

Phrase-Based Statistical Machine Translation: A Level of Detail Approach

Hendra Setiawan^{1,2}, Haizhou Li¹, Min Zhang¹, and Beng Chin Ooi²

¹ Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{stuhs,hli,mzhang}@i2r.a-star.edu.sg

² School of Computing
National University of Singapore
Singapore 117543
{hendrase,ooibc}@comp.nus.edu.sg

Abstract. The merit of phrase-based statistical machine translation is often reduced by the complexity to construct it. In this paper, we address some issues in phrase-based statistical machine translation, namely: the size of the phrase translation table, the use of underlying translation model probability and the length of the phrase unit. We present Level-Of-Detail (LOD) approach, an agglomerative approach for learning phrase-level alignment. Our experiments show that LOD approach significantly improves the performance of the word-based approach. LOD demonstrates a clear advantage that the phrase translation table grows only sub-linearly over the maximum phrase length, while having a performance comparable to those of other phrase-based approaches.

1 Introduction

Early approach to statistical machine translation relies on the word-based translation model to describe the translation process [1]. However, the underlying assumption of word-to-word translation often fails to capture all properties of the language, i.e. the existence of the phrase where a group of words often function together as a unit. Many researchers have proposed to move from the word-based to the phrase-based translation model [2] [3] [4]. A phrase-based approach offers many advantages as a phrase translation captures word context and local re-ordering inherently [3]. It has become popular in statistical machine translation applications.

There are typically two groups of approaches to constructing the phrase-based model. The first group learns phrase translation directly from the sentence pair. It learns both word and phrase units simultaneously. Although these approaches appear intuitive, it usually suffers from a prohibitive computational cost. It might have to consider all possible multi-word sequences as phrase candidates and all possible pairings as phrase translations at the same time.

The second group of approaches learns phrase translations through word-level alignment: alignment template [2] and projection extension [6], just to name a few. In general, these approaches take the word-level alignment, a by-product of the word-based translation model, as their input and then utilize a heuristic measurement to learn the phrase translation. The heuristic measurement contains all possible configurations of word-level alignment on a phrase translation.

It is noted that the underlying word-level alignment is just an approximation to the exact alignment. The approximation is reflected by a probability produced by the word-based translation model. The majority of approaches do not make use of this probability, whereas it may provide a valuable clue leading to a better phrase translation from a statistical point of view. Koehn, et. al [8] compared the representative of both groups and reported that learning phrase translation using a simple heuristic from word alignment yields a better translation performance than learning phrase translation directly from the sentence pair.

Many approaches try to learn all phrase translations in one step, either directly from the sentence pair or through word alignment. As a result, they may encounter a huge amount of phrase translation candidates at once. Usually, they limit the maximum phrase length to reduce the choice of candidates. Although this method is sufficient to satisfy the computational requirement, it comes with the cost of not finding the good phrases longer than the imposed limit. Additionally, to reduce the candidates, those approaches use a threshold to separate good phrase translation from the rest. The threshold is ad-hoc and often not capable of making a clear separation. Therefore, the use of threshold often comes with the cost of the inclusion of undesired phrase translations and the absence of good phrase translations in the phrase translation table. The cost may be reflected from the size of the phrase translation table that often grows almost linearly over the phrase length limit [6][8]. The growth implies a non-intuitive behavior: two phrases with different length introduce an equal number of additional entries to the phrase translation table. As longer phrases occur less often, there should be fewer entries introduced into the phrase translation table.

We propose an agglomerative approach to learn phrase translations. Our approach is motivated by the second group, which is to learn phrase translation through word-alignment, while addressing the common issues: the size of the phrase translation table, the use of underlying translation model probability and the length of the phrase unit.

Only a few approaches move away from one-step learning. Melamed [13] presented an agglomerative approach to learn the phrases progressively from a parallel corpus by using sub-phrase bigram statistics. Moore [14] proposed a similar approach which identifies the phrase candidates by parsing the raw training data. Our idea differs from these approaches in that we look into the association of the alignments rather than the association of the words to discover the phrases.

