A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation

Masao Utiyama and **Hitoshi Isahara**
National Institute of Information and Communications Technology
3-5 Hikari-dai, Soraku-gun, Kyoto 619-0289 Japan
{mutiyama,isahara}@nict.go.jp

Abstract

We compare two pivot strategies for phrase-based statistical machine translation (SMT), namely *phrase translation* and *sentence translation*. The phrase translation strategy means that we directly construct a phrase translation table (phrase-table) of the source and target language pair from two phrase-tables; one constructed from the source language and English and one constructed from English and the target language. We then use that phrase-table in a phrase-based SMT system. The sentence translation strategy means that we first translate a source language sentence into n English sentences and then translate these n sentences into target language sentences separately. Then, we select the highest scoring sentence from these target sentences. We conducted controlled experiments using the Europarl corpus to evaluate the performance of these pivot strategies as compared to directly trained SMT systems. The phrase translation strategy significantly outperformed the sentence translation strategy. Its relative performance was 0.92 to 0.97 compared to directly trained SMT systems.

1 Introduction

The rapid and steady progress in corpus-based machine translation (Nagao, 1981; Brown et al., 1993) has been supported by large parallel corpora such as the Arabic-English and Chinese-English parallel corpora distributed by the Linguistic Data Consortium and the Europarl corpus (Koehn, 2005), which consists of 11 European languages. However, large parallel corpora do not exist for many language pairs. For example, there are no publicly available Arabic-Chinese large-scale parallel corpora even though there are Arabic-English and Chinese-English parallel corpora.

Much work has been done to overcome the lack of parallel corpora. For example, Resnik and Smith (2003) propose mining the web to collect parallel corpora for low-density language pairs. Utiyama and Isahara (2003) extract Japanese-English parallel sentences from a noisy-parallel corpus. Munteanu and Marcu (2005) extract parallel sentences from large Chinese, Arabic, and English non-parallel newspaper corpora.

Researchers can also make the best use of existing (small) parallel corpora. For example, Nießen and Ney (2004) use morpho-syntactic information to take into account the interdependencies of inflected forms of the same lemma in order to reduce the amount of bilingual data necessary to sufficiently cover the vocabulary in translation. Callison-Burch et al. (2006a) use paraphrases to deal with unknown source language phrases to improve coverage and translation quality.

In this paper, we focus on situations where no parallel corpus is available (except a few hundred parallel sentences for tuning parameters). To tackle these extremely scarce training data situations, we propose using a pivot language (English) to bridge the

Proceedings of NAACL HLT 2007, pages 484–491,
Rochester, NY, April 2007.

source and target languages in translation. We first translate source language sentences or phrases into English and then translate those English sentences or phrases into the target language, as described in Section 3. We thus assume that there is a parallel corpus consisting of the source language and English as well as one consisting of English and the target language. Selecting English as a pivot language is a reasonable pragmatic choice because English is included in parallel corpora more often than other languages are, though any language can be used as a pivot language.

In Section 2, we describe a phrase-based statistical machine translation (SMT) system that was used to develop the pivot methods described in Section 3. This is the shared task baseline system for the 2006 NAACL/HLT workshop on statistical machine translation (Koehn and Monz, 2006) and consists of the Pharaoh decoder (Koehn, 2004), SRILM (Stolcke, 2002), GIZA++ (Och and Ney, 2003), mkcls (Och, 1999), Carmel,[1] and a phrase model training code.

2 Phrase-based SMT

We use a phrase-based SMT system, Pharaoh, (Koehn et al., 2003; Koehn, 2004), which is based on a log-linear formulation (Och and Ney, 2002). It is a state-of-the-art SMT system with freely available software, as described in the introduction. The system segments the source sentence into so-called phrases (a number of sequences of consecutive words). Each phrase is translated into a target language phrase. Phrases may be reordered.

