Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation

Jiajun Zhang and Chengqing Zong
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Currently, almost all of the statistical machine translation (SMT) models are trained with the parallel corpora in some specific domains. However, when it comes to a language pair or a different domain without any bilingual resources, the traditional SMT loses its power. Recently, some research works study the unsupervised SMT for inducing a simple word-based translation model from the monolingual corpora. It successfully bypasses the constraint of bitext for SMT and obtains a relatively promising result. In this paper, we take a step forward and propose a simple but effective method to induce a phrase-based model from the monolingual corpora given an automatically-induced translation lexicon or a manually-edited translation dictionary. We apply our method for the domain adaptation task and the extensive experiments show that our proposed method can substantially improve the translation quality.

1 Introduction

During the last decade, statistical machine translation has made great progress. Novel translation models, such as phrase-based models (Koehn et a., 2007), hierarchical phrase-based models (Chiang, 2007) and linguistically syntax-based models (Liu et a., 2006; Huang et al., 2006; Galley, 2006; Zhang et al, 2008; Chiang, 2010; Zhang et al., 2011; Zhai et al., 2011, 2012) have been proposed and achieved higher and higher translation performance. However, all of these state-of-the-art translation models rely on the parallel corpora to induce translation rules and estimate the corresponding parameters.

It is unfortunate that the parallel corpora are very expensive to collect and are usually not available for resource-poor languages and for many specific domains even in a resource-rich language pair.

Recently, more and more researchers concentrated on taking full advantage of the monolingual corpora in both source and target languages, and proposed methods for bilingual lexicon induction from non-parallel data (Rapp, 1995, 1999; Koehn and Knight, 2002; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011) and proposed unsupervised statistical machine translation (bilingual lexicon is a byproduct) with only monolingual corpora (Ravi and Knight, 2011; Nuhn et al., 2012; Dou and Knight, 2012).

In the bilingual lexicon induction (Koehn and Knight, 2002; Haghighi et al., 2008; Daumé III and Jagarlamudi, 2011), with the help of the orthographic and context features, researchers adopted an unsupervised method, such as canonical correlation analysis (CCA) model, to automatically induce the word translation pairs between two languages from non-parallel data only requiring that the monolingual data in each language are from a fairly comparable domain.

The unsupervised statistical machine translation method (Ravi and Knight, 2011; Nuhn et al., 2012; Dou and Knight, 2012) viewed the translation task as a decipherment problem and designed a generative model with the objective function to maximize the likelihood of the source language monolingual data. To tackle the large-scale vocabulary, they mainly considered the word-based model (e.g. IBM Model 3) and applied the Bayesian method with Gibbs sampling or slice sampling. Finally, they used the learned translation model directly to translate unseen data (Ravi and Knight, 2011; Nuhn et al., 2012) or incorporated the learned bilingual lexicon as a new in-domain translation resource into the phrase-based model which is trained with out-of-domain data to improve the domain adaptation performance in machine translation (Dou and Knight, 2012).

We can easily see that these unsupervised methods can only induce the word-based translation rules (bilingual lexicon) at present. It is a big challenge that whether we can induce phrase

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1425–1434,
Sofia, Bulgaria, August 4-9 2013.

1, word reordering example:
本 发明 的 目的 在于 ||| the purpose of the invention is to ||| 0-0 0-3 1-4 2-2 3-1 4-5 4-6

2, idiom example:
辨识 ***真伪*** 的 ||| distinguish the ***true*** from the ***false*** ||| 0-0 1-2 1-5 2-1 2-4

3, unknown word translation:
发光 ***二极管*** 芯片 的 ||| of the light-emitting ***diode*** chip ||| 0-2 1-2 2-4 3-0 3-1

Table 1: Examples of new translation knowledge learned with the proposed phrase pair induction method. For the three fields separated by "|||", the first two are respectively Chinese and English phrase, and the last one is the word alignment between these two phrases.

level translation rules and learn a phrase-based model from the monolingual corpora.

In this paper, we focus on exploring this direction and propose a simple but effective method to induce the phrase-level translation rules from monolingual data. The main idea of our method is to divide the phrase-level translation rule induction into two steps: bilingual lexicon induction and phrase pair induction.

Since many researchers have studied the bilingual lexicon induction, in this paper, we mainly concentrate ourselves on phrase pair induction given a probabilistic bilingual lexicon and two in-domain large monolingual data (source and target language). In addition, we will further introduce how to refine the induced phrase pairs and estimate the parameters of the induced phrase pairs, such as four standard translation features and phrase reordering feature used in the conventional phrase-based models (Koehn et al., 2007). The induced phrase-based model will be used to help domain adaptation for machine translation.

