Phrase-Based Statistical Machine Translation with Pivot Languages

Nicola Bertoldi, Madalina Barbaiani†, Marcello Federico, Roldano Cattoni

FBK-irst - Ricerca Scientifica e Tecnologica
Via Sommarive 18, 38100 Povo (TN), Italy
{bertoldi, federico, cattoni}@fbk.eu

† Research Group on Mathematical Linguistics, Rovira i Virgili University
Pl. Imperial Tárraco 1, Tarragona 43005, Spain
madalina.barbaiani@estudiants.urv.cat

Abstract

Translation with pivot languages has recently gained attention as a means to circumvent the data bottleneck of statistical machine translation (SMT). This paper tries to give a mathematically sound formulation of the various approaches presented in the literature and introduces new methods for training alignment models through pivot languages. We present experimental results on Chinese-Spanish translation via English, on a popular traveling domain task. In contrast to previous literature, we report experimental results by using parallel corpora that are either disjoint or overlapped on the pivot language side. Finally, our original method for generating training data through random sampling shows to perform as well as the best methods based on the coupling of translation systems.

1. Introduction

Statistical machine translation (SMT) is concerned with the machine learning task of designing and developing statistical models and algorithms to translate texts from a source language F into a target language E. Training algorithms for SMT generally rely on a large sample of human translations between F and E. This paradigm has proven to be successful for language pairs for which large parallel corpora are available, such as Chinese-English, Arabic-English and French-English. The largest collections of parallel texts typically come from national and international organizations that publish multilingual documents, e.g. the United Nations, European Parliament, Canadian Parliament, news agencies, etc. Unfortunately, there are many relevant language pairs for which such fundamental language resources are available only to a limited extent.

To circumvent the data bottleneck, research on SMT has been recently investigating the use of so-called *pivot* or *bridge* languages. The assumptions are simple to state: (i) there is lack of parallel texts between E and F, while (ii) there exists a language G for which the are abundant parallel texts between F and G and between G and E.

A realistic working condition with pivot languages is that the parallel corpora for F-G and G-E are *independent*, in the sense that they do not derive from the same set of sentences. Recent research has often focused indeed on the use of parallel corpora, such as the Europarl Corpus, which provides instead multiple translations of the same texts. While such data can be regarded as interesting for the sake of performing contrastive experiments, namely to compare translations obtained with and without bridge languages, they do not reflect the general case and results should be interpreted carefully.

This paper presents a theoretical formulation of SMT with pivot languages, that embraces several approaches in the literature and a few original methods. Extensive experiments are reported that compare performance of each bridging when using dependent and independent parallel data. Experiments were conducted on the IWSLT 2008 benchmark, namely the translation of traveling domain expressions from Chinese to Spanish via English.

2. Previous Work

The use of *pivot* or *bridge* languages has been advocated for different purposes, such as rule-based machine translation systems [1], translation lexicon induction [2, 3], word alignment [4, 5, 6], cross language information retrieval [7].

Concerning statistical machine translation, pivot language translation has been investigated for instance by [8] in order to extend an interlingua based speech translation system to a new language. In [9], Catalan-English translation is bridged through Spanish. The authors compared two coupling strategies: cascading of two translation systems versus training of system from parallel texts whose target part has been automatically translated from pivot to target. System cascading was recently investigated in [10], too.

In [11] word alignment systems are combined from multiple bridge languages by multiplying posterior probability matrices. This technique requires the existence of parallel

data for several languages, like the proceedings of United Nations or European Parliament.

An approach based on phrase table multiplication is discussed in [10, 12]. Scores of the new phrase table are computed by combining corresponding translation probabilities in the source-pivot and pivot-target phrase-tables. Finally, in [13] a similar approach is described, but for the sake of improving translation probabilities through triangulation with other languages.

3. SMT through pivot languages

SMT with bridge languages is concerned about how to optimally perform translation from F to E, by taking advantage of the available language resources. We can device two general approaches to apply bridge languages in SMT, namely bridging at translation time or bridging at training time, which we briefly overview now.

