IJICTR International Journal of Information & Communication Technology Research

Volume 6- Number 3-Summer 2014 (63-76)

Phrase Alignments in Parallel Corpus Using Bootstrapping Approach

Leila Tavakoli
School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
leila.tavakoli@ut.ac.ir

Heshaam Faili
School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
hfaili@ut.ac.ir

Received: May 13, 2014 - Accepted: August 5, 2014

Abstract— **Word choice and word order problems are considered as fundamental barriers in statistical machine translation (SMT). These barriers are more pronounced in deficiencies of training corpus. Phrase-Based SMT has advantages in word choice and local word ordering process; so phrase alignment is effective in improving translation quality. In this paper, an approach for automatic alignment is proposed in which boosts up the machine translation quality. Since, alignment problem is more problematic with lack of training data, so we make corpus of phrase alignment with high precision. For this purpose, a novel phrase alignment approach in a bootstrapping manner is proposed. By bootstrapping on alignment model via using a number of features, the accuracy of the phrase table is improved iteratively. These features are based on the phrase table extracted from Moses, IBM Model 3 alignment probabilities, Google translator and fertility of candidates. Our experiments on English-Persian translation show an improvement about 4.17 BLEU points over the PB-SMT as baseline system.**

Keywords- Phrase-Based SMT, scarce training corpus, log-linear models, Maximum Entropy, Bootstrapping Approach, Fertility

I. INTRODUCTION

Automatic alignment could be defined as the problem of determining a translational correspondence between aligned sentences in a parallel corpus. This correspondence may be carried out in various levels such as word, chunk, phrase, and sentence.

Word is the basic unit of the alignment process in fertility-based models [1]. Alignment of words for idiomatic expressions, free translations, and missing contents faces with some challenging. In this regard, most current statistical models [1][2][3] treat the aligned sentences in the corpus as sequences of words. By using pairs of corresponding sequences of words extracted from parallel corpora, the translation process is modeled by Phrase Based Statistical Machine Translation (PB-SMT) systems. PB-SMT use local contextual information from phrase table and language model. So, reordering issues and local context information are considered more in phrase-based translation than word-based translation. As a result, translating expressions in both target side and source side in phrase-based system in comparison with word-based system leads to a better way. Thus, the translation quality in PB-SMT excels to Word Based Statistical Machine Translation (WB-SMT) [2].

Several studies have proposed to focus on the translation of phrases. In this regards, most systems (e.g., Moses) induce synchronous phrases from parallel corpora using a heuristic two-step pipeline. At

[1] Fertility-based models such as IBM models 3, 4 and 5 enable alignments between one word and several words.

first, this pipeline aligns a parallel corpus at the word level with grow-diag-final-and heuristic that is used for word alignment combination [4]. Second, this pipeline extracts translation rules from word-aligned sentence pairs. In this heuristic, by first running GIZA++ in both directions [5], symmetric word alignments are created and then grow-diag-final-and heuristic is applied. This method extracts corresponding phrase pairs from corresponding word pairs. This approach for inducing synchronous phrases from parallel corpora has a major problem: if a pair word is aligned by Giza++ incorrectly, phrase alignment extraction will be incorrect as well. In this paper, a method for phrase alignment extraction is proposed to alleviate this problem.

In the past, various generative models were proposed to learn translation rules directly from sentence pairs without the direct use of word alignments. Marcu and Wong [6] applied a joint probability model for Statistical Machine Translation (SMT). They illustrated that translation pairs, which are produced by using the joint model are more accurate than those produced by using IBM model 4. Cherry and Lin made use of phrasal Inversion Transduction Grammar (ITG) as an alternative to joint phrasal translation models [7]. They showed that the phrasal translation tables produced by the ITG are superior to those of the flat joint phrasal model. Zhang et al. combined the strengths of Bayesian modeling with the synchronous grammar in unsupervised learning of basic translation phrase pairs [8]. They evaluated their model by using traditional phrase extraction and found that a phrase table based on their system improves MT results over a phrase table extracted from traditional word level alignments. Simultaneously with Zhang et al. [8], DeNero et al. [9], introduced the first tractable Gibbs sampling procedure for estimating phrase pair frequencies. They used the formalism of ITGs, which was better able to explore the full space of phrase translation. Cohn and Blunsom proposed a generative Bayesian model of tree to string translation, which induces grammars that are both smaller and produce better translations than the previous heuristic two-stage approach that employs a separate word alignment step [10]. Neubig et al. [11] used an unsupervised model for joint phrase alignment and extraction using non-parametric Bayesian methods and inversion ITGs. They demonstrated that their proposed model reaches the same accuracy of traditional grow-diagonal heuristic while reduces the phrase table to a fraction of the original size. Levenberg et al. presented a non-parametric Bayesian model, which is able to directly model many-to-many alignments and align discontiguous phrases in both source and target languages [12]. In addition, their model has no restrictions on the length of a rule.

In the present task, we aim to extract phrase alignment directly from sentence pairs without the direct use of word alignments. For this purpose, all possible candidates of phrase pairs are extracted. The main idea of phrase alignment is based on training a log-linear model, which estimates the probability of phrase alignment candidates using sophisticated features. By bootstrapping on this alignment model, the accuracy of the phrase table will be improved iteratively. We use our model in English to Persian translation. Experimental results show that the phrasal translation table produced by our method is superior to those of a phrase table extracted from traditional word level alignments. In this paper, we use an English-Persian parallel corpus, which contains some novels and news text.

As mentioned above, we propose a log-linear model to incorporate all features in this research work. Log-linear models are appropriate to incorporate additional knowledge sources. These models have interested to researchers. Liu et al. [13] presented a framework for word alignment based on log-linear models. They used IBM Model 3 alignment probabilities, POS correspondence, and bilingual dictionary coverage as features. They showed that log-linear models significantly outperform IBM translation models. In addition, Xiao et al., presented a global log-linear model for synchronous grammar induction [14]. This model has capability of incorporating arbitrary features. They significantly outperform a competitive hierarchical phrase-based baseline system by using learned synchronous grammar rules. Huy et al. [15] used a log-linear translation model to integrating WSD features. They could bootstrap WSD models using unlabeled data in phrase-based SMT. Their results on English-Vietnamese translation showed that BLEU scores have been improved significantly.

Here, we present a framework for phrase alignment based on log-linear models. We employ four types of knowledge sources as feature functions. These phrasal features are included:

- **Bilingual phrase dictionary:** We make a bilingual phrase dictionary by using phrase pairs in phrase table[2] of Moses. This dictionary is considered as the base feature of our log-linear model.
- **IBM Model 3 alignment probabilities:** this feature is similar to lexical probability with some differences. In this feature we extract two directions of IBM model 3 from the parallel corpus used in training and bootstrapping. Then, a probabilistic dictionary is generated using two direction of IBM model 3. This feature will be described in section III.
- **Google probabilistic dictionary:** this feature is an additional knowledge source, which is extracted by using Google translator and all of English and Persian words in the mentioned parallel corpus.
- **Fertility:** this feature is trained by a word-level hand- aligned parallel corpus.

[2] This phrase table is extracted from parallel corpus, which is used in training and bootstrapping.

Statistical approaches are sensitive to the quality of the parallel corpora. This problem is more challenging when parallel training data is scarce. The idea of bootstrapping on a small size of training data rises from this perspective. Bootstrapping is a method for assigning measures of accuracy such as confidence intervals to estimate samples. In other words, bootstrapping approach is the best method for each task that there are some samples and a confidence measure to select the best samples [16].