In the rest of this paper, we first explain with examples to show what new translation knowledge can be learned with our proposed phrase pair induction method (Section 2), and then we introduce the approach for probabilistic bilingual lexicon acquisition in Section 3. In Section 4 and 5, we respectively present our method for phrase pair induction and introduce an approach for phrase pair refinement and parameter estimation. Section 6 will show the detailed experiments for the task of domain adaptation. We will introduce some related work in Section 7 and conclude this paper in Section 8.

2 What Can We Learn with Phrase Pair Induction?

Readers may doubt that if phrase pair induction is performed only using bilingual lexicon and monolingual data, what new translation knowledge can be learned?

The bilingual lexicon can only express the translation equivalence between source- and target-side word pair and has little ability to deal with word reordering and idiom translation. In contrast, phrase pair induction can make up for this deficiency to some extent. Furthermore, our method is able to learn some unknown word translations.

From the induced phrase pairs with our method, we have conducted a deep analysis and find that we can learn three kinds of new translation knowledge: 1) word reordering in a phrase pair; 2) idioms; and 3) unknown word translations. Table 1 gives examples for each of the three kinds. For the first example, the source and target phrase are extracted respectively from monolingual data, each word in the source phrase has a translation in the target phrase, but the word order is different. The word order encoded in a phrase pair is difficult to learn in a word-based SMT. In the second example, the *italic* source word corresponds to two target words (in *italic*), and the phrase pair is an idiom which cannot be learned from word-based SMT. In the third example, as we learn from the source and target monolingual text that the words around the *italic* ones are translations with each other, thus we cannot only extract a new phrase pair but also learn a translation pair of unknown words in *italic*.

3 Probabilistic Bilingual Lexicon Acquisition

In order to induce the phrase pairs from the in-domain monolingual data for domain adaptation, the probabilistic bilingual lexicon is essential.

In this paper, we acquire the probabilistic bilingual lexicon from two approaches: 1) build a bilingual lexicon from large-scale out-of-domain parallel data; 2) adopt a manually collected in-domain lexicon. This paper uses Chinese-to-English translation as a case study and electronic data is the in-domain data we focus on.

In Chinese-to-English translation, there are lots of parallel data on News. Here, we utilize about 2.08 million sentence pairs[1] in News domain to learn a probabilistic bilingual lexicon. Basically, we can use GIZA++ (Och, 2003) to get the probabilistic lexicon. However, the problem is that each source-side word associates too many possible translations which contain much noise. For instance, in the lexicon obtained with GIZA++, each source-side word has about 13 translations on average. The noise of the lexicon can influence the accuracy of the induced phrase pairs to a large extent. To learn a lexicon with a high precision, we follow Munteanu and Marcu (2006) to apply *Log-Likelihood-Ratios* (Dunning, 1993; Melamed, 2000; Moore, 2004a, 2004b) to estimate how strong the association is between a source-side word and its aligned target-side word. We employ the same algorithm used in (Munteanu and Marcu, 2006) which first use the GIZA++ (with grow-diag-final-and heuristic) to obtain the word alignment between source and target words, and then calculate the association strength between the aligned words. After using the log-likelihood-ratios algorithm[2], we obtain a probabilistic bilingual lexicon with bidirectional translation probabilities from the out-of-domain data. In the final lexicon, the number of average translations is only 5. We call this lexicon *LLR-lex*.

In the electronic domain, we manually collected a lexicon which contains about 140k entries. It should be noted that there is no translation probability in this lexicon. In order to assign probabilities to each entry, we apply the *Corpus Translation Probability* which used in (Wu et al., 2008): given an in-domain source language monolingual data, we translate this data with the phrase-based model trained on the out-of-domain News data, the in-domain lexicon and the in-domain target language monolingual data (for language model estimation). With the source language data and its translation, we estimate the bidirectional translation probabilities for each entry in the original lexicon. For the entries whose translation probabilities are not estimated, we just assign a uniform probability. That is if a source word has *n* translations, then the translation probability of target word given the source word is 1/*n*. We call this lexicon *Domain-lex*.

[1] LDC category numbers are: LDC2000T50, LDC2003E14, LDC2003E07, LDC2004T07, LDC2005T06, LDC2002L27, LDC2005T10 and LDC2005T34.

[2] Following Moore (2004b), we use the threshold 10 on LLR to filter out unlikely translations.

We combine *LLR-lex* and *Domain-lex* to obtain the final probabilistic bilingual lexicon for phrase pair induction.

4 Phrase Pair Induction Method

Given a probabilistic bilingual lexicon and two monolingual data, we present a simple but effective method for phrase pair induction in this section.