3.1. Bridging at Translation Time

Under this framework, we try to integrate or couple two levels of translation within the same decoding problem:

$$\begin{array}{ccccc} \text{source text} & & \text{pivot text} & & \text{target text} \\ \mathbf{f} & \rightarrow & \mathbf{g} & \rightarrow & \mathbf{e} \end{array}$$

namely, from the source text to the pivot text, and from the pivot text to the target text:

The corresponding statistical decision criterion can be derived by modeling the pivot text as a hidden variable and by assuming independence between the target and the source strings, given the pivot string:

$$\begin{aligned} \mathbf{f} \rightarrow \hat{\mathbf{e}} &= \operatorname*{argmax}_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f}) \\ &= \operatorname*{argmax}_{\mathbf{e}} \sum_{\mathbf{g}} p(\mathbf{e}, \mathbf{g} \mid \mathbf{f}) \\ &= \operatorname*{argmax}_{\mathbf{e}} \sum_{\mathbf{g}} p(\mathbf{g} \mid \mathbf{f})\, p(\mathbf{e} \mid \mathbf{g}) \\ &\approx \operatorname*{argmax}_{\mathbf{e}} \max_{\mathbf{g}} p(\mathbf{g} \mid \mathbf{f})\, p(\mathbf{e} \mid \mathbf{g}) \end{aligned} \quad (1)$$

Notice that in the last step we also apply the usual max approximation, to reduce the complexity of the search procedure. By assuming standard phrase-based models for each of the probability expressions in the right-hand side of the last equation, we have to extend the search with other two hidden variables: $\mathbf{a}$, $\mathbf{b}$ as follows:

$$\operatorname*{argmax}_{\mathbf{e},\mathbf{a}} \max_{\mathbf{g},\mathbf{b}} p(\mathbf{g}, \mathbf{b} \mid \mathbf{f})\, p(\mathbf{e}, \mathbf{a} \mid \mathbf{g}) \quad (2)$$

The two variables $\mathbf{a}$ and $\mathbf{b}$, respectively, model phrase segmentation and re-ordering for each considered translation direction. Figure 1 shows the two level alignments for a simple example involving translations from Chinese to Italian, through English. Horizontal segments show that the English string is segmented differently when it is generated from Chinese than when it is translated into Italian.

3.1.1. Coupling Independent Alignments

From a computational complexity view, the two level translation problem is at least as hard as the Spoken Language Translation (SLT) problem [14]. Briefly, in SLT source strings are sequences of acoustic observations $\mathbf{x}$, pivot strings are transcription hypotheses $\mathbf{f}$ of $\mathbf{x}$, and the target strings are translations $\mathbf{e}$ of $\mathbf{f}$. By taking advantage of approximations proposed for the SLT case, we can reduce the computational burden of (2) by limiting the pivot translations $\mathbf{g}$ to a limited subset $\mathcal{G}(\mathbf{f})$:

$$\operatorname*{argmax}_{\mathbf{e},\mathbf{a}} p(\mathbf{e}, \mathbf{a} \mid \mathbf{g}) \max_{(\mathbf{g},\mathbf{b}) \in \mathcal{G}(\mathbf{f})} p(\mathbf{g}, \mathbf{b} \mid \mathbf{f}) \quad (3)$$

Natural candidates to represent such subsets of pivot translations are n-best lists and word-graphs produced by the source-to-pivot translation engine.

