the phrase-boundary grammar also needs defined word classes. Word classes were defined by POS tags and automatic word clustering. SENNA tool (Collobert et al., 2011) was used to label POS tags. MKCLS tool (Och, 1999) was used to cluster English words. This tool automatically clusters words to the arbitrary number of classes which are not context dependent.

5.1. *Features and parameters*

The baseline model was configured with the default settings of the hierarchical phrase-based model. These settings limit lexical rules to a length of 10 and hierarchical rules to a length of 5, including a maximum of two nonterminals. The following default features for grammar rules were also used:

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

The length of the rules and the features selected for the phrase-boundary model were the same as those of the baseline model, but the pattern-penalty feature (Section 4.3) was used as well as other features when the phrase-boundary model was filtered by the proposed filters.

5.2. *Summary of translation results*

This section presents the translation results of different translation models using the best parameters investigated in the next sections. Then, the results of significance tests on experiments with larger language models are presented. Finally, post-filtering of extracted grammar along with filtered rule extraction methods is examined.

[1] Available at http://statmt.org/wmt13/translation-task.html.

Table 3
The best translation results of the hierarchical phrase-based model (baseline), the Zollmann and Vogel (2011) model (Z&V) and the phrase-boundary model.

Model/*filtering*	Persian-English			French-English			Spanish-English		
	Rules	BLEU	Time	Rules	BLEU	Time	Rules	BLEU	Time
baseline	11	11.75	0.2	176	27.07	0.1	206	23.36	0.2
Z&V	30	11.73	16.4	969	28.78	15.0	1183	24.53	67.9
boundary	29	12.27	21.2	916	28.66	13.8	1115	24.54	54.8
boundary/*non-lexical*	7	12.29	10.1	57	27.98	9.4	65	23.13	8.9
boundary/*monotonic*	16	11.95	7.3	249	28.63	2.0	295	24.18	1.3

5.2.1. *Results using the best parameters*

According to the results (Sections 5.3 and 5.4), the best phrase lengths for extraction of hierarchical rules in the hierarchical phrase-based (baseline), phrase-boundary, monotonic filtered and non-lexical filtered models are 10, 12, 12 and 14 respectively. Also, the results in Section 5.4 illustrate that the non-lexical filter has better performance with fine grained nonterminals and the monotonic filter has better performance with $boundary_2$ filtering pattern. The best results of experiments with the phrase-boundary model (filtered and non-filtered) are presented in Table 3. Each set of values in this table respectively presents the number of rules in millions, the resulting case-insensitive BLEU score and the average decoding time per sentence in seconds.

As mentioned before, to achieve more precision, the $\underline{Glue}$ notation (Definition 2) could be selected based on the corresponding phrase pair. Table 3 also presents the models named Z&V that define the $\underline{Glue}$ notation as Zollmann and Vogel (2011) to name nonterminal of phrase pair $< f_i^j, e_m^n >$:

$$\underline{Glue} = \begin{cases} '-', & n = m + 1 \\ '..' & n > m + 1 \end{cases} \tag{20}$$

In other words, this more precise labeling distinguishes two word phrases from longer ones.

The results indicate that the non-filtered phrase-boundary model has several times larger size than the hierarchical phrase-based model and the non-lexical filtered model has the minimum size among different models. Filtered models have considerably less decoding time than non-filtered ones.

5.2.2. *Significance tests*

As mentioned, 3-gram target language models are used in the experiments reported in Table 3; Translation results using larger language models are presented in Table 4. The BLEU scores in this table are also computed with 95% confidence interval (score*) using bootstrap resampling method (Koehn, 2004) to study the statistical significance of the improvements. Larger language models are 4-grams trained on part 2008 of MultiUn corpus (Eisele and Chen, 2010) (51M words) and the target sides of the parallel training corpora.

The score* in this table is computed by paired bootstrap resampling script in the Moses decoder (Koehn et al., 2007) between each model and the baseline model. In Persian-English translation, the phrase-boundary model has less improvement over the baseline model due to sparseness of the model with more precise labeling. Other results indicate that improvement of the phrase-boundary model (non-filtered and monotonic filtered) over the hierarchical

Table 4
BLEU scores using 4-gram language models also computed with 95% confidence interval (Score*).