Input: Probabilistic bilingual lexicon V (each source word s maps a translation set $V[s]$)
Source language monolingual data $S=\{s_n\}$ n=1...N
Target language monolingual data $T=\{t_m\}$ m=1...M
Output: Phrase pairs P

1: For each distinct source-side phrase s_i^j in S:
2: If each $s_k \in s_i^j$ in V:
3: Collect $V[s_k]_{k=i}^{j}$
4: For each permutation $s_{i'}^{j'}$ of s_i^j :
5: If $t_{i'}^{j'}$ in T: $\triangleright$ $t_{k'} \in V[s_{k'}]$ $k' \in [i, j]$
6: Add phrase pair $\left(s_i^j, t_{i'}^{j'}\right)$ into P

Figure 1: a naïve algorithm for phrase pair induction.

4.1 A Naïve Method

We first introduce a relatively naïve way to extract phrase pairs from the given resources. For a source phrase (word sequence), we can reorder the words in the phrase (permutation) first, and then obtain the target phrases with the bilingual lexicon (translation), and finally check if the target phrase is in the target monolingual data. The algorithm is given in Figure 1.

Figure 1 shows that the naïve algorithm is very easy to implement. However, the time complexity is too high. For each source phrase s_i^j (with $(j-i+1)!$ permutations), suppose a source word has C translations on average and checking whether the target phrase $t_{i'}^{j'}$ in T needs time $O(|T|)$, then, phrase pair induction for a single source phrase needs time $O\left(C^{|j-i+1|}|T|(j-i+1)!\right)$. It is very time consuming. One may design smarter algorithms. For example, one can collect distinct n-grams from source and target monolingual data. Then, for a source-side phrase with length L, one can find the best translation candidate using the probabilistic bilingual lexicon from the target-side phrases with the same length L. The biggest disadvantage of these algorithms is that they can only induce phrase pair (with the

same length) encoding word reordering, but cannot learn phrase pairs in different length. Furthermore, they cannot learn idioms and unknown word translations from monolingual data. Obviously, these kind of approaches is not optimal.

4.2 Phrase Pair Induction with Inverted Index

In order to make the phrase pair induction both effective and efficient, we propose a method using inverted index data structure which is usually a central component of a typical search engine.

The inverted index is employed to represent the target language monolingual data. For a target language word, the inverted index not only records the sentence position in monolingual data, but also records the word position in a sentence. Some examples are shown in Table 2. By doing this, we do not need to iterate all the permutations of source language phrase s_i^j to explore possible phrase pairs encoding word reordering. Furthermore, it is possible to learn idiom translation and unknown word translations. We will elaborate how to induce phrase pairs with the help of inverted index.

Target Language Word	Position
communication	(2,5), (106,20), …, (23022, 12)
…	…
zoom	(90,2), (280,21), …, (90239,15)

Table 2: Some examples of inverted index for target language words, (2,5) means that "communication" occurs at the 5th word of the 2nd sentence in the target monolingual data.

The new algorithm for phrase pair induction is presented in Figure 2. Line 1 iterates all the distinct phrases in the source-side monolingual data. It can be implemented by collecting all the distinct n-grams in which n is the phrase length we are interested in (3 to 7 in this paper). For each distinct source-side phrase, Line 2-5 efficiently collects all the positions in the target monolingual data for the translations of each word in the source phrase. Line 6 sorts the positions so that we can easily find the position sequence belonging to a same sentence. Line 8-9 discards all the position sub-sequences that lack translations for more than one source-side words. That is to say we allow at most one unknown word in an induced phrase pair in order to make the induction more accurate. Line 10 and Line 12 is the core of this algorithm. We first define a constraint before detailing the algorithm.

Input: Probabilistic bilingual lexicon V (each source word s maps a translation set $V[s]$)
Source language monolingual data $S=\{s_n\}$ n=1...N
Inverted index representing target language monolingual data *IMap*
Output: Phrase pairs P

1: For each distinct source-side phrase s_i^j in S:
2: *positionArray* = []
3: For each $s_k \in s_i^j$:
4: For each $t \in V[s_k]$:
5: add *IMap*[t] into *positionArray*
6: Sort *positionArray*
7: For each sequence in a same sentence in *positionArray*:
8: If more than 1 word in s_i^j has no trans in the seq:
9: Discard this seq and continue
10: ***Probability smoothing for single word gap***
11: For all continuous position sub-sequence:
12: ***Find the one*** t_h^k ***with maximum probability***
13: Add phrase pair $\left(s_i^j, t_h^k\right)$ into P

Figure 2: Phrase pair induction using inverted index.