In this paper, we propose the Level of Detail (LOD) approach for learning of phrase translations in phrase-based statistical machine translation. Section 2 discusses the background and motivation and then formulates the LOD approach

while section 3 describes the learning process in details. Section 4 describes the experimental results. In this section, we compare LOD with state-of-the-art word-based approach in translation tasks. Finally, section 5 concludes this paper by providing some discussion in comparison with other related works.

2 Statistical Machine Translation: A Level of Detail

2.1 Motivation and Background

It is often not intuitive to model the translation of a phrase using the word-based translation model. First, the literal translation of phrase constituents is often inappropriate from a linguistic point of view. The word-based translation model treats a phrase as a multi-word. One such example is the case where a phrase appears as an idiom. The translation of an idiom cannot be synthesized from the literal translation of its constituents but rather from the semantic translation of the whole. Besides, the literal translation of an idiom detracts from the intended meaning. In one such example, the literal translation of French "*manger sur le pouce*" is "*to eat on the thumb*". This detracts from the correct translation "*to grab a bite to eat*". In addition, to produce the correct translation, the word-based translation model might have to learn that "*manger*" is translated as "*eat*" or "*pouce*" is translated as "*thumb*". Although it may serve the translation purpose, it will introduce many non-intuitive entries to the dictionary.

Second, even if it is possible to translate a phrase verbatim, modeling phrase translation using the word-based translation model suffers from a disadvantage: the number of word alignments required to synthesize the phrase translation is large. It requires four word alignments to model the translation between "*une minute de silence*" and "*one minute of silence*", whereas one phrase alignment is adequate. The introduction of more alignments also implies the requirement to estimate more parameters for the translation model. The implication often comes with the cost of learning wrong word alignments.

Third, a phrase often constitutes some spurious words. The word-based translation model often has trouble in modeling spurious words, such as function words. Function words may appear freely in any position and often may not be translated to any word. We observe that many of these function words appear inside a phrase. It is beneficial to realize these spurious words inside a phrase unit so as to improve statistical machine translation performance and also to remove the necessity to model them explicitly. All these suggest that, ideally, a phrase translation should be realized as a phrase alignment, where the lexical correspondence is established on phrase level rather than on its word constituents.

The discussion above suggests that phrase-based translation is a wise choice. Practically, as a phrase is not a well defined lexical entry, a mechanism is needed to judge what constitutes a phrase in the context of statistical machine translation. In this paper, we advocate an approach to look into the phrase discovery process at different level of details. The level of detail refers to the size of a

phrase unit. At its finest level of detail, a phrase translation uses the word-based translation model where a phrase is modeled through its word constituent. At a coarser level of detail, a sub-phrase unit is introduced as a sequence of words, making it a constituent of the phrase. The coarsest level of detail refers to the status of a phrase where all word constituents converge into a whole unit.

Our Level-Of-Detail (LOD) approach views the problem of phrase-based translation modeling through a LOD process. It starts from the finest word-level alignment and transforms the phrase translation into its coarsest level of detail.

2.2 Formulation

Let $\langle \mathbf{e}, \mathbf{f} \rangle$ be a sentence pair of two sequences of words with \mathbf{e} as an English sentence and \mathbf{f} as its translation in French³. Let $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle$ represents the same sentence pair but with the phrase as its atomic unit rather than the word. To generalize the notation, we treat word and phrase unit similarly by considering a word as a phrase of length one. Therefore, $\langle \mathbf{e}, \mathbf{f} \rangle$ hereafter will be referred as $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(0)}$, which represents the finest level of detail, and $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle$ as $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(N)}$, which represents the coarsest level of detail. Let each tuple in the sentence pair of any level of detail n , $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$ be $\tilde{\mathbf{e}}^{(n)} = \{\tilde{e}_0^{(n)}, \tilde{e}_1^{(n)}, \dots, \tilde{e}_i^{(n)}, \dots, \tilde{e}_{l^{(n)}}^{(n)}\}$ and $\tilde{\mathbf{f}}^{(n)} = \{\tilde{f}_0^{(n)}, \tilde{f}_1^{(n)}, \dots, \tilde{f}_j^{(n)}, \dots, \tilde{f}_{m^{(n)}}^{(n)}\}$ where $\tilde{e}_0^{(n)}, \tilde{f}_0^{(n)}$ represent the special token *NULL* as suggested in [1] and $l^{(n)}, m^{(n)}$ represent the length of the corresponding sentence. Let $T^{(n)}$ be a set of alignment defined over the sentence pair $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$ with $t_{ij}^{(n)} = [\tilde{e}_i^{(n)}, \tilde{f}_j^{(n)}]$ as its member. The superscript in all notations denotes the level of detail where 0 represents the finest and N represents the coarsest level of detail.