Model/*filtering*	Persian-English			French-English			Spanish-English		
	BLEU	Score*	p-value	BLEU	Score*	p-value	BLEU	Score*	p-value
baseline	12.16	12.22	n/a	28.21	28.09	n/a	24.07	23.89	n/a
Z&V	12.23	12.43	0.204	29.53	29.40	0.0	25.03	24.88	0.0
boundary	12.51	12.61	0.110	29.53	29.40	0.0	24.83	24.67	0.001
boundary/*non-lexical*	12.42	12.48	0.091	28.60	28.52	0.073	23.59	23.40	0.045
boundary/*monotonic*	11.98	12.03	0.149	29.68	29.59	0.0	24.68	24.43	0.012

Table 5
Translation results of monotonic filtering and post-filtering with ‘min-rule-count’ threshold (MRC > 1).

Model/*filtering*	MRC	French-English			Spanish-English		
		Rules	BLEU	Time	Rules	BLEU	Time
boundary	1	916	29.53	12.8	1115	24.83	52.5
boundary	2	128	29.32	9.4	161	24.29	9.1
boundary	3	52	29.25	5.8	64	24.17	6.2
boundary/*monotonic*	1	249	29.68	2.7	295	24.68	3.1
boundary/*monotonic*	2	47	29.29	0.8	58	24.86	1.0

phrase-based model is statistically significant with “p-value < 0.05″. However there is no obvious difference between the translation quality of default naming (Definition 2) and naming of Zollmann and Vogel (2011) (Z&V), the *Glue* notation could be selected based on the training corpus to achieve better performance.

5.2.3. Post-filtering of extracted grammar

A known post-filtering of extracted grammar is introduced by Zollmann et al. (2008). They discarded rare hierarchical rules occurred fewer times than a given threshold and showed that increasing this threshold decreases the translation quality. Bearing some impacts on translation quality, Zollmann and Vogel (2011) set this threshold to six for a large training corpus. The parameter ‘min-rule-count’ in Thrax 2.0 (Post et al., 2013) – the grammar extraction tool – is used for discarding rare rules. This way, the rules with frequencies lower than ‘min-rule-count’ are discarded from extracted grammar. Due to the large model size for the European language pairs, we apply this post-filtering on the monotonic filtered grammar in French-English and Spanish-English translation systems. Table 5 presents the results of these experiments using the 4-gram language models along with some previous results, for easy comparison.

The first three models in this table are not filtered in the extraction step. The “MRC > 1″ means post-filtering of extracted grammar with ‘min-rule-count’ threshold. The last model in this table is extracted with the monotonic filter and then singleton rules are discarded from the extracted grammar. In comparison to the models of similar sizes (MRC = 3), the monotonic filtered models have less decoding time due to the existence of more useful rules in the grammar. It is noteworthy that the grammar extraction requires less time and memory using filtered rule extraction methods such as monotonic filter.

5.3. Results of non-filtered models

The best translation results of the non-filtered phrase-boundary model and the hierarchical phrase-based model (baseline) were presented in Table 3. Using the same measurement units, Table 6 presents the translation results of these models with different phrase lengths. The phrase length in this table is the maximum length of phrase pairs included in the extraction of hierarchical rules. The length of 10 words is the default phrase length of the hierarchical phrase-based model; the models are examined with this phrase length and some longer phrase lengths that may improve translation quality.

Table 6
Translation results of the hierarchical phrase-based (baseline) and the phrase-boundary models with different phrase lengths.

Model/phrase length	Persian-English		French-English		Spanish-English	
	Rules	BLEU	Rules	BLEU	Rules	BLEU
baseline/**10**	11	11.75	176	27.07	206	23.36
baseline/12	11	11.71	202	27.09	237	22.99
baseline/14	11	11.40	224	26.86	n/a	n/a
boundary/10	26	12.06	684	28.39	828	23.90
boundary/**12**	29	12.27	916	28.66	1115	24.54
boundary/14	31	12.11	1140	27.41	n/a	n/a

Table 7
Results of Persian-English translation with word clustering.

Phrase length	Word classes	Rules	BLEU
10	5	22	11.95
10	10	24	12.19
10	15	25	10.77
12	10	27	**12.11**
14	10	29	11.93

In these experiments, POS tags are used to label rules in the training of the phrase-boundary model. The results indicate that phrases longer than 10 words in length cannot improve the translation quality of the hierarchical phrase-based model, while the phrase length of 12 words results in a better translation quality with the non-filtered phrase-boundary model. The longer phrase length supports handling of long distance dependencies, but it also increases the grammar ambiguity due to higher nonterminal coverage. In the proposed phrase-boundary model, nonterminal coverage is restricted by the boundary word classes of aligned phrase pairs.

One language pair is selected for other experiments with the non-filtered phrase-boundary model using word clustering instead of POS tags. Table 7 presents the results of these experiments in Persian-English translation including the number of word classes used in the training of the phrase-boundary model.