Some of previous works of alignment are used a bootstrapping approach. Previously, Ma et al., simplified the task of automatic word alignment as several consecutive words, which are together correspond to a single word in the opposite language by bootstrapping on its output [17]. Their experiments on Chinese-English translation showed that BLEU scores have been improved on the state-of-the art phrase-based SMT system. Moreover, Pal and Bandyopadhyay analyzed the effects of chunk alignment on the results of the English-Bengali phrase based statistical machine translation system [18]. The objective of their approach was to align the chunks in a bootstrapping manner and then modifying the model by inserting the aligned chunks to the parallel corpus after each iteration of the bootstrapping process. Their approach system improved on the results of the English-Bengali translation task.

By having hand-aligned phrases in low size and an appropriate confidence measure to choose the best phrase pairs among the entire scored candidates of the phrase pair, bootstrapping approach is an appropriate method for phrase alignment. In this paper, we use a log linear model with some of knowledge sources as features to scoring all of candidates. Scores of candidates along with the value of precision are used to determine confidence measure.

The idea of improving phrase-based translation is based on proposed features on log-linear model in two parallel sentences. This task is started with hand-aligned phrases, which are obtained from small number of sentences. Then by using corresponding phrases, new phrases are recognized and added to the previous group in a bootstrapping manner. Produced phrase table has a lower size and higher precision in comparison with phrase table of Moses. The produced phrase table by our method can improve the local reordering.

This research work is based on bootstrapping approach and summarized in the following steps:

- **Candidate extraction:** all possible candidates in the sentence pair for phrase alignment are extracted. For this purpose, we consider candidates of alignments for word and phrase levels.
- **Filtering candidates:** All extracted candidates are filtered by some rules. In this step, candidates that do not align in a high probability, is recognized and eliminated.
- **Scoring candidates:** By using the proposed log-linear model with defined features, a probabilistic number is assigned to each candidate of phrase pair.
- **Bootstrapping approach on the training data:** in this step, by using development set, the threshold for final filtering is determined. This threshold is set by considering high precision low recall filtering heuristic. Then, by using ranked candidates and the intended threshold, the best candidates are extracted and added to the training data.

This research has several contributions as follow:

- In the present task, a novel method to determine of phrase alignment directly from sentence pairs without the direct use of word alignments is proposed.
- In this research work, some filtering rules are defined to eliminate the noisy candidates from all possible phrase alignment candidate
- To score all candidates of a source phrase, four knowledge sources as feature functions are defined.
- We present a framework to score the candidates based on log-linear model. By this model, phrasal features are integrated. This features on log-linear model lead to extracting the best target phrase for the source phrase.
- By using hand aligned phrase table in small size and proposed framework on log-linear model, the size of the training data increases in a bootstrapping manner. Phrase table based on our approach outperforms the SMT results over a phrase table extracted from traditional word level alignments (grow-diag-final-and heuristic).
- In addition to BLEU metric, different evaluations on SMT such as Alignment Error Rate (AER) and TER are reported.

The rest of this paper is organized as follows:

Section 2 briefly describes the related works. Our approach is described in Section 3. The evaluation of the data set statistics and the performed experiments by using our model and the achieved results will be discussed in Section 4. Finally, in Section 5 conclusions are drawn along with future works.

II. PREVIOUS WORKS

As mentioned, the main objective of the present work is to make a phrase table, which has better quality compared to previous heuristic two-stage approach[3] that employs a separate word alignment step.

Some researchers have proposed to learn translation rules directly from sentence pairs without word alignments. Marcu and Wong [6] used a joint probability model for SMT. They illustrated that translations produced by using the joint model were

[3] This approach produces a phrase table.

more accurate than translations produced using IBM model 4.

Cherry and Lin made use of phrasal inversion transduction grammar as an alternative to joint phrasal translation models [7]. They showed that the phrasal translation tables produced by their approach are superior to traditional phrase table. Zhang et al. combined the strengths of Bayesian modeling and synchronous grammar in unsupervised learning of basic translation phrase pairs [8]. They tested their model by using traditional phrase extraction method and find that a phrase table based on their system improves MT results over a phrase table extracted from traditional word level alignments. Simultaneously with Zhang et al. [8], DeNero et al. [9], introduced an approach for estimating phrase pair frequencies. Phrase table weights learned under their model yield an improvement in BLEU score over the word alignment based baseline. Cohn and Blunsom [10] proposed a generative Bayesian model of tree to string translation, which induces grammars that are smaller and produce better translations than the previous heuristic two-stage approach. Neubig et al. [11] used an unsupervised model for joint phrase alignment and extraction using non-parametric Bayesian methods and inversion transduction grammars (ITGs). They demonstrated that the proposed model matches the accuracy of traditional grow-diag-final-and heuristic while reducing the phrase table to a fraction of the original size. Levenberg et al. presented a non-parametric Bayesian model that is able to directly model many-to-many alignments and align discontiguous phrases in both source and target languages [12].

In this research, phrase-alignment candidates are extracted, and then some features are defined that are integrated by the log-liner model to score the candidates. Log-linear models are appropriate to incorporate additional dependencies. Log-linear models have interested to researchers. Previously, Cherry and Lin developed a statistical model to find word alignments [19]. This model, integrates context-specific features easily. Liu et al., presented a log-linear model with some language resources for word alignment in English-Chinese [13]. They used IBM Model 3 alignment probabilities, POS correspondence, and bilingual dictionary coverage as features. In addition, IBM translation model is determined as a base feature. Xiao et al. presented a global log-linear model for synchronous grammar induction that is capable of incorporating arbitrary features [14]. Using learned synchronous grammar rules, they significantly outperform a competitive hierarchical phrase-based baseline system.

It should be mentioned that, approaches based on statistics frequently have infirmity against defects of parallel and specific domain corpora. The idea of bootstrapping on a small size of training data rises from this perspective.

In the past, some of previous works on alignment have used a bootstrapping approach. Yanjun Ma et al. proposed word alignment with bootstrapping approach for word packing in each step [17]. In fact, a simple approach was introduced for statistical alignment of word packing. Pal and Bandyopadhuau have investigated that automatic bootstrapping on the alignment of various chunks makes gains over the previous best English-Bengali probabilistic SMT system [18]. Their focus was on chunk alignment. A chunk is a non-recursive core of an intra-clausal constituent. In this paper, our focus is on phrase alignment. Statistical models [1][2][3] consider the parallel sentence like a sequence of words. A phrase refers to a sequence of words (or sometimes a single word). In 2013, Hien Vu Huy et al. [15] investigated that by bootstrapping WSD models using unlabeled data, they can improve an SMT system. The experiments have been carried out on English-Vietnamese translation and they showed that BLEU scores have been improved significantly.

In our approach, automatic bootstrapping on the alignment of various phrases is used. To scoring, all candidates of phrase alignments, some features are proposed. These features are integrated by log-linear model.

III. Our Approach

In our approach, all possible candidates of phrase pairs are extracted. The main idea of phrase alignment is based on training a log-linear model, which estimates the candidates of phrase alignment probability by using some sophisticated features. This task is started with a few hand-aligned phrases, which are obtained from small number of sentences. Then using corresponding phrases, new phrases are recognized and added to the previous group in a bootstrapping manner. In each iteration, the accuracy of the phrase table is improved.

The system follows four steps:

- In the first step, it generates all possible candidates of phrase alignment for the parallel corpus. Then, some rules are imposed on all candidates to remove a number of them. Candidates of phrase pair are eliminated that are wrongly aligned with high probability. It means that, we use some heuristics for estimating wrongly aligned candidates.
- The second step scores all candidates by some additional knowledge sources as phrasal features in a log-linear model.
- In the third step, the best threshold for filtering all candidates is determined. By the determined threshold, the best candidates are selected to be added into the training corpus in the first iteration.
- Finally, the whole process is repeated in a bootstrapping manner to achieve best phrase level alignments as well as best SMT model.