Constraint: we require that there exists at most one phrase in a target sentence that is the translation of the source-side phrase.

According to our analysis, it is not often to find that two phrases (length larger than 2) in a same sentence have the same meaning. Even if it happens, it is reasonable to keep the one with the highest probability. Given a position sequence belonging to a same sentence, Line 10 smoothes the probability of the single word gap according to the probabilities of the around words. Single word gap means that this word is not aligned but its left and right words are aligned with the words of the source-side phrase. Suppose the target sub-sequence is $t_i \cdots t_{i+r} \cdots t_j$ and t_{i+r} is the only word that is not aligned with source-side words. We smooth the probability $p(t_{i+r} \mid null)$ as follows:

$$p(t_{i+r} \mid null) = \begin{cases} \dfrac{\min\left\{p(t_i \mid s_{t_i}), p(t_j \mid s_{t_j})\right\}}{2}, & if\ r=1 or\ r+1=j \\ \dfrac{p(t_{i+r-1} \mid s_{t_{i+r-1}}) + p(t_{i+r+1} \mid s_{t_{i+r+1}})}{2}, & otherwise \end{cases} \quad (1)$$

The above formula means that if the left or the right side only has one word, then the smoothed probability is one half of the minimum of the probabilities of the two neighbors, otherwise the smoothed probability is the average of the probabilities of the two neighbors. This smoothing strategy encourages that if more words around the un-aligned word are translations of the source-side phrase, then the gap word is more likely to belong to the translations of the source-side phrase.

After probability smoothing of the single gap word, we are ready to extract the candidate translation of the source-side phrase. Similar with Line 9 in Figure 2, we further filter the target continuous phrase if more than one word in source-side phrase has no translation in this target phrase. After that, we just choose the continuous target phrase with the largest probability if two or more continuous target phrases exist in the same target sentence. The probability of a target-side phrase given the source-side phrase is computed similar to that of (Koehn et al., 2003) except that we impose length normalization:

$$p_{lex}(t \mid s,a) = \left[\prod_{i=1}^{n} \frac{1}{\left| \{ j \mid (i,j) \in a \} \right|} \sum_{\forall (i,j) \in a} p(t_i \mid s_j) \right]^{\frac{1}{n}} \quad (2)$$

where the alignment a is produced using probabilistic bilingual lexicon. If a target word in t is a gap word, we suppose there is a word alignment between the target gap word and the source-side *null*.

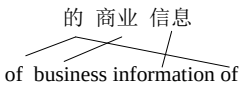
Similarly, we can compute the probability of source-side phrase given the target-side phrase $p_{lex}(s \mid t,a)$. Then, we find the target-side phrase which has the biggest value of $p_{lex}(t \mid s,a) \cdot p_{lex}(s \mid t,a)$. Line 13 in Figure 2 collects the induced phrase pairs.

For the time complexity, it depends on the length of *positionArray*, since the time complexity of the core algorithm (Line 7-13) is proportional to the length of *positionArray*. If *positionArray* contains almost all the positions in the target monolingual data *T*, then the worst time complexity will be $O(|T| \log |T|)$ (for array sort). However, we find in the target monolingual data (1 million sentences) that each distinct word happens 110 times on average. Then, for a sources-side phrase with 7 words, the average length of *positionArray* will be 3850, since each source word has averagely 5 target translations (mentioned in Section 3). Therefore, the algorithm is relatively efficient in the average case.

5 Phrase Pair Refinement and Parameterization

5.1 Phrase Pair Refinement

Some of the phrase pairs induced in Section 4 may contain noise. According to our analysis, we find that the biggest problem is that in the target-side of the phrase pair, there are two or more identical words aligned to the same source-side word. For example, we extract a phrase pair as follows:



In the above phrase pair, there are two words "of" in the target side and the first one is redundant. The phrase pair induction algorithm presented in Section 4 cannot deal with this situation. In this section, we propose a simple approach to handle this problem. For each entry in *LLR-lex*, such as (的, of), we can learn two kinds of information from the out-of-domain word-aligned sentence pairs: one is whether the target translation is before or after the translation of the preceding source-side word (Order); the other is whether the target translation is adjacent with the translation of the preceding source-side word (Adjacency). If the source-side word is the beginning of the phrase, we calculate the corresponding information with the succeeding word instead of the preceding word. For the entries in *Domain-lex*, we constrain that the target translation should be adjacent with the translations of its source-side neighbors and translation order is the same with the source-side words.