LOD algorithm iteratively transforms $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(0)}$ to $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(N)}$ through re-alignment of phrases and re-estimation of phrase translation probability. At n -th iteration, LOD harvests all bi-directional alignments from the sentence pair $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$. The alignment is obtained by a typical word-based translation model, such as the IBM model, while treating a sub-phrase at n -th iteration as a word. We refer to those alignments as $\mathcal{B}^{(n)}$, a pool of sub-phrase alignments unique to the particular iteration. Afterwards, LOD generates all possible phrase alignment candidates $\mathcal{C}^{(n)}$ for a coarser level of detail from these sub-phrase alignments. A resulting phrase alignment candidate is basically a joining of two adjacent sub-phrase alignments subject to a certain criterion. It represents the future coarser level alignment. Up to this point, two sets of alignment are obtained over $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$: a pool of sub-phrase alignments $\mathcal{B}^{(n)}$ at the current level and a pool of phrase alignment candidates $\mathcal{C}^{(n)}$ at a coarser level. From these two sets of alignments $\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)}$, we would like to derive a new set of alignments $T^{(n+1)}$ that best describes the training corpus with the re-estimated statistics obtained at n -th iteration. LOD constructs $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n+1)}$ from the new set of alignment. Algorithm 1 provides the general overview of LOD algorithm.

³ Subsequently, we will refer \mathbf{e} as source sentence and \mathbf{f} as target sentence, but the term does not always reflect the translation direction

Algorithm 1. An overview of LOD approach in learning phrase translation. The LOD approach takes a sentence pair at its finest level of detail as its input, learns the phrase-level alignment iteratively and outputs the same sentence pair at its coarsest level of detail along with its phrase translation table.

input $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(0)}$
for $n = 0$ **to** $(N - 1)$ **do**
 - Generate bi-directional sub-phrase level alignments $\mathcal{B}^{(n)}$ from $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$
 - Identify phrase-level alignment candidates $\mathcal{C}^{(n)}$ from $\mathcal{B}^{(n)}$
 - Estimate the alignment probability in $\mathcal{B}^{(n)}$ and $\mathcal{C}^{(n)}$
 - Learn coarser level alignment $T^{(n+1)}$ from $\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)}$ and construct $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n+1)}$
output $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(N)}$ and $T^{(N)}$

3 Learning Phrase Translation

In this section, we discuss the steps of LOD algorithm in detail. As presented in Algorithm 1, moving from one level of alignment to its coarser level, LOD follows four simple steps:

1. Generation of bi-directional sub-phrase level alignments ⁴
2. Identification of phrase level alignment candidates
3. Estimation of alignment probability
4. Learning coarser level alignment

3.1 Generation of Bi-directional Sub-phrase Level Alignments

LOD follows the common practice to utilize the IBM translation model for learning the phrase translation. That is to harvest all alignments from both translation directions. For the sake of clarity, LOD defines the following notation for these alignments, as follows:

Let $\Gamma_{ef}^{(n)} : \tilde{e}_i^{(n)} \longrightarrow \tilde{f}_j^{(n)}$ be an alignment function represents all alignments from translating the source English sentence to the target French sentence, and $\Gamma_{fe}^{(n)} : \tilde{f}_j^{(n)} \longrightarrow \tilde{e}_i^{(n)}$ be the reversed translation direction. Then, bi-directional sub-phrase alignment $\mathcal{B}^{(n)}$ includes all possible alignment by both functions:

$$\mathcal{B}^{(n)} = \{t_{ij}^{(n)} = [\tilde{e}_i^{(n)}, \tilde{f}_j^{(n)}] | (\Gamma_{ef}^{(n)}(\tilde{e}_i^{(n)}) = \tilde{f}_j^{(n)}) \cup (\Gamma_{fe}^{(n)}(\tilde{f}_j^{(n)}) = \tilde{e}_i^{(n)})\}$$

Let us denote *NULL* alignments, $\mathcal{N}^{(n)}$, a subset of alignments in $\mathcal{B}^{(n)}$ in which the special token *NULL* is involved.