In the present task, we need to make sure that the extracted phrases do not harm the quality of translation. Therefore, we are more interested in precision than recall. Moreover, we use phrase table in Moses format for ranking each chosen phrase pair in translation task. As a result, our proposed method is error prone.

A. Bootstrapping Approach

The present task is run in a bootstrapping manner. Automatic bootstrapping is carried out on the alignment of various phrases.

At first, a small initial training data is built by hand-aligned data. Initial training data is based on 1100 sentences from the parallel corpus, which contains 48983 phrase pairs. For each phrase pair in training data, count of co-occurrence in parallel corpus is stored. This number not only is used for training based on log-linear translation model in each iteration, but also is considered as a feature in this model. Then, we extract all candidates of phrase alignment from remaining sentences in mentioned parallel corpus. In each iteration of bootstrapping process, each candidate of phrase pair gets a score number based on sophisticated features in a log-linear translation model. Then the threshold is selected by considering high precision low recall heuristic. By using ranked candidates and the intended threshold, the best candidates with their co-occurrence are extracted and added to the training data. An increase of phrase pairs to the training corpus is carried out until the bootstrapping process identifies no new phrase alignment or the quality of translation not to improve significantly.

B. Candidates of Phrase Alignment

In this section, we present a method that produce phrase pair candidates. More specifically, we (1) extract candidates for phrase alignment; (2) estimate the reliability of these candidates and filtering them.

1) Candidate extraction

Statistical models [1][2][3] consider the parallel sentence as a sequence of words. Our approach consists of packing consecutive words together to see whether they correspond to a single phrase. In other words, this bilingually packing of words changes the basic unit of the alignment process. Therefore, the task of automatic alignment is simplified and then the quality of translation is improved.

Candidates of phrase alignment are extracted according to following cases:

- A phrase refers to a sequence of words (or sometimes a single word). According to this definition, all of the possible phrases in each sentence pairs are extracted.
- Moses produces a phrase table with a maximum length of seven words on each side by default. Therefore, we extracted phrase alignment candidates up to length seven words on both sides.
- In the present work, an alignment ***a*** is defined as a subset of the Cartesian product of the phrase position. Candidates can have common words[4]. In our approach, the best candidates from them could be extracted from them.

2) Candidate Reliability Estimation

In this section, the candidates that are more likely to be wrong are estimated. It means that, we extract most probable candidates for phrase alignment.

By studying a phrase table extracted from traditional word level alignments[5] and a phrase table that extracted from hand-aligned words, we found some rules for candidates pruning. As a result, for estimating the candidate reliability we consider the following rules:

(i) Given a phrase pair ***(s, t)*** in which ***s*** and ***t*** refer to the source and the target phrases respectively:

$$s = \{e_1, \dots, e_n\} \qquad 1 \le n \le 7$$
$$t = \{f_1, \dots, f_m\} \qquad 1 \le m \le 7 \quad (1)$$
$$d = |n - m|$$

Here, ***d*** is equal to the difference between the numbers of words in each side of a phrase pair.

We observed that in all of the phrase pairs extracted from hand-aligned words, ***d*** is smaller than five words.

In Moses, there is a length limit of seven words by default for extracted phrases, but in Moses, there is no limit on the difference between the number of words in the source and target phrases.

In the present work, we have extracted all candidates of phrase alignment with the length of one to seven words in each side, but we do not extract phrase pairs which ***d*** is greater than four words. For example, a phrase pair with two words in one side and length of seven words in other side is not extracted.

ii) The phrase table extracted by Moses uses IBM model 3 in two directions of alignment.

Since, our concentration in this paper is on phrase alignment so we consider two translation directions of IBM Model 3 alignment to filter most probable candidates.

We use following ***Definition*** for word alignment [13].

Definition: We have an English sentence $e = e_1^I = e_1, \dots, e_i, \dots, e_I$ and a Persian sentence $f = f_1^J = f_1, \dots, f_j, \dots, f_J$. A link **l= (i, j)** to exist if e_i and f_j are translation (or part of a

[4] Phrases can overlap each other

[5] The phrase table is acquired from Moses.

translation) of each other. If e_i does not correspond to a translation for any Persian word in ***f*** we determine a null link as l= (i, 0). The null link l= (0,j) is defined similarly [13].

In this section, symmetric word alignments are created by running GIZA++ based on IBM model 3, in both directions. For each phrase pair candidate, we consider all of links between words of source phrase and target phrase. In addition, we consider the phrase with the minimum number of words between source phrase and target phrase. A candidate of phrase pair is extracted, if at least half of words for mentioned phrase[6] have link to the some words of the other phrase. By using the symmetric word alignments, the second condition is shown with the following algorithm:

Input: Count, $s = \{e_1, \dots, e_n\}$, $t = \{f_1, \dots, f_m\}$

Output: prob, $s = \{e_1, \dots, e_n\}$, $t = \{f_1, \dots, f_m\}$

1. Start with Count=0.
2. Do for each phrase pairs ***(s,t),*** which is extracted by first condition:
   For each e_i and f_j:
   If e_i and f_j are translation (or part of a translation) in each of direction IBM Model 3, do Count=Count+1;
3. $\text{prob} = \frac{\text{Count}}{\min(n,m)}$
4. If prob ≥ 0.5 then this candidate is extracted.

A candidate is extracted, while it passes both mentioned steps.

C. Scoring Candidates

For scoring all candidates of phrase alignment, we define some knowledge sources as features. To incorporate these features, we propose a log-linear model.

- log-linear model

The source language sentence ***e*** and the target language sentence ***f*** are the essential knowledge sources for the task of finding word alignment and likewise phrase alignment. The use of linguistic information is one of effective approach to improve alignment strategies. The log-linear models can be used as an appropriate approach to incorporating this information.

According to ***Definition***, a subset of the Cartesian product of the word position is determined as an alignment ***a***:

$$a \subseteq \{(i,j): i = 0, \dots, I, j = 0, \dots, J\} \quad (2)$$

By considering the phrase, (sometimes phrase is a single word) as the unit base of alignment instead of words, the above ***Definition*** is extended to phrases. Therefore, phrases can have common words. It means that phrases can have overlap with each other.

We define the phrase alignment problem as finding the alignment ***a*** that maximizes ***Pr(a|s,t)*** [13].

Maximum entropy is considered as a well-founded framework for modeling the probability ***Pr(a|s,t)*** [20].

- *Feature functions*

In this paper, we use bilingual phrase dictionary as the base feature of our log-linear models. In addition, we also make use of additional knowledge sources such as IBM probabilistic dictionary and Google probabilistic dictionary. We also make use of fertility information that extracted from hand-aligned words as a feature.

1) Bilingual Phrase Dictionary

In 2005, Liu et al. used IBM Model 3 as the base feature of their log-linear models [13]. In addition, they used IBM Model 3 as a baseline approach. Thus, we use a part of phrase table that extracted from Moses as base feature of our log-linear model.

A bilingual phrase dictionary is extracted from phrase table. It means that in this dictionary we do not consider features of phrase table[7], while phrase pairs in phrase table are placed in dictionary.

We define a bilingual phrase dictionary as a set of entries: **D**= *{(s,t)}*, *s* is a source language phrase and *t* is a target language phrase in bilingual phrase dictionary. This bilingual phrase dictionary can be considered an additional knowledge source.