Let $\mathbf{f}$ be a source sentence (e.g, French) and $\mathbf{e}$ be a target sentence (e.g., English), the SMT system outputs an $\hat{\mathbf{e}}$ that satisfies

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) \quad (1)$$

$$= \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \quad (2)$$

where $h_m(\mathbf{e}, \mathbf{f})$ is a feature function and λ_m is a weight. The system uses a total of eight feature functions: a trigram language model probability of the target language, two phrase translation probabilities (both directions), two lexical translation probabilities (both directions), a word penalty, a phrase penalty, and a linear reordering penalty. For details on these feature functions, please refer to (Koehn et al., 2003; Koehn, 2004; Koehn et al., 2005). To set the weights, λ_m, we carried out minimum error rate training (Och, 2003) using BLEU (Papineni et al., 2002) as the objective function.

[1] http://www.isi.edu/licensed-sw/carmel/

3 Pivot methods

We use the phrase-based SMT system described in the previous section to develop pivot methods. We use English $\mathbf{e}$ as the pivot language. We use French $\mathbf{f}$ and German $\mathbf{g}$ as examples of the source and target languages in this section.

We describe two types of pivot strategies, namely *phrase translation* and *sentence translation*.

The phrase translation strategy means that we directly construct a French-German phrase translation table (phrase-table for short) from a French-English phrase-table and an English-German phrase-table. We assume that these French-English and English-German tables are built using the phrase model training code in the baseline system described in the introduction. That is, phrases are heuristically extracted from word-level alignments produced by doing GIZA++ training on the corresponding parallel corpora (Koehn et al., 2003).

The sentence translation strategy means that we first translate a French sentence into n English sentences and translate these n sentences into German separately. Then, we select the highest scoring sentence from the German sentences.

3.1 Phrase translation strategy

The phrase translation strategy is based on the fact that the phrase-based SMT system needs a phrase-table and a language model for translation. Usually, we have the language model of a target language. Consequently, we only need to construct a phrase-table to train the phrase-based SMT system.

We assume that we have a French-English phrase-table T_{FE} and an English-German phrase-table T_{EG}. From these tables, we construct a French-German phrase-table T_{FG}, which requires estimating four feature functions; phrase translation probabilities for both directions, $\phi(\bar{f}|\bar{g})$ and $\phi(\bar{g}|\bar{f})$ and lexical translation probabilities for both directions,

$p_w(\bar{f}|\bar{g})$ and $p_w(\bar{g}|\bar{f})$, where $\bar{f}$ and $\bar{g}$ are French and German phrases that are parts of phrase translation pairs in T_{FE} and T_{EG}, respectively.[2]

We estimate these probabilities using the probabilities available in T_{FE} and T_{EG} as follows.[3]

$$\phi(\bar{f}|\bar{g}) = \sum_{\bar{e} \in T_{FE} \cap T_{EG}} \phi(\bar{f}|\bar{e})\phi(\bar{e}|\bar{g}) \quad (3)$$

$$\phi(\bar{g}|\bar{f}) = \sum_{\bar{e} \in T_{FE} \cap T_{EG}} \phi(\bar{g}|\bar{e})\phi(\bar{e}|\bar{f}) \quad (4)$$

$$p_w(\bar{f}|\bar{g}) = \sum_{\bar{e} \in T_{FE} \cap T_{EG}} p_w(\bar{f}|\bar{e})p_w(\bar{e}|\bar{g}) \quad (5)$$

$$p_w(\bar{g}|\bar{f}) = \sum_{\bar{e} \in T_{FE} \cap T_{EG}} p_w(\bar{g}|\bar{e})p_w(\bar{e}|\bar{f}) \quad (6)$$

where $\bar{e} \in T_{FE} \cap T_{EG}$ means that the English phrase $\bar{e}$ is included in both T_{FE} and T_{EG} as part of phrase translation pairs. $\phi(\bar{f}|\bar{e})$ and $\phi(\bar{e}|\bar{f})$ are phrase translation probabilities for T_{FE} and $\phi(\bar{e}|\bar{g})$ and $\phi(\bar{g}|\bar{e})$ are those for T_{EG}. $p_w(\bar{f}|\bar{e})$ and $p_w(\bar{e}|\bar{f})$ are lexical translation probabilities for T_{FE} and $p_w(\bar{e}|\bar{g})$ and $p_w(\bar{g}|\bar{e})$ are those for T_{EG}.