⁴ The process starts with word level alignment. A word here is also referred to as a sub-phrase.

3.2 Identification of Phrase Alignment Candidates

LOD applies a simple heuristic to identify a phrase alignment candidate. First, LOD considers every combination of two distinct sub-phrase alignments and assesses its candidacy. Here, we define a phrase alignment candidate $\langle t_{ij}^{(n)}, t_{i'j'}^{(n)} \rangle \in \mathcal{C}^{(n)}$ as follows:

Let $\langle t_{ij}^{(n)}, t_{i'j'}^{(n)} \rangle$ be a set of two tuples, where $t_{ij}^{(n)} \in \mathcal{B}^{(n)}$ and $t_{i'j'}^{(n)} \in \mathcal{B}^{(n)}$. Then $\langle t_{ij}^{(n)}, t_{i'j'}^{(n)} \rangle$ is a phrase alignment candidate **if and only if**

1. **not** $((i, i') \neq 0)$ **or** $(|i - i'| = 1)$
2. **not** $((t_{ij}^{(n)} \in \mathcal{N}^{(n)})$ **and** $(t_{i'j'}^{(n)} \in \mathcal{N}^{(n)}))$

In the definition above, the first clause defines a candidate as a set of two whose source sub-phrases are adjacent. The second clause forbids the consideration of two *NULL* alignments.

As LOD considers only two alignments for each phrase alignment candidate, it implies that, at the n -th iteration, the length of the longest possible phrase is bounded by 2^n . Apparently, we do not have to examine sub-phrase alignment trunks of more than two sub-phrases because the iteration process guarantees LOD to explore phrases of any length given sufficient iteration. This way, the search space at each iteration can be manageable at each iteration.

3.3 Estimation of Alignment Probability

Joining the alignment set $\mathcal{B}^{(n)}$ derived in Section 3.1 and the coarser level alignment $\mathcal{C}^{(n)}$ derived in Section 3.2, we form a candidate alignment set $\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)}$. Assuming that there are two alignments $x \in \mathcal{B}^{(n)}$, $y \in \mathcal{B}^{(n)}$, and a candidate alignment $\langle x, y \rangle \in \mathcal{C}^{(n)}$, we derive the probability $p(x)$ and $p(y)$ from the statistics as the count of x and y normalized by the number of alignments in the corpus, and we derive the joint probability $p(\langle x, y \rangle)$ in a similar way.

If there is a genuine association between the two alignments, x and y , then we expect that $p(\langle x, y \rangle) \gg p(x)p(y)$. If there is no interesting relationship between x and y , then $p(\langle x, y \rangle) \approx p(x)p(y)$ where we say that x and y are independent. If x and y are in a complementary relationship, then we expect to see that $p(\langle x, y \rangle) \ll p(x)p(y)$. These statistics allow us to discover a genuine sub-phrase association.

The probability is estimated by the count of observed events normalized by the corpus size. Note that the alignment from the IBM translation model is derived using a Viterbi-like decoding scheme. Each observed event is counted as one. This is referred to as hard-counting. As the alignment is done according to probability distribution, another way of counting the event is to use the fractional count that can be derived from the translation model. We refer to it as soft-counting.