For each candidate of phrase dictionary, we check the whether source phrase and target phrase occur in the bilingual phrase dictionary. Therefore, the feature function using a bilingual phrase dictionary (for each phrase pair *(s, t)*) is as follows:

$$h(s,t,D) = \begin{cases} \frac{1}{|1 - \sum_t occur(s,t)|} & if (s,t) occurs\ in\ D \\ 0 & else\ if\ s\ exist\ in\ D, but\ (s,t)\ no\ occur\ in\ D \\ \frac{1}{N} & else\ if\ s\ not\ exist\ in\ D \end{cases} \quad (3)$$

$$occur(s,t,D) = \begin{cases} 1 & if (s,t)\ occurs\ in\ D \\ 0 & else \end{cases} \quad (4)$$

Where ***N*** is a total number of candidates that their source phrases are equal to ***s***. As it is shown in the Eq.3, for each candidate of phrase pair, if source phrase does not exist in bilingual phrase dictionary, this feature treats identically for all of candidates with this source phrase.

Since some of candidates in this feature have no value, so we use Laplace Smoothing for entire candidates. That is to say, $\frac{1}{V}$ is added to feature value of each candidate. Where ***V*** is equal to all of candidates. At the end, Eq.3 is converted as follows:

[6] Phrase with the minimum number of words between source and target phrase.

[7] These features can be considered for experiments on baseline.

$$h(s,t,D)=\begin{cases}\frac{2}{|1-\sum_t occur(s,t)|+v} & if(s,t)occurs\ in\ \boldsymbol{D}\\ \frac{1}{V} & else\ if\ \boldsymbol{s}\ exist\ in\ \boldsymbol{D},but\ (\boldsymbol{s},\boldsymbol{t})\ no\ occur\ in\ \boldsymbol{D}\\ \frac{2}{N+V} & else\ if\ \boldsymbol{s}\ not\ exist\ in\ \boldsymbol{D}\end{cases} \quad (5)$$

2) IBM Probabilistic dictionary

In 1993, Brown et al. proposed a series of statistical models of the translation process [1]. The translation probability $\Pr(f_1^J|e_1^I)$, which explains the relationship between a source language sentence e_1^I and a target language sentence f_1^J is modeling by IBM translation models.

In this paper, second feature for our log-linear model is based on IBM Model 3.

As respects, our focus is on phrase alignment, so we use two translation directions to utilize model 3 as feature function.

To define this feature, first we make a bilingual probabilistic dictionary based on IBM translation model 3. For making this dictionary, we carried out following steps:

1. By running GIZA++ in both directions, IBM Model 3 in two translation directions is extracted.
2. Combining two translation directions of IBM Model 3 as follows:
   For each entry[8] (e,f) in each direction of IBM Model 3:

$$\Pr(e,f)=\begin{cases}\Pr(f|e)*\Pr(e|f) & if(\Pr(f|e)\,\&\,\Pr(e|f))<>0\\ \Pr(f|e) & else\ if\ \Pr(f|e)<>0\\ \Pr(e|f) & else\ if\ \Pr(e|f)<>0\end{cases} \quad (6)$$

We define a bilingual dictionary as a set of entries: ***D= {(e, f, prob)}***, ***e*** is a source language word and ***f*** is a target language word in mentioned bilingual probabilistic dictionary. We normalize the amount of *prob* in this probabilistic dictionary. That is to say, for each source word as *e*, $\sum_{k=1}^{|} \Pr(e,f_k)=1$. This bilingual probabilistic dictionary can be considered as an additional knowledge source.

By considering produced bilingual probabilistic dictionary, we calculate ***n***th root of lexical weighting (or geometric mean of translation probability of each word from source phrase to target phrase ($\sum_{j=1}^{m} p(e_i|f_j)$)) for each candidate of phrase alignment:

For each given a phrase pair ***(s,t)*** where $\boldsymbol{s}=\{e_1,\dots,e_n\}$ is a source phrase and $\boldsymbol{t}=\{f_1,\dots,f_m\}$ is a target phrase, we use an approach that introducing by Vogel et al. [21], to calculate lexical weights based on a statistical lexicon (extracted from mentioned bilingual probabilistic dictionary) for constituent word in the phrase:

$$P_w(t|s)=\prod_{i=1}^{n}\sum_{j=1}^{m} p(f_i|e_j) \quad (7)$$

Where $p(f_i|e_j)$ is word probability estimated using produced probabilistic dictionary from two directions of IBM alignment Models with considering the position alignment.

Therefore, the feature function using a bilingual probabilistic IBM Model 3 dictionary (for each phrase pairs (s, t)) and Eq.7 is:

$$h(a,s,t,D)=\frac{\sqrt[n]{P_w(s|t)}}{\frac{n}{\acute{n}}} \quad (8)$$

As a Section 1, since some of candidates in this feature have no value, so we use Laplace Smoothing for entire candidates. That is to say, $\frac{1}{V}$ is added to feature value of each candidate.

Here, ***a*** is a phrase alignment between source language phrase and target language phrase. ***n*** is a count of words in phrase target and $\acute{n}$ is a count from ***n*** that have at least one link from/to words of source phrase. This division is carried out to considering the count of words in target phrase that have link to/from words of source phrase. For more explanation, two examples are illustrated in Fig. 1 and Fig. 2.

In Fig. 1, a phrase pair and links[9] between their words are indicated. For this phrase pair, ***n*** is equal to 6 and $\acute{n}$ is equal to 5. In addition, in Fig. 2, another phrase pair is indicated. For this phrase pair, ***n*** is equal to 7 and $\acute{n}$ is equal to 5. Clearly, phrase pair in Fig. 1 is better than phrase pair in Fig. 2. Therefore, by this feature, phrase pair in Fig. 1 gets a high score in comparison with phrase pair in Fig. 2.

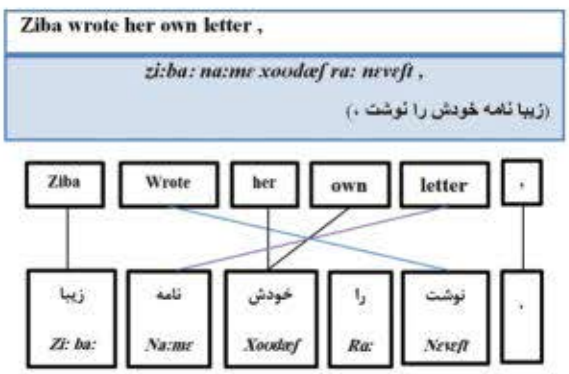

Figure 1. The sample of candidate of phrase alignment in the present task

[8] Word pair

[9] These links are obtained from IBM probabilistic dictionary

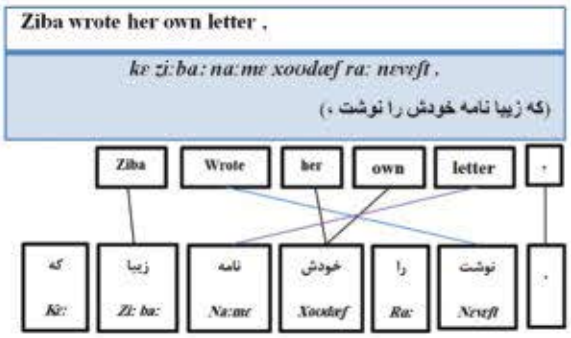

Figure 2. The sample of candidate of phrase alignment in present work

3) Google Probabilistic Dictionary

Despite two proposed features, there are still cases in which local context cannot achieve correctly the meaning of source phrases. One idea is to consider different meanings of words with their probabilities inside each phrase. We can still define some phrasal features as bilingual probabilistic dictionaries to synthesize them for considering local context more than before.

In third feature, we use a probabilistic dictionary, which is extracted from Google translator[10].