The definitions of the phrase and lexical translation probabilities are as follows (Koehn et al., 2003).

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}'} \text{count}(\bar{f}', \bar{e})} \quad (7)$$

where $\text{count}(\bar{f}, \bar{e})$ gives the total number of times the phrase $\bar{f}$ is aligned with the phrase $\bar{e}$ in the parallel corpus. Eq. 7 means that $\phi(\bar{f}|\bar{e})$ is calculated using maximum likelihood estimation.

The definition of the lexical translation probability is

$$p_w(\bar{f}|\bar{e}) = \max_{\mathbf{a}} p_w(\bar{f}|\bar{e}, \mathbf{a}) \quad (8)$$

$$p_w(\bar{f}|\bar{e}, \mathbf{a}) = \prod_{i=1}^{n} E_w(f_i|\bar{e}, \mathbf{a}) \quad (9)$$

$$E_w(f_i|\bar{e}, \mathbf{a}) = \frac{1}{|\{j|(i,j) \in \mathbf{a}\}|} \sum_{\forall (i,j) \in \mathbf{a}} w(f_i|e_j) \quad (10)$$

[2]Feature functions scores are calculated using these probabilities. For example, for a translation probability of a French sentence $\mathbf{f} = \bar{f}_1 \dots \bar{f}_K$ and a German sentence $\mathbf{g} = \bar{g}_1 \dots \bar{g}_K$, $h(\mathbf{g}, \mathbf{f}) = \log \prod_{i=1}^{K} \phi(\bar{f}_i|\bar{g}_i)$, where K is the number of phrases.

[3]Wang et al. (2006) use essentially the same definition to induce the translation probability of the source and target language word alignment that is bridged by an intermediate language. Callison-Burch et al. (2006a) use a similar definition for a paraphrase probability.

$$w(f|e) = \frac{\text{count}(f, e)}{\sum_{f'} \text{count}(f', e)} \quad (11)$$

where $\text{count}(f, e)$ gives the total number of times the word f is aligned with the word e in the parallel corpus. Thus, $w(f|e)$ is the maximum likelihood estimation of the word translation probability of f given e. $Ew(f_i|\bar{e}, \mathbf{a})$ is calculated from a word alignment $\mathbf{a}$ between a phrase pair $\bar{f} = f_1 f_2 \dots f_n$ and $\bar{e} = e_1 e_2 \dots e_m$ where f_i is connected to several ($|\{j|(i,j) \in \mathbf{a}\}|$) English words. Thus, $Ew(f_i|\bar{e}, \mathbf{a})$ is the average (or mixture) of $w(f_i|e_j)$. This means that $Ew(f_i|\bar{e}, \mathbf{a})$ is an estimation of the probability of f_i in $\mathbf{a}$. Consequently, $p_w(\bar{f}|\bar{e}, \mathbf{a})$ estimates the probability of $\bar{f}$ given $\bar{e}$ and $\mathbf{a}$ using the product of the probabilities $Ew(f_i|\bar{e}, \mathbf{a})$. This assumes that the probability of f_i is independent given $\bar{e}$ and $\mathbf{a}$. $p_w(\bar{f}|\bar{e})$ takes the highest $p_w(\bar{f}|\bar{e}, \mathbf{a})$ if there are multiple alignments $\mathbf{a}$. This discussion, which is partly based on Section 4.1.2 of (Och and Ney, 2004), means that the lexical translation probability $p_w(\bar{f}|\bar{e})$ is another probability estimated using the word translation probability $w(f|e)$.