3.4 Learning a Coarser Level Alignment

From section 3.1 to 3.3, we have prepared all the necessary alignments with their probability estimates. The next step is to re-align $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$ into $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n+1)}$ using alignment phrases in $\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)}$ with their newly estimated probability distribution. The re-alignment is considered as a constrained search process. Let $p(t_{ij}^{(n)})$ be the probability of a phrase alignment $t_{ij}^{(n)} \in (\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)})$ as defined in Section 3.3, $T^{(n)}$ be the potential new alignment sequence for $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$, we have the likelihood for $T^{(n)}$ as

$$\log P(\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)} | T^{(n)}) = \sum_{t_{ij}^{(n)} \in T^{(n)}} \log p(t_{ij}^{(n)}) \quad (1)$$

The constrained search is to decode an alignment sequence that produces the highest likelihood possible in the current iteration, subject to the following constraints:

1. to preserve the phrase ordering of the source and target languages
2. to preserve the completeness of word or phrase coverage in the sentence pair
3. to ensure the mutual exclusion between alignments (except for the special *NULL* tokens)

The constrained search can be formulated as follows:

$$T^{(n+1)} = \underset{\forall T^{(n)}}{\operatorname{argmax}} \log P(\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)} | T^{(n)}) \quad (2)$$

In Eq.(2), we have $T^{(n+1)}$ as the best alignment sequence to re-align sentence pair $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$ to $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n+1)}$.

The constraints are to ensure that the search leads to a valid alignment result. The search is essentially a decoding process, which traverses the sentence pair along the source language and explores all the possible phrase alignments with the target language. In practice, LOD tries to find a phrase translation table that maximizes Eq.(2) as formulated in Algorithm 2. As the existing alignment for $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{f}} \rangle^{(n)}$ in the n -th iteration is a valid alignment subject to three constraints, it also serves as one resolution to the search. In the worst case, if the constrained search can not discover any new alignment other than the existing one, then the existing alignment in the current iteration will stand through the next iteration.

In Algorithm 2, we establish the lattice along the source language. In the case of English to French translation, we follow the phrases in the English order. However, it can be done along the target language as well since our approach follows a symmetric many-to-many word alignment strategy.

This step ends with the promotion of all phrase alignment candidates in the best alignment sequence $T^{(n+1)}$. The promotion includes the merging of the two sub-phrase alignments and the concerning sub-phrases. The merged unit will be considered as a unit in the next iteration.

Algorithm 2. A stack decoding algorithm to explore the best alignment path between source and target languages by considering all alignment candidates in $\mathcal{B}^{(n)} \cup \mathcal{C}^{(n)}$ at n -th iteration.

1. Initialize a lattice of $l^{(n)}$ slots for $l^{(n)}$ sub-phrase in source language.
 2. Starting from $i=1$, for all phrases in source language e_i ;
 - 1) Register all the alignments $t_{ij}^{(n)}$ that map source phrases ending with e_i , including e_i itself, into slot i in the lattice;
 - 2) Register the probability of alignment $p(t_{ij}^{(n)})$ together with the alignment entry $t_{ij}^{(n)}$
 - 3) Repeat 1) and 2) until $i=l^{(n)}$
 3. Apply stack decoding [15] process to find the top n -best paths subject to the three constraints. During the decoding processing, the extension of partial path is subject to a connectivity test to enforce the three constraints.
 4. Output the top best alignment result as the final result.
-

4 Experiments

The objective of our experiments is to validate our LOD approach in machine translation task. Additionally, we are interested in investigating the following: the effect of soft-counting in probability estimation, and the behavior of LOD approach in every iteration, in terms of the length of the phrase unit and the size of the phrase translation table. We report all our experiments using BLEU metrics [10]. Furthermore, we report confidence intervals with 95% statistical significance level of each experiments, as suggested by Koehn [16].

We validate our approach through several experiments using English and French language pairs from the Hansard corpus. We restrict the sentence length to at most 20 words to obtain around 110 thousands sentence pairs. Then we randomly select around 10 thousands sentence pair as our own testing set. In total, the French corpus consists of 994,564 words and 29,360 unique words; while the English corpus consists of 1,055,167 words and 20,138 unique words. Our experiment is conducted on both English-to-French (e2f) and French-to-English (f2e) tasks under open testing set-up. We use these available tools: GIZA++⁵ for word-based IBM 4 model training and ISI ReWrite⁶ for translation test. For measuring the BLEU score and deriving the confidence intervals, we use the publicly available tools⁷.