Google translator is one of the state of the art commercial machine translation systems, which is used nowadays. By using Google translator, given text from one language to another language is translated. Most of the words in Google translator have different meanings with various probabilities. Since, our purpose is providing phrase table with high quality, so we used two translation directions in Google translator. Therefore, in this feature, we adopt Google translator for English-to-Persian translation and Persian-to-English translation.

We define a probabilistic Google dictionary as a set of entries: **D**= {(***e***, ***f***, prob)}. ***e*** is a source language word. ***f*** is a target language word and ***prob*** is probabilities the alignment from source word to target word.

For making bilingual probabilistic dictionary to use this feature, we carried out following steps as Section 2:

1. All meanings of English and Persian words in mentioned English-Persian parallel corpus are extracted by using Google translator. First, all of English and Persian words are extracted from mentioned parallel corpus. Then by using Google translator all meanings of English and Persian words are extracted.
2. Two translation directions that extracted from Google Translator are combined each other as follows:

For each word pair ***(e,f)*** in each direction of Google Translation:

$$\Pr(e,f) = \begin{cases} \Pr(f|e) \times \Pr(e|f) & if(\Pr(f|e) \,\&\, \Pr(e|f)) <> 0 \\ \Pr(f|e) & else\ if\ \Pr(f|e) <> 0 \\ \Pr(e|f) & else\ if\ \Pr(e|f) <> 0 \end{cases} \quad (9)$$

Finally, we define a bilingual probabilistic dictionary as a set of entries: **D**= *{(e, f, prob)}*, *e* is a source language word and *f* is a target word in language word and ***prob*** is a probability the alignment from source word to target word in mentioned bilingual probabilistic dictionary. Google probabilistic dictionary can be considered as an additional knowledge source.

By considering produced Google probabilistic dictionary that extracted from Google translator and according to Eq.7, the feature function for each phrase pairs (s, t) is:

$$h(a,s,t,D) = \frac{\sqrt[n]{P_w(t|s)}}{\frac{n}{\bar{n}}} \quad (10)$$

We also use Laplace Smoothing for this feature.

4) Fertility

Another feature exploited in our log-linear model is fertility.

The rule of this feature is extracted from hand-aligned phrases. First, we have 2,672 sentences, with the hand-aligned word alignment. Then, we extracted all of phrase-pairs[11] from these sentences by using grow-diag-final-and heuristic. Phrase pairs ***(s,t),*** which are consistent with the word alignment in koehn et al. [4] are as follow:

1. There are links from words of ***s*** to words of ***t***.
2. For every word outside ***s***, there is no link to any word of ***t***.
3. For every word outside ***t***, there is no link to any word of ***s***.

In present work, we extracted statistics of phrase pairs with various lengths in two sides. In Table I, we report the frequencies of the different types of alignments in hand aligned data in English-Persian language. As shown in this table, the most frequent alignment in English-Persian phrase alignment is one word to one word. The second most frequent phrase alignment includes phrase pairs with length of two words on each side and the third frequent phrase alignment is English phrase with length of two words aligned to Persian phrase with length of one word. There are lots of samples for alignment of two English words to one Persian word such as "The word" in English phrase can be aligned to the word of "کلمه/*kælæme*" in Persian phrase.

[10] http://translate.google.com/

[11] Phrase-pairs extracted from word-pairs with grow-diag-final-and heuristics.

By considering Zipf's law and acquired statistics of various alignment types for each phrase pair we reach the following feature function:

$$h(a,s,t,) = \frac{1}{score_{d_s - d_t}} \quad (11)$$

In this feature, Zipf's law has been used for scoring based on statistics of various alignment types for English-Persian phrase pairs. The amount of fertility feature function for the ten different types of most frequent alignments is shown in Table I.

D. The Selection of Best Candidates

In this step, best threshold is determined to select the best phrase pair. For determining this threshold, we use development set. This development set is extracted from PCTS. This data set consist of 772 Persian and English sentences that all of parallel sentences are reviewed and revised. Then GIZA++ is run for extracting word alignment from parallel corpus. The obtained word alignment is revised manually. At the end, all of the phrase alignments are extracted by grow-diag-final-and heuristic approach. Therefore, the gold alignment is made.

By hand-aligned phrase level alignment [12] in development set, we use precision, recall, and AER for determining the best threshold of our approach against gold alignments [5].

To determine the threshold in order to select phrase pairs to add the training data, we use a precision measure. By using development set precision score is calculated. For this purpose, all candidates of phrase alignment are extracted from parallel corpus that is used for development set.

Finally, the threshold with higher precision between all of the intended thresholds as a best threshold is considered. By the determined threshold, the best candidates are selected and then added to training corpus.

TABLE I. Value of fertility feature function for ten of most frequent of alignment type

English-Persian phrase pair	Percentage of various alignment types(Phrase pairs)	h(a,s,t)
1-1	8.9%	1
2-2	4.3%	½
2-1	3.6%	1/3
3-3	2.7%	¼
3-2	2.6%	1/5
1-2	2.5%	1/6
2-3	1.9%	1/7
4-4	1.9%	1/8
4-3	1.8%	1/9
5-5	1.5%	1/10

[12] Sometimes, a phrase refers to a single word.

IV. Evaluation

In order to evaluate our method, we performed a comparison between our method and the baseline for phrase extraction, which is bassed on grow diag-and-final heuristic. First, data sets and tools that are used in experiments are introduced. Then, results of conducted experiments on English-to-Persian translation tasks, are presented. For this comparison several measures are selected and the proposed method is evaluated according to them. Experimental results show the capability of the proposed method in improvement of word and phrase alignments and thereby improvement of translation quality of PB-SMT.

A. Corpora and Tools

The corpus used in our experiments is an English-Persian bilingual corpus collected from several different fields such as News and Novels. This corpus includes approximately 1,090,000 sentence pairs. Since all candidates of phrase pairs make a large set, and so the experiments would not be outperformed in a responsible time, we investigated three parallel corpuses with different sizes including 10,000, 50,000, and 150,000 sentences. Each of them will be used in a different experiment.

Statistics of these corpuses are shown in TABLE II. These corpuses are used in training and the bootstrapping approach.

TABLE II. Statics of training corpus for different size of parallel corpus

	Number of sentences	Number of words	Number of candidates of phrase pair
The parallel corpus that used in Bootstrapping manner for increasing training corpus (150,000 sentences)			
English	151,094	2,290,167	571,134,000
Persian	151,094	2,335,200	
50,000 sentences			
English	50,000	734,148	116,378,000
Persian	50,000	770,429	
10,000 sentences			
English	10,000	163,355	113,689,000
Persian	10,000	163,326	

For evaluating the proposed system we used some metrics include Bilingual Evaluation Understudy (BLEU), Translation Edit Rate (TER), and AER. For this purpose, having two distinct test sets is necessary.

1. The first test set was used for evaluating BLEU and TER scores. This test set, which is called PCTS (Parallel Corpus Test Set), was extracted randomly from the main parallel corpus and were excluded from training set. This corpus includes 393 sentence pairs and is divided into two parts: development set and test set. In Table III statistics of the PCTS corpus is reported. As this table shows, there are four Persian references for each English sentence. Development set has 193 sentences and test set has 200 sentences. We used

the development set in the evaluation of MERT of SMT system. MERT is the process of finding the optimal weights for a log-linear model[13]. These optimal weights are used to maximize the translation quality on our small development set.