The justification of Eqs. 3–6 is straightforward. From the discussion above, we know that the probabilities, $\phi(\bar{f}|\bar{e})$, $\phi(\bar{e}|\bar{f})$, $\phi(\bar{g}|\bar{e})$, $\phi(\bar{e}|\bar{g})$, $p_w(\bar{f}|\bar{e})$, $p_w(\bar{e}|\bar{f})$, $p_w(\bar{g}|\bar{e})$, and $p_w(\bar{e}|\bar{g})$ are probabilities in the ordinary sense. Thus, we can derive $\phi(\bar{f}|\bar{g})$, $\phi(\bar{g}|\bar{f})$, $p_w(\bar{f}|\bar{g})$, and $p_w(\bar{g}|\bar{f})$ by assuming that these probabilities are independent given an English phrase $\bar{e}$ (e.g., $\phi(\bar{f}|\bar{g}, \bar{e}) = \phi(\bar{f}|\bar{e})$).

We construct a T_{FG} that consists of all French-German phrases whose phrase and lexical translation probabilities as defined in Eqs. 3–6 are greater than 0. We use the term *PhraseTrans* to denote SMT systems that use the phrase translation strategy described above.

3.2 Sentence translation strategy

The sentence translation strategy uses two independently trained SMT systems. We first translate a French sentence $\mathbf{f}$ into n English sentences $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$ using a French-English SMT system. Each $\mathbf{e}_i$ $(i = 1 \dots n)$ has the eight scores calculated from the eight feature functions described in Section 2. We denote these scores $h^e_{i1}, h^e_{i2}, \dots h^e_{i8}$. Second, we translate each $\mathbf{e}_i$ into n German sentences $\mathbf{g}_{i1}, \mathbf{g}_{i2}, \dots, \mathbf{g}_{in}$ using an English-German

SMT system. Each $\mathbf{g}_{ij}$ $(j = 1 \ldots n)$ has the eight scores, which are denoted as $h^g_{ij1}, h^g_{ij2}, \ldots, h^g_{ij8}$. This situation is depicted as

$$\begin{array}{rcll} \mathbf{f} & \rightarrow & \mathbf{e}_i & (h^e_{i1}, h^e_{i2}, \ldots, h^e_{i8}) \\ & \rightarrow & \mathbf{g}_{ij} & (h^g_{ij1}, h^g_{ij2}, \ldots, h^g_{ij8}) \end{array}$$

We define the score of $\mathbf{g}_{ij}$, $S(\mathbf{g}_{ij})$, as

$$S(\mathbf{g}_{ij}) = \sum_{m=1}^{8} (\lambda^e_m h^e_{im} + \lambda^g_m h^g_{ijm}) \qquad (12)$$

where λ^e_m and λ^g_m are weights set by performing minimum error rate training[4] as described in Section 2. We select the highest scoring German sentence

$$\hat{\mathbf{g}} = \arg\max_{\mathbf{g}_{ij}} S(\mathbf{g}_{ij}) \qquad (13)$$

as the translation of the French sentence $\mathbf{f}$.

A drawback of this strategy is that translation speed is about $O(n)$ times slower than those of the component SMT systems. This is because we have to run the English-German SMT system n times for a French sentence. Consequently, we cannot set n very high. When we used $n = 15$ in the experiments described in Section 4, it took more than two days to translate 3064 test sentences on a 3.06GHz LINUX machine.

Note that when $n = 1$, the above strategy produces the same translation with the simple sequential method that we first translate a French sentence into an English sentence and then translate that sentence into a German sentence.

We use the terms *SntTrans15* and *SntTrans1* to denote SMT systems that use the sentence translation strategy with $n = 15$ and $n = 1$, respectively.

4 Experiments

We conducted controlled experiments using the Europarl corpus. For each language pair described below, the Europarl corpus provides three types of parallel corpora; the source language–English, English–the target language, and the source language–the target language. This means that we can directly train an SMT system using the source and target language parallel corpus as well as pivot SMT systems using English as the pivot language. We use the term *Direct* to denote directly trained SMT systems. For each language pair, we compare four SMT systems; *Direct*, *PhraseTrans*, *SntTrans15*, and *SntTrans1*.[5]

4.1 Training, tuning and testing SMT systems

We used the training data for the shared task of the SMT workshop (Koehn and Monz, 2006) to train our SMT systems. It consists of three parallel corpora: French-English, Spanish-English, and German-English.