4.1 Soft-counting vs. Hard-counting

Table 1 summarizes our experiments in analyzing the effect of soft-counting and hard-counting in the probability estimation on the BLEU score. Case I

⁵ <http://www.fjoch.com/>

⁶ <http://www.isi.edu/licensed-sw/rewrite-decoder/>

⁷ <http://www.nist.gov/speech/tests/mt/resources/scoring.htm> and <http://projectile.is.cs.cmu.edu/research/public/tools/bootStrap/tutorial.htm>

demonstrates the BLEU score of the experiment using the underlying translation model probability or soft-counting, while Case II demonstrates the score of hard-counting. The experimental results suggest that the use of the underlying translation model probability is beneficial as it gives consistently higher BLEU scores in all the iterations. The comparison using paired bootstrap resampling [16] also confirms the conclusion.

Table 1. Summary of experiment showing the contribution of using the translation model probability. The experiments are conducted on English-to-French task. Case I indicates the BLEU score of the LOD approach using soft-counting whereas Case II indicates the BLEU score of hard-counting. The value in the column indicates the BLEU score. The range inside the bracket indicates the confidence intervals with 95% statistical significance level.

iteration	Case I	Case II
1	29.60 (29.01-30.14)	28.80 (28.20-29.38)
2	30.72 (30.09-31.29)	30.11 (29.48-30.67)
3	31.52 (30.87-32.06)	30.70 (30.05-31.32)
4	31.93 (31.28-32.50)	30.93 (30.30-31.51)
5	31.90 (31.45-32.68)	31.07 (30.39-31.62)

4.2 LOD Behavior over Iteration

Table 2 summarizes the performance of our LOD approach for the first 10 iterations in comparison with the baseline IBM 4 word-based approach. The results show that the LOD approach produces a significant improvement over IBM 4 consistently. The first iteration yields the biggest improvement. We achieve an absolute BLEU score improvement of 5.01 for the English-to-French task and 5.48 for the French-to-English task from the first iteration. The subsequent improvement is obtained by performing more iterations and capturing longer phrase translation, however, the improvement gained is less significant compared to that of the first iteration.

Table 2 also summarizes the maximum phrase length and the behavior of the phrase translation table: its size and its increment over iteration. It shows that the phrase length is soft-constrained by the maximum likelihood criterion in Eq. (2) rather than limited. As iteration goes on, longer phrases are learnt but their probabilities are less probable than shorter one. Consequently, longer phrases introduce fewer entries to the phrase translation table. Table 2 captures the behavior of the phrase translation table. The first iteration contributes the highest increment of 12.5 % to the phrase translation table while the accumulated increment of table size up to 10th iteration only contributes 27.5% increment over the original size. It suggests that as iteration goes and longer phrases are captured, fewer additional entries are introduced to the phrase translation table.

Table 2. Summary of experiments showing the behavior of LOD approach and the characteristics of the phrase translation table in each iteration. The table shows the translation performance of the word-based IBM 4 approach and the first 10 iteration of LOD approach in BLEU score. The value in the columns indicate the BLEU score while the range inside the bracket represents the confidence intervals with 95% statistical significance level. The table also shows the trend of the phrase translation table: the maximum phrase length, its size, and its increase over iterations.

Iteration	Max Phrase Length	Table Size	Increase	BLEU with confidence intervals	
				e2f	f2e
IBM 4	1	216,852	-	24.59 (24.12-25.21)	26.76 (26.15-27.33)
1	2	244,097	27,245	29.60 (29.01-30.14)	32.24 (31.58-32.83)
2	4	258,734	14,637	30.72 (30.09-31.29)	32.93 (32.28-33.57)
3	7	266,209	7,475	31.52 (30.87-32.06)	33.88 (33.22-34.49)
4	7	270,531	4,322	31.93 (31.28-32.50)	34.14 (33.46-34.76)
5	10	271,793	1,262	31.90 (31.45-32.68)	34.26 (33.56-34.93)
6	11	273,589	1,796	32.14 (31.48-32.72)	34.50 (33.78-35.16)
7	12	274,641	1,052	32.09 (31.43-32.68)	34.55 (33.81-35.18)
8	12	275,399	758	32.07 (31.39-32.60)	34.43 (33.71-35.09)
9	13	275,595	196	31.98 (31.32-32.55)	34.65 (33.93-35.29)
10	14	276,508	913	32.22 (31.55-32.79)	34.61 (33.91-35.26)