3.1.2. Coupling Constrained Alignments

Besides limiting the translation candidates $\mathbf{g}$, another alternative proposed in the literature is to constrain the alignments $\mathbf{a}$ and $\mathbf{b}$ to share exactly the same segmentation, and $\mathbf{b}$ to be monotonic. An example of the effect of these constraints is shown in Figure 1. With these restrictions, search can be carried out in a single step by pre-computing the product of the involved phrase-tables. Assuming two phrase tables with entries $(\tilde{f}, \tilde{g})$ and $(\tilde{g}, \tilde{e})$, and scores $t(\tilde{f}, \tilde{g})$ and $t(\tilde{g}, \tilde{e})$, respectively, we can build a new phrase table with entries $(\tilde{f}, \tilde{e})$ and scores computed according to one of the following criteria:

- Integration:

$$t(\tilde{e} \mid \tilde{f}) = \sum_{\tilde{g}} t(\tilde{f}, \tilde{g}) \times t(\tilde{g}, \tilde{e}) \quad (4)$$

- Maximization:

$$t(\tilde{e} \mid \tilde{f}) = \max_{\tilde{g}}\ t(\tilde{f}, \tilde{g}) \times t(\tilde{g}, \tilde{e}) \quad (5)$$

Search with constrained alignments requires the target language model and a single distortion model that directly maps source to target positions. At first sight, it seems rather difficult to compute the distortion model by combining the distortion models of the two translation steps. As there are no parallel data to train a lexicalized distortion model, we opted for a plain exponential distortion model.

3.2. Bridging at Training Time

Another way to exploit parallel training corpora F-G and G-E is to use them to develop and train a translation system from F to E.

3.2.1. Bridging Alignment Models

We will focus here on possible extension of the standard training criterion of IBM alignment models [15], which assumes a parallel corpus $(F, E) = \{(\mathbf{f}_i, \mathbf{e}_i)\}$ and looks for

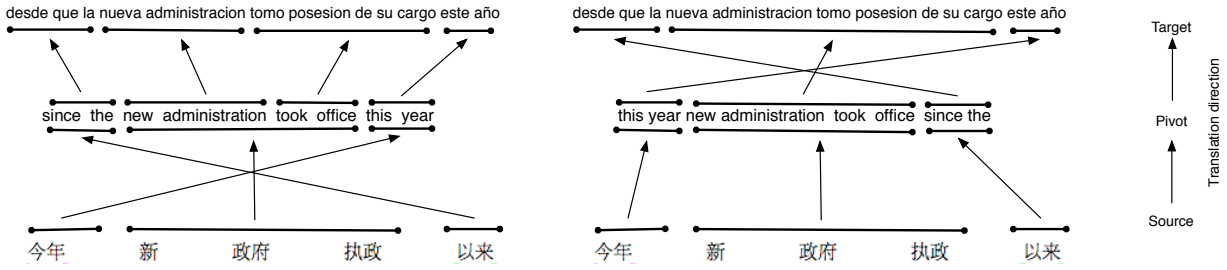

Figure 1: Phrase-based translation from Chinese to Spanish, through English, with independent alignments (left) and constrained alignments (right).

parameters maximizing

$$\theta^*_{FE} = \underset{\theta_{FE}}{\operatorname{argmax}} \prod_i P_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_i) \tag{6}$$

Let us assume instead the availability of a parallel corpus $(F, G) = \{(\mathbf{f}_i, \mathbf{g}_i)\}$ and of an already trained translation system from G to E, that models the posterior probability $P(\mathbf{e} \mid \mathbf{g})$. The probability $P(\mathbf{f} \mid \mathbf{g})$ can be written as the marginal distribution:

$$P(\mathbf{f} \mid \mathbf{g}) = \sum_{\mathbf{e}} P(\mathbf{f} \mid \mathbf{e}) P(\mathbf{e} \mid \mathbf{g}) \tag{7}$$

were we assume independence between the target and source strings, given the pivot string. If $\tilde{P}(\mathbf{e} \mid \mathbf{g})$ is the given translation model and $P_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_i)$ is an alignment model to be estimated, the following training criterion can be applied:

$$\theta^*_{FE} = \underset{\theta_{FE}}{\operatorname{argmax}} \prod_i \sum_{\mathbf{e}_i} P_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_i) \tilde{P}(\mathbf{e}_i \mid \mathbf{g}_i) \tag{8}$$

In the new formulation $\mathbf{e}_i$ becomes now a hidden random variable generated by the translation system $\tilde{P}(\mathbf{e} \mid \mathbf{g})$, trained on a parallel corpus (G, E). Notice that our formulation leaves us free about the way the latter model is built. Namely, we will assume $\tilde{P}(\mathbf{e} \mid \mathbf{g})$ to be the best phrase-based model we can develop from training data (G, E).