We used these three corpora to extract a set of sentences that were aligned to each other across all four languages. For that purpose, we used English as the pivot. For each distinct English sentence, we extracted the corresponding French, Spanish, and German sentences. When an English sentence occurred multiple times, we extracted the most frequent translation. For example, because "Resumption of the session" was translated into "Wiederaufnahme der Sitzungsperiode" 120 times and "Wiederaufnahme der Sitzung" once, we extracted "Wiederaufnahme der Sitzungsperiode" as its translation. Consequently, we extracted 585,830 sentences for each language. From these corpora, we constructed the training parallel corpora for all language pairs.

We followed the instruction of the shared task baseline system to train our SMT systems.[6] We used the trigram language models provided with the shared task. We did minimum error rate training on the first 500 sentences in the shared task development data to tune our SMT systems and used the

[4] We use a reranking strategy for the sentence translation strategy. We first obtain n^2 German sentences for each French sentence by applying two independently trained French-English and English-German SMT systems. Each of the translated German sentences has the sixteen scores as described above. The weights in Eq. 12 are tuned against reference German sentences by performing minimum error rate training. These weights are in general different from those of the original French-English and English-German SMT systems.

[5] As discussed in the introduction, we intend to use the pivot strategies in a situation where a very limited amount of parallel text is available. The use of the Europarl corpus is not an accurate simulation of the intended situation because it enables us to use a relatively large parallel corpus for direct training. However, it is necessary to evaluate the performance of the pivot strategies against that of *Direct* SMT systems under controlled experiments in order to determine how much the pivot strategies can be improved. This is a first step toward the use of pivot methods in situations where training data is extremely scarce.

[6] The parameters for the Pharaoh decoder were "-dl 4 -b 0.03 -s 100". The maximum phrase length was 7.

3064 test sentences for each language as our test set.

Our evaluation metric was %BLEU scores, as calculated by the script provided along with the shared task.[7] We lowercased the training, development and test sentences.

4.2 Results

Table 1 compares the BLEU scores of the four SMT systems; *Direct*, *PhraseTrans*, *SntTrans15*, and *SntTrans1* for each language pair. The columns SE and ET list the BLEU scores of the *Direct* SMT systems trained on the source language–English and English–the target language parallel corpora. The numbers in the parentheses are the relative scores of the pivot SMT systems, which were obtained by dividing their BLEU scores by that of the corresponding *Direct* system. For example, for the Spanish–French language pair, the BLEU score of the *Direct* SMT system was 35.78, that of the *PhraseTrans* SMT system was 32.90, and the relative performance was $0.92 = (32.90/35.78)$. For the *SntTrans15* SMT system, the BLEU score was 29.49 and the relative performance was $0.82 = (29.49/35.78)$.

The BLEU scores of the *Direct* SMT systems were higher than those of the *PhraseTrans* SMT systems for all six source-target language pairs. The *PhraseTrans* SMT systems performed better than the *SntTrans15* SMT systems for all pairs. The *SntTrans15* SMT systems were better than the *SntTrans1* SMT systems for four pairs. According to the sign test, under the null hypothesis that the BLEU scores of two systems are equivalent, finding one system obtaining better BLEU scores on all six language pairs is statistically significant at the 5 % level. Obtaining four better scores is not statistically significant. Thus, Table 1 indicates

$$\textit{Direct} > \textit{PhraseTrans} > \textit{SntTrans15} \sim \textit{SntTrans1}$$

where "$>$" and "$\sim$" means that the differences of the BLEU scores of the corresponding SMT systems are statistically significant and insignificant, respectively.

[7] Callison-Burch et al. (2006b) show that in general a higher BLEU score is not necessarily indicative of better translation quality. However, they also suggest that the use of BLEU is appropriate for comparing systems that use similar translation strategies, which is the case with our experiments.