2. Another development set and test set, which were used in experiments, are based on PCTS. To determine the best threshold for extracting the best candidates, we need a test set, which is the same format with training corpus. It means that, a hand aligned phrase pairs is needed to determine the most appropriate threshold. This test set is divided into two parts: development set and test set. The test set contains 800 English and Persian sentences and development set consists of 772 sentence pairs. This development set is used for determining the threshold for extracting the best phrase pairs from all candidates, in each iteration. The test set is used to report the value of AER. More details about these development set and test set are reported in TABLE IV. Preparing these sets is summarized as following steps:

   First, all of parallel sentences from PCTS were reviewed and revised manually. Then, word alignments were extracted by Giza++. Next, all of the word alignments were revised manually. At the end, all of the phrase alignments were extracted by grow-diag-final-and heuristic approach from word alignments.

The software used for learning phrase translations, is Moses [22]. We used GIZA++ [23] and SRILM [24] for creating word alignment and learning language models. In addition, the UMIACS word alignment interface[14] has been used to modify word alignment, which is extracted by Giza++. The MaxEnt Grammar tool[15] has been used to perform training and testing in MaxEnt classifier.

TABLE III. Statics on the PCTS data set

	Number of words
Developing corpus	
English	2,139
Reference 1(Persian)	2,212
Reference 2(Persian)	2,310
Reference 3(Persian)	2,203
Reference 4(Persian)	2,215
Testing corpus	
English	2,713
Reference 1(Persian)	2,673
Reference 2(Persian)	3,031
Reference 3(Persian)	2,971
Reference 4(Persian)	2,638

[13] By using this linear model, Moses scores translation hypothesis in during decoding.
[14] Available at http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm
[15] Available at http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool

TABLE IV. Statics for development corpus (PCTS) and testing corpus (PCTS) used for calculating Precision, Recall and AER

	Number of sentences	Number of words	Number of phrase pairs (hand aligned)
Development set			
English	772	8,716	23,279
Persian	772	8,749	
Test set			
English	800	11,031	20,392
Persian	800	10,833	

B. *Experiments and results*

Metrics used for evaluating the proposed system include BLEU, TER, and AER. In the following, we present the results of log-linear model in a bootstrapping manner for phrase alignment. We used Moses package to train IBM translation models. The base feature of our log-linear model, bilingual phrase dictionary, takes the phrase alignment generated by Moses as entry of generated dictionary. The second feature of our log-linear models, derivation of IBM model 3, takes the parameters generated by GIZA++ as parameters for itself. The third feature is based on Google translator and the fourth feature is based on fertility each phrase pair.

In the first iteration, there is hand-aligned phrase table that is used for training set. The best candidates of phrase pairs are added to this phrase table iteratively. We took the following steps to prepare this data set:

- All of parallel sentences are reviewed and revised manually.
- Word alignments are extracted by Giza++.
- All of the word alignments are revised manually.
- All of the phrase alignments are extracted by using grow-diag-final-and heuristic approach.

Fig. 3 shows the results of various features of our log-linear model in the first iteration. As shown in this figure, adding each proposed feature improves the BLEU results. In other words, adding each proposed feature to the model leads to increase the scores of the best phrase pairs.

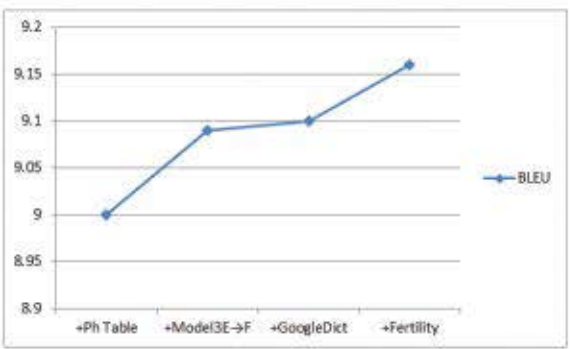

Figure 3. Results of BLEU using log-linear models (feature selection) in the first iteration (150,000 sentences)

In all experiments, we used the phrase table that is extracted from Moses as the baseline: this phrase table is based on some features such as phrase translation probabilities and lexical smoothing of phrases in both directions [25].

As mentioned in the previous section, we selected three sets of our parallel corpus including 10,000, 50,000, and 150,000 sentences respectively.

In the proposed approach, a threshold is defined in each iteration. By this threshold, all candidates that their score is above the threshold are added to the training phrase table. To determine the threshold, second development set[16] is used. It is need to make sure that the extracted phrases do not harm the quality of translation. So to determine the threshold, precision is much more important than the recall. Fig. 4 shows values of precision, recall, and AER for different thresholds in parallel corpus with 150,000 sentences. In this paper, a confidence measure as a threshold is determined according to Precision, Recall, and AER. Since the precision is more important than recall in alignment task, we intent to select a threshold, which has highest precision among the all threshold. It is clear that if a threshold value is higher, the precision value will be increased and the recall value will be decreased. In terms of our strategy, we select the threshold with highest and most acceptable precision. That is to say, the threshold is not selected, which its recall is close to zero. Accordingly, in our approach, the appropriate threshold equals to 0.5. So, this threshold is selected to extract the best candidates.

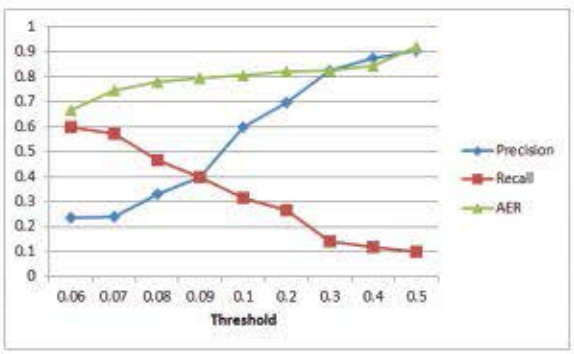

Figure 4. Precision, Recall and AER for different thresholds in first iteration

In parallel corpus with 150,000 sentences, there are about 511 million candidates for phrase alignment, which are extracted from the parallel corpus. A number of these candidates have been aligned by our system in each iteration. In Fig. 5, size of the generated phrase table in each iteration is reported. As shown in this figure, in the end of final iteration, we have a phrase table with 3,112,953 entries, while the phrase table extracted by Moses has 6,605,196 entries. As a result, the phrase table generated by our approach is smaller than a phrase table that is extracted by Moses. Meanwhile, our phrase table leads to a better translation compared to the phrase table extracted by the Moses.

To calculate the BLEU and TER scores, we transformed all of chosen phrase pairs in phrase table to the Moses Format. In other words, to evaluate reliability of each phrase pairs to be used in the translation task, different features have been considered in this phrase table, which has been produced by new Moses format. It can be observed in Fig. 6 that the baseline system produces a BLEU score of 10.09. Also Fig. 6 shows the values of BLEU measure in each loop of bootstrap method. Hence, experiments with 150,000 sentences are converged in 18th iteration. By adding selected phrase pairs to hand-aligned phrase table in each iteration, the quality of statistical machine translation is improved. As shown in this figure, the BLEU score of the proposed method in 150,000 parallel sentences achieves 14.26.