The results indicate a better translation quality for the phrase-boundary model with about 10 word classes, which is lower than the number of English POS tags. More word classes cause a sparser model, because words with the same POS tag may have different classes by word clustering. According to the results shown in Tables 6 and 7, there is no significant difference in translation quality between the models which are trained by word clustering and those trained by POS tags. Of course, defining word classes using word clustering requires multiple runs of training and evaluation to determine the optimum number of word classes.

5.4. Results of filtered models

The best translation results of the phrase-boundary model extracted with the non-lexical filtering were reported in Table 3. Using the same measurement units, Table 8 presents the translation results of the non-lexical filtered models with different phrase lengths and different number of word classes. The reduction ratio (RR) presents the percentage of reduction in the model size computed based on the size of non-filtered phrase-boundary grammar (Table 6) of the closest phrase length.

Table 8
Translation results of non-lexical filtering along with the Reduction Ratio.

System	Phrase length	Word classes	Rules	BLEU	RR
Persian-English	12	POS	5	11.21	83
		50	7	11.84	76
	14	POS	5	11.12	84
		40	6	11.36	81
		50	7	**12.29**	77
		60	7	12.00	77
French-English	12	70	46	26.95	95
		80	49	27.14	95
		90	51	27.21	94
		100	54	27.68	94
		110	56	27.45	94
	14	100	57	**27.98**	95
Spanish-English	12	90	57	23.07	95
		100	62	23.19	94
		110	63	22.55	94
	14	100	65	**23.13**	94

Table 9
Translation results of monotonic filtering along with the Reduction Ratio.

	Persian-English			French-English			Spanish-English		
Filtering Pattern	Rules	BLEU	RR	Rules	BLEU	RR	Rules	BLEU	RR
boundary$_1$	11	11.20	62	136	28.29	85	158	23.58	86
floating$_1$	13	11.49	55	215	28.43	77	253	23.86	77
boundary$_2$	16	11.95	45	249	28.63	73	295	24.18	74

Table 10
Results of applying filters on all phrase pairs along with the Reduction Ratio.

System	Filtering pattern	Phrase length	Word classes	Rules	BLEU	RR
Persian-English	*non-lexical*	14	50	6	11.55	81
French-English	*boundary*$_2$	12	POS	246	28.00	73

Trainings are performed using POS tags and word clustering on the target side of the parallel training corpus. As expected, non-lexical filtered models achieve a better translation quality using fine grained nonterminals obtained from a large number of automatically extracted word classes. Although the phrase length of 14 words improves the performance of non-lexical filter, this filter affects the quality of French-English and Spanish-English translations due to high ambiguity of non-lexical rules.

The best translation results of the phrase-boundary model extracted with the monotonic filtering were reported in Table 3. Table 9 presents the translation results of the monotonic filtered models using different filtering patterns (Table 1).

According to the results, the monotonic filter has a better performance with *boundary*$_2$ filtering pattern (containing at most two nonterminals). Indeed, ignoring rules with more parts (including strings of words or nonterminals) slightly increases the ambiguity of the grammar. As most of the aligned phrase pairs in the training corpora for the European language pairs are monotone 2-decomposable, the monotonic filter reduces the size of French-English and Spanish-English translation models more than the size of the Persian-English translation model. Note that the model size of French-English and Spanish-English translations without filtering (Table 6) are tens of times greater than that of the Persian-English translation because of the monotone alignment of words in European language pairs.

The restricted extraction of rules from all phrase pairs is examined in the next experiments. Table 10 presents the results of applying the proposed filtering patterns (*non-lexical* and *boundary*$_2$) on all phrase pairs instead of 2-decomposable ones.

The results indicate that the translation quality is significantly affected by applying the *non-lexical* and *boundary*$_2$ filtering patterns on all phrase pairs, while the additional reduction of the model size is not remarkable. This shows that rules with multiple strings of words and nonterminals must exist in order to cover the decomposition of phrase pairs that are not 2-decomposable.

6. Conclusion

This paper proposed a novel hierarchical model for statistical machine translation that labels nonterminals by boundary word classes of aligned phrases. In addition to POS tags, labeling of the rules can be done by clustering words on the target corpus, but selecting the optimum number of classes requires multiple iterations of training and evaluation. The proposed model with more precise labeling outperformed the hierarchical phrase-based model with one generic nonterminal. Compared with syntactic models, the phrase-boundary model is applicable for under-resourced languages which lack linguistic tools such as a parser. Moreover, naming nonterminals based on boundary word classes is straightforward for all phrases and extracting word classes is faster than extracting parse trees.