As expected, the *Direct* SMT systems outperformed the other systems. We regard the BLEU scores of the *Direct* systems as the upperbound. The *SntTrans15* SMT systems did not significantly outperform the *SntTrans1* SMT systems. We think that this is because $n = 15$ was not large enough to cover good translation candidates.[8] Selecting the highest scoring translation from a small pool did not always lead to better performance. To improve the performance of the sentence translation strategy, we need to use a large n. However, this is not practical because of the slow translation speed, as discussed in Section 3.2.

The *PhraseTrans* SMT systems significantly outperformed the *SntTrans15* and *SntTrans1* systems. That is, the phrase translation strategy is better than the sentence translation strategy. Since the phrase-tables constructed using the phrase translation strategy can be integrated into the Pharaoh decoder as well as the directly extracted phrase-tables, the *PhraseTrans* SMT systems can fully exploit the power of the decoder. This led to better performance even when the induced phrase-tables were noisy, as described below.

The relative performance of the *PhraseTrans* SMT systems compared to the *Direct* SMT systems was 0.92 to 0.97. These are very promising results. To show how these systems translated the test sentences, we translated some outputs of the Spanish-French *Direct* and *PhraseTrans* SMT systems into English using the French-English *Direct* system. These are shown in Table 3 with the reference English sentences.

The relative performance seems to be related to the BLEU scores for the *Direct* SMT systems. It was relatively high (0.95 to 0.97) for the difficult (in terms of BLEU) language pairs but relatively low (0.92) for the easy language pairs; Spanish–French and French–Spanish. There is a lot of room for improvement for the relatively easy language pairs. This relationship is stronger than the relationship between the BLEU scores for SE/ET and those for the *PhraseTrans* systems, where no clear trend exists.

Table 2 shows the number of phrases stored in the phrase-tables. The *Direct* SMT systems had 7.3 to

[8] A typical reranking approach to SMT (Och et al., 2004) uses a 1000–best list.

Source–Target	*Direct*		*PhraseTrans*		*SntTrans15*		*SntTrans1*	SE	ET
Spanish–French	35.78	>	32.90 (0.92)	>	29.49 (0.82)	>	29.16 (0.81)	29.31	28.80
French–Spanish	34.16	>	31.49 (0.92)	>	28.41 (0.83)	>	27.99 (0.82)	27.59	29.07
German–French	23.37	>	22.47 (0.96)	>	22.03 (0.94)	>	21.64 (0.93)	22.40	28.80
French–German	15.27	>	14.51 (0.95)	>	14.03 (0.92)	<	14.21 (0.93)	27.59	15.81
German–Spanish	22.34	>	21.76 (0.97)	>	21.36 (0.96)	>	20.97 (0.94)	22.40	29.07
Spanish–German	15.50	>	15.11 (0.97)	>	14.46 (0.93)	<	14.61 (0.94)	29.31	15.81

Table 1: BLEU scores and relative performance

	No. of phrases ("M" means 10^6)			R	P
	Direct	*PhraseTrans*	common		
S–F	18.2M	190.8M	6.3M	34.7	3.3
F–S	18.2M	186.8M	6.3M	34.7	3.4
G–F	7.3M	174.9M	3.1M	43.2	1.8
F–G	7.3M	168.2M	3.1M	43.2	1.9
G–S	7.5M	179.6M	3.3M	44.1	1.9
S–G	7.6M	176.6M	3.3M	44.1	1.9

"S", "F", and "G" are the acronyms of Spanish, French, and German, respectively. "X–Y" means that "X" is the source language and "Y" is the target language.

Table 2: Statistics for the phrase-tables

18.2 million phrases, and the *PhraseTrans* systems had 168.2 to 190.8 million phrases. The numbers of phrases stored in the *PhraseTrans* systems were very large compared to those of *Direct* systems.[9] However, this does not cause a computational problem in decoding because those phrases that do not appear in source sentences are filtered so that only the relevant phrases are used during decoding.