With the Order and Adjacency information, we first check the order information, and then check the adjacency information if the duplicates cannot be handled using order information. For example, since (的, of) is an entry in *LLR-lex* and we have learned that "of" is much more likely to be behind the translation of the succeeding word. Thus, the first word "of" can be discarded. This refinement can be applied before finding the phrase pair with maximum probability (Line 12 in Figure 2) so that the duplicate words do not affect the calculation of translation probability of phrase pair.

5.2 Translation Probability Estimation

It is well known that in the phrase-based SMT there are four translation probabilities and the reordering probability for each phrase pair.

The translation probabilities in the traditional phrase-based SMT include bidirectional phrase translation probabilities and bidirectional lexical weights. For the lexical weights, we can use the $p_{lex}(s \mid t,a)$ and $p_{lex}(t \mid s,a)$ computed in the above section without length normalization. However, for the phrase-level probability, we cannot use maximum likelihood estimation since the phrase pairs are not extracted from parallel sentences.

In this paper, we borrow and extend the idea of (Klementiev et al., 2012) to calculate the phrase-level translation probability with context information in source and target monolingual corpus. The value is calculated using a vector space model. With source and target vocabularies $(s_1, s_2, \cdots, s_N)$ and $(t_1, t_2, \cdots, t_M)$, the source-side phrase *s* and target-side phrase *t* can be respectively represented in an *N*- and *M*-dimensional vector. The *k*-th component of *s*'s contextual vector is computed using the method of (Fung and Yee, 1998) as follows:

$$w_k = n_{s,k} \times \left(\log\left(n_{\max} / n_k\right) + 1\right) \quad (3)$$

where $n_{s,k}$ and n_k denotes the number of times s_k occurs in the context of *s* and in the entire source language monolingual data, and $n_{\max}$ is the maximum number of occurrence of any source-side word in the source language monolingual data. The *k*-th element of *t*'s vector can be computed with the same method. We finally normalize these vectors with L_2-norm.

With the *s*'s and *t*'s contextual vector representations, we calculate two similarities: 1) project *s*'s vector into target side $\hat{t}$ with the lexical mapping *p(t|s)*, and then get the similarity by computing the cosine of two angles between *t*'s and $\hat{t}$'s vectors; 2) project *t*'s vector into source side $\hat{s}$ with the lexical mapping *p(s|t)*, and then obtain the similarity between *s*'s and $\hat{s}$'s vectors. These two contextual similarities will serve as two phrase-level translation probabilities.

5.3 Reordering Probability Estimation

For the reordering probabilities of newly induced phrase pairs, we can also follow Klementiev et al. (2012) to estimate these probabilities using source and target monolingual data. The method is to calculate six probabilities for monotone, swap or discontinuous cases. For the phrase pair (的 商业 信息, business information of), we find a source sentence containing 的 商业 信息, and find a target sentence containing *business information of*. If there is another phrase pair $(\bar{s}, \bar{t})$, $\bar{t}$ exactly follows *business information of* and $\bar{s}$ occurs in the same source sentence with 的 商业 信息, then we compare the position relationship between $\bar{s}$ and 的 商业 信息. We increment the swap count if $\bar{s}$ is just before 的 商业 信息. After counting, we finally use maximum likelihood estimation method to compute the reordering probabilities.

6 Related Work

As far as we know, few researchers study phrase pair induction from only monolingual data.

There are three research works that are most related with ours. One is using an in-domain probabilistic bilingual lexicon to extract sub-sentential parallel fragments from comparable corpora (Munteanu and Marcu, 2006; Quirk et al., 2007; Cettolo et al., 2010). Munteanu and Marcu (2006) first extract the candidate parallel sentences from the comparable corpora and further extract the accurate sub-sentential bilingual fragments from the candidate parallel sentences using the in-domain probabilistic bilingual lexicon. Compared with their work, our focus is to induce phrase pairs directly from monolingual data rather than comparable data. Thus, finding the candidate parallel sentences is not possible in our situation.

Another is to make full use of monolingual data with transductive learning (Ueffing et al., 2007; Schwenk, 2008; Wu et al., 2008; Bertoldi and Federico, 2009). For the target-side monolingual data, they just use it to train language model, and for the source-side monolingual data, they employ a baseline (word-based SMT or phrase-based SMT trained with small-scale bitext) to first translate the source sentences, combining the source sentence and its target translation as a bilingual sentence pair, and then train a new phrase-base SMT with these pseudo sentence pairs. This method cannot learn idiom translations and unknown word translations.

The third is to estimate the translation parameters and reordering parameters using monolingual data given the phrase pairs (Klementiev et al., 2012). Their work supposes the phrase pairs are already given and then corresponding parameters can be learned with monolingual data. Different from their work, we concentrate ourselves on inducing phrase pairs from monolingual data and then borrow some ideas from theirs for parameter estimation. Furthermore, we extend their contextual similarity between source and target phrases to both directions.