Diversity of nonterminals increases the precision of the phrase-boundary model, but as a drawback increases its size which is several times larger than the hierarchical phrase-based model. To tackle this challenge, two filtered rule extraction methods were proposed based on the alignment pattern of phrase pairs; namely non-lexical and monotonic filter. Both filters restrict the extraction of hierarchical rules from phrase pairs that are decomposable to two aligned subphrases. The non-lexical and monotonic filters reduce considerably the required time and memory in the training and decoding processes. Our experiments showed that the filtered models have about 70% smaller size, while hierarchical rules are extracted from all aligned phrase pairs.

References

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

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

Chiang, D., 2005. A hierarchical phrase-based model. In: Proceedings of the 43rd Annual Meeting of the ACL, pp. 263–270.

Chiang, D., 2007. Hierarchical Phrase-Based Translation. Association for Computational Linguistics, pp. 201–228.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res., 2461–2505.

Eisele, A., Chen, Y., 2010. MultiUN: A Multilingual Corpus from United Nation Documents. LREC, pp. 2868–2872.

Haque, R., Naskar, S.K., Bosch, A., Way, A., 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 (AMTA 2010).

He, Z., Liu, Q., Lin, S., 2008. Improving statistical machine translation using lexicalized rule selection. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 321–328.

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

Huck, M., Peitz, S., Freitag, M., Ney, H.,2012. Discriminative reordering extensions for hierarchical phrase-based machine translation. In: Proceedings of the 16th EAMT Conference. European Association for Machine Translation, Trento, Italy, pp. 313–320.

Iglesias, G., Gispert, A., Banga, E.R., Byrne, W., 2009. Rule filtering by pattern for efficient hierarchical translation. In: Proceedings of EACL 2009, pp. 380–388.

Koehn, P., 2004. Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 388–395.

Koehn, P., 2005. Europarl: a parallel corpus for statistical machine translation. In: Proceedings of the tenth Machine Translation Summit, AAMT, pp. 79–86.

Koehn, P., Och, F.J., Marcu, D., 2003. Statistical phrase-based translation. In: Proceedings of the HLT-NAACL, pp. 127–133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

Lee, S.-W., Zhang, D., Li, M., Zhou, M., Rim, H.-C., 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, pp. 291–295.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W., Weese, J., Zaidanv, O., 2009. Joshua: an open source toolkit for parsing-based machine translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 135–139.

Och, F.J., 1999. An efficient method for determining bilingual word classes. In: Ninth conference on European chapter of the Association for Computational Linguistics, pp. 71–76.

Och, F.J., 2003. Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting of the ACL, pp. 160–167.

Och, F.J., Ney, H., 2000. Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the ACL, pp. 440–447.

Och, F.J., Ney, H., 2002. Discriminative training and maximum entropy models for statistical machine translation. In: 40th Annual Meeting of the ACL, pp. 295–302.

Och, F.J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Comput. Linguist., 19–51.

Och, F.J., Ney, H., 2004. The alignment template approach to statistical machine translation. Comput. Linguist. 30, 417–449.

Papineni, K., Roukos, S., Ward, T., Zhu, W., 2002. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Association for Computational Linguistics, pp. 311–318.

Pauls, A., Klein, D., 2011. Faster and smaller n-gram language models. In: Proceedings of the Association for Computational Linguistics, pp. 258–267.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., 2013. Joshua 5.0: sparser, better, faster, server. Eighth Workshop on Statistical Machine Translation, pp. 206–212.

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

Sankaran, B., Haffari, G., Sarkar, A., 2012. Compact rule extraction for hierarchical phrase-based translation. In: The 10th biennial conference of the Association for Machine Translation in the Americas (AMTA).

Supreme Council of ICT, 2013. Mizan English-Persian Parallel Corpus. I.R. Iran, Tehran.

Vilar, D., Stein, D., Peitz, S., Ney, H., 2010. If I only had a parser: poor man's syntax for hierarchical machine translation. In: International Workshop on Spoken Language Translation, pp. 345–352.

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

Zens, R., Och, F.J., Ney, H., 2002. Phrase-based statistical machine translation. In: Proceedings of the German Conference on Artificial Intelligence, pp. 18–32.

Zhou, B., Xiang, B., Zhu, X., Gao, Y., 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. In: Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2), pp. 19–27.

Zollmann, A., Venugopal, A., 2006. Syntax augmented machine translation via chart parsing. In: Proceedings of the Workshop on Statistical Machine Translation, HLT/NAACL.

Zollmann, A., Vogel, S., 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, pp. 1–11.

Zollmann, A., Venugopal, A., Och, F.J., Ponte, J., 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 1145–1152.