The figures in the *common* column are the number of phrases common to the *Direct* and *PhraseTrans* systems. R (recall) and P (precision) are defined as follows.

$$\text{R} = \frac{\text{No. of common phrases} \times 100}{\text{No. of phrases in } \textit{Direct} \text{ system}}$$

$$\text{P} = \frac{\text{No. of common phrases} \times 100}{\text{No. of phrases in } \textit{PhraseTrans} \text{ system}}$$

Recall was reasonably high. However, the upper bound of recall was 100 percent because we used a multilingual corpus whose sentences were aligned to each other across all four languages, as described in Section 4.1. Thus, there is a lot of room for improvement with respect to recall. Precision, on the other hand, was very low. However, translation performance was not significantly affected by this low precision, as is shown in Table 1. This indicates that recall is more important than precision in building phrase-tables.

5 Related work

Pivot languages have been used in rule-based machine translation systems. Boitet (1988) discusses the pros and cons of the pivot approaches in multilingual machine translation. Schubert (1988) argues that a pivot language needs to be a natural language, due to the inherent lack of expressiveness of artificial languages.

Pivot-based methods have also been used in other related areas, such as translation lexicon induction (Schafer and Yarowsky, 2002), word alignment (Wang et al., 2006), and cross language information retrieval (Gollins and Sanderson, 2001). The translation disambiguation techniques used in these studies could be used for improving the quality of phrase translation tables.

In contrast to these, very little work has been done on pivot-based methods for SMT. Kauers et al. (2002) used an artificial interlingua for spoken language translation. Gispert and Mariño (2006) created an English-Catalan parallel corpus by automatically translating the Spanish part of an English-Spanish parallel corpus into Catalan with a Spanish-Catalan SMT system. They then directly trained an SMT system on the English-Catalan corpus. They

[9] In Table 2, the *PhraseTrans* systems have more than 10x as many phrases as the *Direct* systems. This can be explained as follows. Let f_i be the *fanout* of an English phrase i, i.e., f_i is the number of phrase pairs containing the English phrase i in a phrase-table, then the size of the phrase-table is $s_1 = \sum_{i=1}^{n} f_i$, where n is the number of distinct English phrases. When we combine two phrase-tables, the size of the combined phrase table is roughly $s_2 = \sum_{i=1}^{n} f_i^2$. Thus, the relative size of the combined phrase table is roughly $r = \frac{s_2}{s_1} = \frac{E(f^2)}{E(f)}$, where $E(f) = \frac{s_1}{n}$ and $E(f^2) = \frac{s_2}{n}$ are the averages over f_i and f_i^2, respectively. As an example, we calculated these averages for the German-English phrase table. $E(f)$ was 1.5, $E(f^2)$ was 43.7, and r was 28.9. This shows that even if an average fanout is small, the size of a combined phrase table can be very large.

showed that this direct training method is superior to the sentence translation strategy (*SntTrans1*) in translating Catalan into English but is inferior to it in the opposite translation direction (in terms of the BLEU score). In contrast, we have shown that the phrase translation strategy consistently outperformed the sentence translation strategy in the controlled experiments.

6 Conclusion

We have compared two types of pivot strategies, namely *phrase translation* and *sentence translation*. The phrase translation strategy directly constructs a phrase translation table from a source language and English phrase-table and a target language and English phrase-table. It then uses this phrase table in a phrase-based SMT system. The sentence translation strategy first translates a source language sentence into n English sentences and translates these n sentences into target language sentences separately. Then, it selects the highest scoring sentence from the target language sentences.

We conducted controlled experiments using the Europarl corpus to compare the performance of these two strategies to that of directly trained SMT systems. The experiments showed that the performance of the phrase translation strategy was statistically significantly better than that of the sentence translation strategy and that its relative performance compared to the directly trained SMT systems was 0.92 to 0.97. These are very promising results.