7 Experiments

7.1 Experimental Setup

Our purpose is to induce phrase pairs to improve translation quality for domain adaptation. We have introduced the out-of-domain data and the electronic in-domain lexicon in Section 3. Here we introduce other information about the in-

domain data. Besides the in-domain lexicon, we have collected respectively 1 million monolingual sentences in electronic area from the web. They are neither parallel nor comparable because we cannot even extract a small number of parallel sentence pairs from this monolingual data using the method of (Munteanu and Marcu, 2006). We further employ experts to translate 2000 Chinese electronic sentences into English. The first half is used as the tuning set (*elec1000-tune*) and the second half is employed as the testing set (*elec1000-test*).

We construct two kinds of phrase-based models using Moses (Koehn et al., 2007): one uses out-of-domain data and the other uses in-domain data. For the out-of-domain data, we build the phrase table and reordering table using the 2.08 million Chinese-to-English sentence pairs, and we use the SRILM toolkit (Stolcke, 2002) to train the 5-gram English language model with the target part of the parallel sentences and the Xinhua portion of the English Gigaword. For the in-domain electronic data, we first consider the lexicon as a phrase table in which we assign a constant 1.0 for each of the four probabilities, and then we combine this initial phrase table and the induced phrase pairs to form the new phrase table. The in-domain reordering table is created for the induced phrase pairs. An in-domain 5-gram English language model is trained with the target 1 million monolingual data.

We use BLEU (Papineni et al., 2002) score with shortest length penalty as the evaluation metric and apply the pairwise re-sampling approach (Koehn, 2004) to perform the significance test.

7.2 Experimental Results

In this section, we first conduct experiments to figure out how the translation performance degrades when the domain changes. To better illustrate the comparison, we first use News data to evaluate the NIST evaluation tests and then use the same News data to evaluate the electronic test sets. For the NIST evaluation, we employ Chinese-to-English NIST MT03 as the tuning set and NIST MT05 as the test set. Table 3 gives the results. It is obvious that, it is relatively high when using the News training data to evaluate the same News test set. However, when the test domain is changed, the translation performance decreases to a large extent.

Given the in-domain bilingual lexicon and two monolingual data, previous works also proposed some good methods to explore the potential of the given data to improve the translation quality. Here, we implement their approaches and use them as our strong baseline. Wu et al. (2008) regards the in-domain lexicon with corpus translation probability as another phrase table and further use the in-domain language model besides the out-of-domain language model. Table 4 gives the results. We can see from the table that the domain lexicon is much helpful and significantly outperforms the baseline with more than 4.0 BLEU points. When it is enhanced with the in-domain language model, it can further improve the translation performance by more than 2.5 BLEU points. This method has made good use of in-domain lexicon and the target-side in-domain monolingual data, but it does not take full advantage of the in-domain source-side monolingual data.

In order to use source-side monolingual data, Ueffing et al. (2007), Schwenk (2008), Wu et al. (2008) and Bertoldi and Federico (2009) employed the transductive learning to first translate the source-side monolingual data using the best configuration (baseline+in-domain lexicon+in-domain language model) and obtain 1-best translation for each source-side sentence. With the source-side sentences and their translations, the new phrase table and reordering table are built. Then, these resources are added into the best configuration. The experimental results are presented in the last low of Table 4. From the results, we see that transductive learning can further improve the translation performance significantly by 0.6 BLEU points.

In tranductive learning, in-domain lexicon and both-side monolingual data have been explored. However, this method does not take full advantage of both-side monolingual data because it uses source and target monolingual data individually. In our method, we explore fully the source and target monolingual data to induce translation equivalence on the phrase level. In order to make the phrase pair induction more efficient, we first sort all the sentences in the both-side monolingual data according to the word hit rate in the bilingual lexicon. Then, we conduct six sets of experiments respectively on the first 100k, 200k, 300k, 500k and whole 1m sentences. All the experiments are run based on the configuration with BLEU 13.41 in Table 4, and we call this configuration *BestConfig*. Note that the unknown words are only allowed if the source-side of a phrase pair has more than 3 words. Table 5 shows the results.

Training Data	Tune Data (NIST MT03)	Test Data (NIST MT05)
2.08M sentence pairs in News	35.79	34.26
	Tune Data (elec1000-tune)	Test Data (elec1000-test)
	7.93	6.69

Table 3: Experimental results using News training data to test NIST evaluation data and electronic data (numbers denote BLEU score points in percent).