The results also show the growth of the size of the phrase translation table is sub-linear and it converges after reasonable number of iterations. This represents a clear advantage of LOD over other related work [6][8].

5 Discussion

In this paper, we propose LOD approach to phrase-based statistical machine translation. The LOD approach addresses three issues in the phrase-based translation framework: the size of phrase translation table, the use of underlying translation model probability and the length of the phrase unit.

In terms of the size of the phrase translation table, our LOD approach presents a sub-linear growth of the phrase translation table. It demonstrates a clear advantage over other reported attempts, such as in [6][8] where the phrase translation table grows almost linearly over the phrase length limit. The LOD approach manages the phrase translation table size in a systematic way as a result of the incorporation of maximum likelihood criterion into the phrase discovery process.

In terms of the use of underlying translation model probability, we propose to use soft-counting instead of hard-counting in the re-estimation processing of probability estimation. In the projection extension algorithm [6], the phrases are learnt based on the presence of alignment in certain configurations. In alignment template[2], two phrases are considered to be translation of each other, if the

word alignments exist within the phrases and not to the words outside. Both methods are based on hard-counting of translation event. Our experiment results suggest the use of soft-counting.

In terms of the length of the phrase unit, we move away from the window-like limit for phrase candidacy [4][9]. The LOD approach is shown to be more flexible in capturing phrases of different length. It gradually explores longer phrases as iteration goes, leading any reasonable length given sufficient iteration as long as they are statistically credible.

It is known that statistical machine translation relies very much on the training corpus. A larger phrase translation table means more training data are needed for the translation model to be statistically significant. In this paper, we successfully introduce the LOD approach to control the process of new phrase discovery process. The results are encouraging.

References

1. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), pp. 263-311.
2. Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20-28, University of Maryland, College Park, MD, June.
3. Franz Josef Och and Hermann Ney. 2000. A Comparison of alignment models for statistical machine translation. In *Proc of the 18th International Conference of Computational Linguistics*, Saarbruken, Germany, July.
4. Daniel Marcu and William Wong. 2002. A phrase-Based, joint probability model for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 133-139, Philadelphia, PA, July.
5. Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation, *Proc. of COLING '96: The 16th International Conference of Computational Linguistics*. pp. 836-841. Copenhagen, Denmark.
6. Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan.
7. Ying Zhang, Stephan Vogel, Alex Waibel. 2003. Integrated phrase segmentation and alignment algorithm for statistical machine translation. in *Proc. of the Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China.
8. Philipp Koehn, Franz Josef Och, Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proc. of the Human Language Technology Conference*, pp. 127-133, Edmonton, Canada, May/June.
9. Ashish Venugopal, Stephan Vogel, Alex Waibel. 2004. Effective phrase translation extraction from alignment models. in *Proc. of 41st Annual Meeting of Association of Computational Linguistics*, pp. 319-326, Sapporo, Japan, July.

10. K. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Report.
11. G. Doddington. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In Proc. of the Conference on Human Language Technology, pp. 138-135, San Diego, CA, USA.
12. Richard Zens, Hermann Ney. 2004. Improvements in phrase-Based statistical machine translation. in Proc. of Conference on Human Language Technology, pp. 257-264, Boston, MA, USA.
13. I. D. Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In Proc. of 2nd Conference on Empirical Methods in Natural Language Processing, Providence, RI.
14. Robert C Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In Proc of Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter, Association for Computational Linguistics, pp. 79-86, Toulouse, France.
15. R Schwartz and Y. L. Chow . 1990. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypothesis. In Proc. of ICASSP 1990, pp. 81-84. Albuquerque, CA.
16. Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 388-395.