Although we used the Europarl corpus for controlled experiments, we intend to use the pivot strategies in situations where very limited amount of parallel corpora are available for a source and target language but where relatively large parallel corpora are available for the source language–English and the target language–English. In future work, we will further investigate the pivot strategies described in this paper to confirm that the phrase translation strategy is better than the sentence translation strategy in the intended situation as well as with the Europarl corpus.[10]

[10] As a first step towards real situations, we conducted additional experiments. We divided the training corpora in Section 4 into two halves. We used the first 292915 sentences to train source-English SMT systems and the remaining 292915 ones to train English-target SMT systems. Based on these source-English and English-target SMT systems, we trained *PhraseTrans* and *SntTrans1* SMT systems. Other experimental conditions were the same as those described in Section 4. The table below shows the BLUE scores of these SMT systems. It indicates that the *PhraseTrans* systems consistently outperformed the *SntTrans1* systems.

Source-Target	*PhraseTrans*	*SntTrans1*
Spanish-French	31.57	28.36
French-Spanish	30.18	27.75
German-French	20.48	19.83
French-German	14.38	14.11
German-Spanish	19.58	18.67
Spanish-German	14.80	14.46

References

Christian Boitet. 1988. Pros and cons of the pivot and transfer approaches in multilingual machine translation. In Dan Maxwell, Klaus Schubert, and Toon Witkam, editors, *New Directions in Machine Translation*. Foris. (appeared in Sergei Nirenburg, Harold Somers and Yorick Wilks (eds.) *Readings in Machine Translation* published by the MIT Press in 2003).

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):263–311.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006a. Improved statistical machine translation using paraphrases. In *NAACL*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006b. Re-evaluating the role of BLEU in machine translation research. In *EACL*.

Adriá de Gispert and José B. Mari no. 2006. Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish. In *Proc. of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages*.

Tim Gollins and Mark Sanderson. 2001. Improving cross language information retrieval with triangulated translation. In *SIGIR*.

Manuel Kauers, Stephan Vogel, Christian Fügen, and Alex Waibel. 2002. Interlingua based statistical machine translation. In *ICSLP*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL*.

Ref	i hope with all my heart , and i must say this quite emphatically , that an opportunity will arise when this document can be incorporated into the treaties at some point in the future .
Dir	i hope with conviction , and put great emphasis , that again is a serious possibility of including this in the treaties .
PT	i hope with conviction , and i very much , insisted that never be a serious possibility of including this in the treaties .
Ref	should this fail to materialise , we should not be surprised if public opinion proves sceptical about europe , or even rejects it .
Dir	otherwise , we must not be surprised by the scepticism , even the rejection of europe in the public .
PT	otherwise , we must not be surprised by the scepticism , and even the rejection of europe in the public .
Ref	the intergovernmental conference - to address a third subject - on the reform of the european institutions is also of decisive significance for us in parliament .
Dir	the intergovernmental conference - and this i turn to the third issue on the reform of the european institutions is of enormous importance for the european parliament .
PT	the intergovernmental conference - and this brings me to the third issue - on the reform of the european institutions has enormous importance for the european parliament .

Table 3: Reference sentences (Ref) and the English translations (by the French-English *Direct* system) of the outputs of the Spanish-French *Direct* and *PhraseTrans* SMT systems (Dir and PT).

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *IWSLT*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *AMTA*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Makoto Nagao. 1981. A framework of a mechanical translation between Japanese and English by analogy principle. In *the International NATO Symposium on Artificial and Human Intelligence*. (appeared in Sergei Nirenburg, Harold Somers and Yorick Wilks (eds.) *Readings in Machine Translation* published by the MIT Press in 2003).

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *HLT-NAACL*.

Franz Josef Och. 1999. An efficient method for determining bilingual word classes. In *EACL*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

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

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*.

Klaus Schubert. 1988. Implicitness as a guiding principle in machine translation. In *COLING*.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *ICSLP*.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL*, pages 72–79.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *COLING/ACL 2006 Main Conference Poster Sessions*.