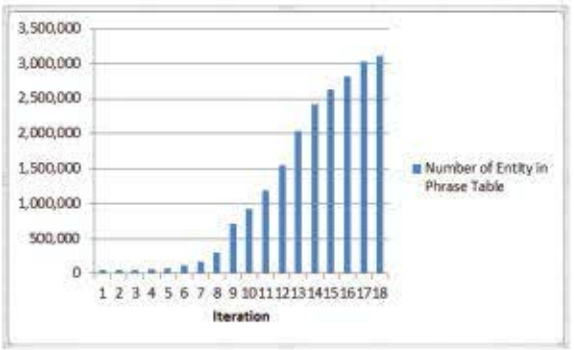

Figure 5. Number of phrase table entries (Unique Phrase Pairs) for 150,000 sentences

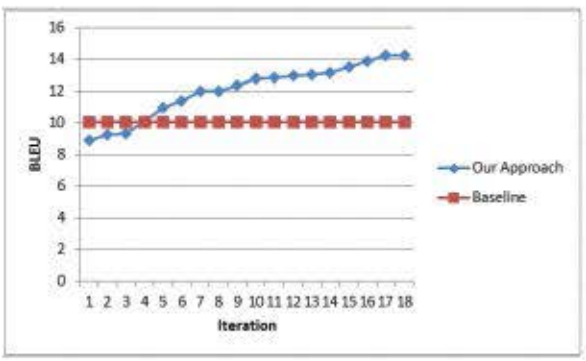

Figure 6. Evaluation results of BLEU for 150,000 sentences in each iteration

Moreover, as indicated from Fig. 7, SMT system[17] leads to higher translation quality with reductions by 3.59 in TER in comparison with Baseline SMT system. Fig. 7 shows that the value of TER in baseline equals about 71 while in final loop of bootstrapping method, equals about 67. As indicated from the Fig. 6 and Fig. 7, that SMT system utilizes our proposed approach leads to the high translation quality with growths by 4.17 in BLEU score and decreases by 3.59 in TER score in comparison with baseline SMT system.

However, the baseline approach of IBM alignments is unsupervised, while our approach is

[16] This development set described in previous section.

[17] This SMT system uses 50K sentences.

semi-supervised relying on hand-aligned bilingual corpus and bootstrapping approach.

To verify experiments, two other sizes of parallel corpus are employed. Fig. 8 and Fig. 9 report the BLEU score and TER score for corpus with 50,000 sentences respectively. As indicated from these figures, the SMT system, which utilizes bootstrapping approach with proposed features, leads to higher translation quality with growths by 3.4 in BLEU score and reductions by 1.84 in TER score in comparison with the Baseline SMT system.

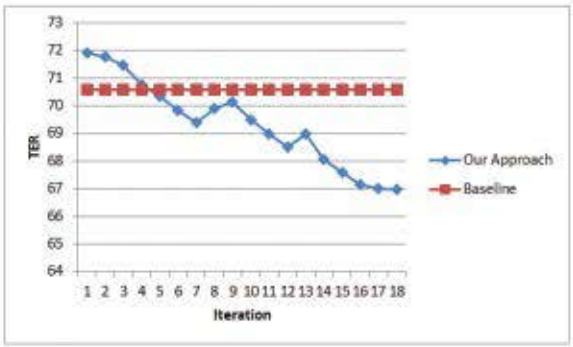

Figure 7. Evaluation results of TER for 150,000 sentences in each iteration

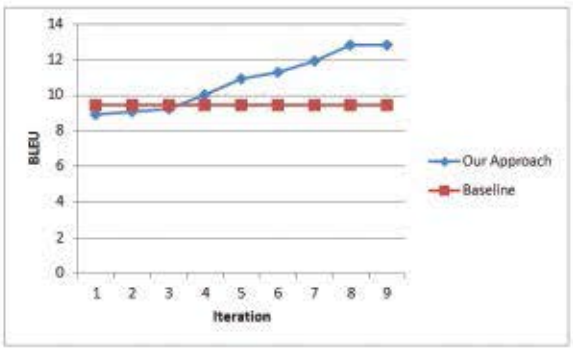

Figure 8. Evaluation results of BLEU for 50,000 sentences in each iteration

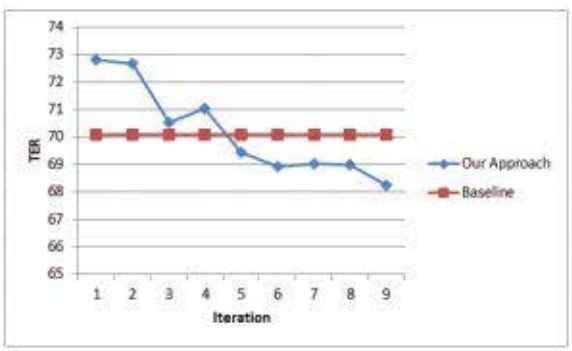

Figure 9. Evaluation results of TER for 50,000 sentences in each iteration

To calculate the AER in each iteration, the second test set has been used. In TABLE V the results of AER for two different sizes of parallel corpus is presented. As shown in TABLE V, AER improves in each iteration. In first, second and third iterations maximum precision and minimum recall is obtained. It means that, a few numbers of phrase pairs are aligned by the proposed system in the test set and almost all of phrase pairs exist in the gold data set. While in the fourth iteration, a significant number of phrase pairs are aligned by the proposed system. So, lower precision and higher recall are obtained relative to the previous iterations. As a result, the value of AER decreases. According to this table, experiments with 50,000 sentences coverage and stop in ninth iteration.

At the next experiment, we used the parallel corpus with the 10,000 sentences for experiments. The value of BLEU for the baseline of SMT system with 10,000 sentences and 650,000 entries in phrase table equals to 7.77. On the other hand, the value of BLEU for hand-aligned phrase table before using proposed approach with 48,000 entries equals to 8.91. So, we do not apply our approach for parallel corpus with 10,000 sentences.

TABLE V. Evaluation results of AER in each iteration

	Size of Training Corpus	
Iteration	**50 K**	**150 K**
1	0.96	0.95
2	0.96	0.95
3	0.96	0.95
4	0.89	0.88
5	0.86	0.84
6	0.82	0.81
7	0.77	0.77
8	0.73	0.72
9	0.72	0.71
10	----------	0.71
11	----------	0.70
12	----------	0.68
13	----------	0.68
14	----------	0.66
15	----------	0.60
16	----------	0.59
17	----------	0.58
Final Iteration	----------	0.58

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we indicated a considerable effect of bootstrapping method with some knowledge sources as features on translation quality of PB-SMT. Statistical approaches are sensitive to the quality of the parallel training data. Therefore, we try to alleviate this problem with bootstrapping approach on training data. The analyses and results on experiments point out that the proposed approach of extracting phrase alignment directly from sentence pairs without using of word alignments contributes to the improvement of translation quality. In our experiments have been used two parallel corpora with lengths of 50,000 and 150,000 sentence pairs. The best system yields 4.17 BLEU points and 3.59 TER points improvement over the baseline. In addition to, AER has been improved iteratively in each loop.

The proposed bootstrapping method for aligning phrases has some pitfalls. The number of candidates for phrase alignment is too many and a few numbers of them[18] are extracted and added to training data. So, a large number of loops are run for convergence of this approach in large parallel corpus. As a result, this approach is time-consuming for large parallel corpus.

[18] Because of the importance of Precision

Accordingly, in order to recognize incorrect candidates and remove them, so we intend to add the number of conditional filtering candidate, which was described in section B. Due to low quality of available parallel corpus and weakness of features for some candidates, a few of incorrect candidates are inserted to training data. So, error cascade is happened and affected on training weights of features in next loop. In the other words, this problem has been published in the next loops of bootstrap method. In the future, we would like to continue experiments with the expansion of the number of features. We would like to add POS Tags transition model, inverse of IBM probabilistic dictionary and inverse of Google probabilistic dictionary as features to enhance the quality of translation. In addition, we would like to improve reordering by using produced phrase table.

REFERENCES

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," Association for Computational Linguistics, vol. 19, No. 2, 1993, pp. 263-311.

[2] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Translation," In Proceedings of COLING, Copenhagen, Denmark, 1996, pp. 836-841.

[3] Y. Deng and W. Byrne, "HMM Word and Phrase Alignment for Statistical Machine Translation," In Proceedings of HLT-EMNLP, Vancouver, Canada, 2005, pp. 169-176.