Method	Tune (elec1000-tune)	Test (elec1000-test)
Baseline	7.93	6.69
baseline + in-domain lexicon	10.97	**10.87**
baseline + in-domain lexicon + in-domain language model	13.72	**13.41**$^{++}$
Transductive Learning	14.13	**14.01***

Table 4: Experimental results using News training data, in-domain lexicon, language model and transductive learning. **Bold** figures mean that the results are statistically significant better than the baseline with $p<0.01$, and "++" denotes the result is statistically significant better than *baseline+in-domain lexicon*. "*" means that the result is statistically significant better than 13.41 with $p<0.05$.

Method	Tune (BLEU %)	Test (BLEU %)
BestConfig	13.72	13.41
+phrase pair induction (100k)	14.23	**14.06**
+phrase pair induction (200k)	14.45	***14.24***
+phrase pair induction (300k)	14.76	***14.83***$^{++}$
+phrase pair induction (500k)	14.98	***15.16***$^{++}$
+phrase pair induction (1m)	15.11	***15.30***$^{++}$

Table 5: Experimental results of our phrase pair induction method. **Bold** figures denotes the corresponding method significantly outperform the BestConfig with $p<0.05$. **Bold** and *Italic* figures means the results are significantly better than that of BestConfig with $p<0.01$. "++" denotes that the corresponding approach performs significantly better than Transductive Learning with $p<0.01$.

Method	Before Filtering	After Filtering
+phrase pair induction (100k)	72,615	8,724
+phrase pair induction (200k)	108,948	12,328
+phrase pair induction (300k)	136,529	17,505
+phrase pair induction (500k)	150,263	19,862
+phrase pair induction (1m)	169,172	21,486

Table 6: the number of phrase pairs induced with different size of monolingual data.

We can see from the table that our method obtains the best translation performance. When using the first 100k sentences for phrase pair induction, it obtains a significant improvement over the BestConfig by 0.65 BLEU points and can outperform the transductive learning method. When we use more monolingual data, the performance becomes even better. The method of phrase pair induction using 300k sentences performs quite well. It outperforms the BestConfig significantly with an improvement of 1.42 BLEU points and it also performs much better than transductive learning method with gains of 0.82 BLEU points. With the monolingual data larger and larger, the gains become smaller and smaller because the word hit rate gets lower and lower. These experimental results empirically show the effectiveness of our proposed phrase pair induction method.

A question remains that how many new phrase pairs are induced with different size of monolingual data. Here, we give respectively the statistics before and after filtering with the 1000 test sentences. Table 6 shows the statistics. We can see from the table that lots of new phrase pairs can be induced since the source and target monolingual data is in the same domain. However, since the source and target monolingual data is

far from parallel, most of the phrase pairs are not long. For example, in the 108,948 distinct phrase pairs, we find that the phrase pair distribution according to source-side length is (3:50.6%, 4:35.6%, 5:3.3%, 6:9.8%, 7:0.7%). It is easy to see that the phrase pairs whose source-side length longer than 4 account for only a very small part.

8 Conclusion and Future Work

This paper proposes a simple but effective method to induce phrase pairs from monolingual data. Given the probabilistic bilingual lexicon and both-side monolingual data in the same domain, the method employs inverted index structure to represent the target-side monolingual data, and induce the translations for each distinct source-side phrase with the help of the bilingual lexicon. We further propose an approach to refine the result phrase pairs to make them more accurate. We also introduce how to estimate the translation and reordering parameters for the induced phrase pairs with monolingual data. Extensive experiments on domain adaptation have shown that our method can significantly outperform previous methods which also focus on exploring the in-domain lexicon and monolingual data.

However, through the analysis we find that our induced phrase pairs still contain some noise, such as the words in source- and target-side of the phrase pair are all aligned but the target-side phrase expresses the different meaning. Furthermore, our proposed method cannot learn expressions which are not lexical translations but are semantic ones. In the future, we will study further on these phenomena and propose new methods to handle these problems.

Acknowledgments

The research work has been funded by the Hi-Tech Research and Development Program ("863" Program) of China under Grant No. 2011AA01A207, 2012AA011101 and 2012AA011102, and also supported by the Key Project of Knowledge Innovation of Program of Chinese Academy of Sciences under Grant No. KGZD-EW-501. We would also like to thank the anonymous reviewers for their valuable suggestions.

References

Nicola Bertoldi and Marcello Federico, 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proc. of the Fourth Workshop on Statistical Machine Translation*, pages 182-189.

Mauro Cettolo, Marcello Federico and Nicola Bertoldi, 2010. Mining parallel fragments from comparable texts. In *Proc. of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 227-234.

David Chiang, 2007. Hierarchical phrase-based translation. *computational linguistics*, 33 (2). pages 201-228.

David Chiang, 2010. Learning to translate with source and target syntax. In *Proc. of ACL 2010*, pages 1443-1452.

Hal Daumé III and Jagadeesh Jagarlamudi, 2011. Domain adaptation for machine translation by mining unseen words. In *Proc. of ACL-HLT 2011*.

Qing Dou and Kevin Knight, 2012. Large Scale Decipherment for Out-of-Domain Machine Translation. In *Proc. of EMNLP-CONLL 2012.*

Ted Dunning, 1993. Accurate methods for the statistics of surprise and coincidence. *computational linguistics*, 19 (1). pages 61-74.

Pascale Fung and Lo Yuen Yee, 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of ACL-COLING 1998.*, pages 414-420.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang and Ignacio Thayer, 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of COLING-ACL 2006*, pages 961-968.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick and Dan Klein, 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. of ACL-08: HLT*, pages 771-779.

Liang Huang, Kevin Knight and Aravind Joshi, 2006. A syntax-directed translator with extended domain of locality. In *Proc. of AMTA 2006*, pages 1-8.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch and David Yarowsky, 2012. Toward statistical machine translation without parallel corpora. In *Proc. of EACL 2012.*, pages 130-140.

Philipp Koehn, 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP 2004.*, pages 388-395, Barcelona, Spain, July 25th-26th, 2004.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst, 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL on Interactive Poster and Demonstration Sessions 2007.*, pages 177-180, Prague, Czech Republic, June 27th-30th, 2007.

Philipp Koehn and Kevin Knight, 2002. Learning a translation lexicon from monolingual corpora. In

Proc. of the ACL-02 workshop on Unsupervised lexical acquisition, pages 9-16.

Yang Liu, Qun Liu and Shouxun Lin, 2006. Tree-to-string alignment template for statistical machine translation. In *Proc. of COLING-ACL 2006*, pages 609-616.

I. Dan Melamed, 2000. Models of translational equivalence among words. *computational linguistics*, 26 (2). pages 221-249.

Rorbert C. Moore, 2004a. Improving IBM word-alignment model 1. In *Proc. of ACL 2004.*

Rorbert C. Moore, 2004b. On log-likelihood-ratios and the significance of rare events. In *Proc. of EMNLP 2004.*, pages 333-340.

Dragos Stefan Munteanu and Daniel Marcu, 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proc. of ACL-COLING 2006.*

Malte Nuhn, Arne Mauser and Hermann Ney, 2012. Deciphering Foreign Language by Combining Language Models and Context Vectors. In *Proc. of ACL 2012.*

Franz Josef Och and Hermann Ney., 2003. A systematic comparison of various statistical alignment models. *computational linguistics*, 29 (1). pages 19-51.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL 2002.*, pages 311-318.

Chris Quirk, Raghavendra Udupa and Arul Menezes, 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proc. of the Machine Translation Summit XI*, pages 377-384.

Reinhard Rapp, 1995. Identifying word translations in non-parallel texts. In *Proc. of ACL 1995*, pages 320-322.

Reinhard Rapp, 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of ACL 1999*, pages 519-526.

Sujith Ravi and Kevin Knight, 2011. Deciphering foreign language. In *Proc. of ACL 2011.*, pages 12-21.

Holger Schwenk, 2008. Investigations on largescale lightly-supervised training for statistical machine translation. In *Proc. of IWSLT 2008*, pages 182-189.

Andreas Stolcke, 2002. SRILM-an extensible language modeling toolkit. In *Proc. of 7th International Conference on Spoken Language Processing*, pages 901-904, Denver, Colorado, USA, September 16th-20th, 2002.

Nicola Ueffing, Gholamreza Haffari and Anoop Sarkar, 2007. Transductive learning for statistical machine translation. In *Proc. of ACL 2007.*

Hua Wu, Haifeng Wang and Chengqing Zong, 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proc. of COLING 2008.*, pages 993-1000.

Feifei Zhai, Jiajun Zhang, Yu Zhou and Chengqing Zong, 2011. Simple but effective approaches to improving tree-to-tree model. In *Proc. of MT Summit XIII* 2011, pages 261-268.

Feifei Zhai, Jiajun Zhang, Yu Zhou and Chengqing Zong, 2012. Tree-based translation without using parse trees. *In Proc. of COLING 2012*, pages 3037-3054.

Jiajun Zhang, Feifei Zhai and Chengqing Zong, 2011. Augmenting string-to-tree translation models with fuzzy use of the source-side syntax. *In Proc. of EMNLP 2011*, pages 204-215.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan and Sheng Li, 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proc. of ACL-08: HLT*, pages 559-567.