[4] P. Koehn, F. Och, and D. Marcu, "Statistical Phrase-based Translation," In Proceedings of HLT-NAACL, Edmonton, Canada, 2003, pp. 48-54.

[5] F. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," Computational Linguistics, Vol. 29, No. 1, 2003, pp. 19-51.

[6] D. Marcu and W. Wong, "A Phrase-based Joint Probability Model for Statistical Machine Translation," In Proceedings of EMNLP, Morristown, NJ, 2002, pp. 133-139.

[7] C. Cherry and D. Lin, "Inversion Transduction Grammar for Joint Phrasal Translation Modeling," In Proceedings of SSST, NAACL-HLT Workshop on Syntax and Structure in Statistical Translation, 2007, pp. 17-24.

[8] H. Zhang, C. Quirk, R. C. Moore, and D. Gildea, "Bayesian Learning of Non-compositional Phrases with Synchronous Parsing," In Proceedings of ACL, 2008, pp. 97-105.

[9] J. DeNero, A. Bouchard-Cote, and D. Klein, "Sampling Alignment Structure under a Bayesian Translation Model," In Proceedings of Empirical Methods in Natural Language Processing, 2008, pp. 314-323.

[10] T. Cohn, and P. Blunsom, "A Bayesian Model of Syntax-directed Tree to String Grammar Induction," In Proceedings of Empirical Methods in Natural Language Processing, 2009, pp. 352-361.

[11] G. Neubig, T. Watanabe, E. Sumita, S. Mori and T. Kawahara, "An Unsupervised Model for Joint Phrase Alignment and Extraction," In Proceedings of ACL, 2011, pp. 632-641.

[12] A. Levenberg, C. Dyer, and P. Blunsom, "A Bayesian Model for Learning SCFGs with Discontiguous Rules," In Proceedings of Empirical Methods in Natural Language Processing, Association for Computer Linguistic, 2012, pp. 223-232.

[13] Y. Liu, Q. Liu, and S. Lin, "Log-linear Models for Word Alignment," In proceedings of the 43rd Annual Meeting of the ACL, 2005, pp. 459-466.

[14] X. Xiao, D. Xiong, Y. Liu, Q. Liu and S. Lin, "Unsupervised Discriminative Induction of Synchronous Grammar for Machine Translation," In Proceedings of COLING, Mumbai, 2012, pp. 2883-2898.

[15] H. V. Huy, P-T Nguyen, T-L Nguyen, and M.L Nguyen, "Bootstrapping Phrase-based Statistical Machine Translation via WSD Integration," International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013, pp. 1042-1046.

[16] K. Singh, and M. Xie, "Bootstrap: A Statistical Method," Unpublished Working Paper, Rutgers University, 2008.

[17] Y. Ma, S. Nicolas, and A. Way, "Bootstrapping Word Alignment Via Word Packing," In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007, pp. 304-311.

[18] S. Pal and S. Bandyopadhyay, "Bootstrapping Method for Chunk Alignment in phrase Based SMT," In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 2012, pp. 93-100.

[19] C. Cherry and D. Lin, "A probability model to improve word alignment," In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistic(ACL), Sapporo, Japan, 2003, pp. 88-95.

[20] A. L. Berger, S. A. Della Pietra, and V. J. DellaPietra, "A maximum entropy approach to natural language processing," Computational Linguistics, vol. 22, pp. 39-72, 1996.

[21] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A.Venogupal, B.Zhao, and A. Waibel, "The CMU Statistical Machine Translation System," In Proceedings of MT-Summit IX. New Orleans, LA, 2003.

[22] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moren, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbbst, "Moses: Open Source Toolkit for Statistical Machine Translation," In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 2007, pp. 177-180.

[23] F. J. Och, and H. Ney, "Improved Statistical Alignment Models," In Proceedings of ACL, pp. 440-447, 2000.

[24] A. Stolcke, "SRILM-An extensible language modeling toolkit," In Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado, 2002, pp. 901-904.

[25] R. Zens, and H. Ney, "Improvements in phrase-based statistical machine translation," In Proceedings of HLT-NAACL, Boston, MA, 2004, pp. 257-264.

APPENDIX A: INTERNATIONAL PHONETIC ALPHABET FOR PERSIAN

IPA	Letter(s)	Examples
ɑ:	ا	part, father
Æ	ا	bat, pad
ɛ	ا ٔ	bed, fell
B	ب	bee, beet
P	پ	pen, spoon
T	ت	stick, tall
Θ	ث	thigh, math

dʒ	جج	giant, jazz
tʃ	چچ	Chip, catch
H	حح	Hat
X	خخ	ugh, loch
D	د	done, den
Z	ذ	Jazz
R	ر	dark, try
Z	ز	thus, bazaar
Zh	ژ	Journal
S	سس	sock, school
ʃ	شش	shah, cash
S	صص	Massage
Z	ضض	Dark
T	ط	Star
Z	ظ	thus, bazaar
E	ع	(No equivalent)
ɣ˜ʁ	غ	French R
F	ف	fast, phl
Q	ق	Scar
K	ک	sky, crack
G	گ	good, bag
L	ل	bell, sleep
M	م	me, man
N	ن	can, no
V	و	verb, we
H	ه ه ه	help, ahead
ɪ	ی	fill, bin
i:	ی	fell, sea
ɑɪ	آی	fine, pie
oʊ	او	foal, bone
ʊ	او	foot, good
u:	اوو	boot, chew

Leila Tavakoli received her B.Sc. degree in software engineering from South Tehran Branch of Islamic Azad University at 2010. Since 2011, she is a M.Sc. student at Tehran University in the field of natural language processing. Her research areas include computational natural language processing, machine translation, pattern recognition, and bilingual alignment.

Hesham Faili received his B.Sc. and M.Sc. in software engineering and his Ph.D. in artificial intelligence from Sharif University of Technology. He is an assistant professor at Tehran University in the School of Electrical and Computer Engineering. His research interests include natural language processing, machine translation, data mining, and social networks.

IJICTR International Journal of Information & Communication Technology Research

Volume 6- Number 3-Summer 2014

Preparation of Papers for IJICTR

Paper Title (use style: *paper title*)

Authors Name/s per 1st Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: e-mail address if desired

Authors Name/s per 2nd Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: e-mail address if desired

***Abstract*—This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document. (Abstract should not be longer than 150 words).**

Keywords-component; formatting; style; styling; insert (Include 5 to 10 words .)

I. INTRODUCTION (HEADING 1)

This template, modified in MS Word 2003 and saved as "Word 97-2003 & 6.0/95 – RTF" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow margins.

(Top:20mm.Bottom:20mm,Left:25mm,Right:25mm)

II. EASE OF USE

Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file for "MSW_USltr_format".

III. PREPARE YOUR PAPER BEFORE STYLING

Space between top of the page and title of the paper has to be 85mm wide with the title centered. Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. *Abbreviations and Acronyms*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. *Units*

- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter", not "webers/m2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".

- Use a zero before decimal points: "0.25", not ".25". Use "cm3", not "cc". (bullet list)

C. *Equations*

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation.

IV. USING THE TEMPLATE

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. *Identify the Headings*

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

B. *Figures and Tables*

Place figures and tables at the top and bottom of columns. Avoid. placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

ACKNOWLEDGMENT (HEADING 5)

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression, "One of us (R. B. G.) thanks . . ." Instead, try "R. B. G. thanks". Put sponsor acknowledgments in the unnum-bered footnote on the first page.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

Note: Please submit your proposals via IJICT website at: http://jounal.itrc.ac.ir