The concern is now how to efficiently estimate the word alignment models with the criterion (8). A first reasonable approximation is to limit the sum over $\mathbf{e}_i$ to the n-best hypotheses, or equivalently to limit the support of $\tilde{P}(\mathbf{e} \mid \mathbf{g})$ to the top n translations. While we leave for future work the extension of the standard training algorithms, at least for the simplest IBM alignment models, in this work we only experimented the simplest case, namely using the 1-best hypothesis (Viterbi) $\mathbf{e}_i$, which trivially results in the criterion (6) with $\mathbf{e}_i$ obtained by taking the top best translation from $\tilde{P}(\mathbf{e}_i \mid \mathbf{g}_i)$.

3.2.2. Random Sampling of Training Data

Another simple method we investigate is to generate a parallel corpus by sampling from an available translation system from G to E. The idea is simple. For each example $(\mathbf{f}_i, \mathbf{g}_i)$ in the training corpus (F, G) we generate a random sample of m translations $\mathbf{e}_{ij}$ of $\mathbf{g}_i$ according to the distribution $\tilde{P}(\mathbf{e} \mid \mathbf{g})$. Given the newly created sample $(F, E) = \{(\mathbf{f}_i, \mathbf{e}_{ij})\}, j = 1, \ldots, k$, we build a translation system from word alignments estimated by maximizing the criterion:

$$\theta^*_{FE} = \underset{\theta_{FE}}{\operatorname{argmax}} \prod_{i,j} P_{\theta_{FE}}(\mathbf{f}_i \mid \mathbf{e}_{ij}) \tag{9}$$

Practically, we generate n-best list of translations $\mathbf{e}$ from $\mathbf{g}_i$ and normalize their translation scores in order to define the posterior $\tilde{P}(\mathbf{e} \mid \mathbf{g}_i)$. Then, we sample with replacement k alternatives from this list according to the posterior distribution. The idea is to get a sample that contains possible duplicates of the most probable translations. In this way, most reliable word alignments are reinforced during training as well as phrase-pairs using words of the most probable translations.

This approach is indeed more sound than just taking the list of n-best, as experimental results will confirm in the following sections.

4. Task description

The approaches introduced in the previous section were evaluated on a benchmark provided by the 2008 International Workshop on Spoken Language Translation[1]. One of the proposed tasks consists in translating from Chinese to Spanish by pivoting through English. Training and evaluation data are from the Basic Travel Expression Corpus (BTEC) [16], a collection of parallel translations in the traveling domain.

Five monolingual corpora were available: two for Chinese (C1 and C2), two for English (E1 and E2) and one for Spanish (S1). C1, E1, and S1 are also aligned at the sentence level, hence they provide a trilingual parallel corpus; C2 and E2 are aligned as well and form a bilingual parallel corpus. The official benchmark for the pivot task of IWSLT consists only in the two non overlapping bilingual parallel corpora CE2 and ES1, while the bilingual parallel corpus CS1 is for contrastive experiments. The benchmark also includes a development set of 506 Chinese sentences, with 16 alternative translations in English and Spanish. In order to evaluated

[1] www.slc.atr.jp/IWSLT2008/

and compare our approaches we extracted a test set of 998 sentences from the trilingual corpus (CES1).

More details about training, dev and test sets are given in Tables 1. Statistics reported refer to texts after tokenization or segmentation (for Chinese), and converting numbers into digits.

		Chi	Eng	Spa
corpus 1	sentences		18,974	
	words	161K	172K	176K
	dictionary	8,017	8,210	10,773
corpus 2	sentences		18,999	
	words	150K	172K	-
	dictionary	8,114	8,631	-
dev	sentences		506	
	words	3,721	3,769	3,774
	dictionary	935	931	1,053
test	sentences		998	
	words	8,588	9,294	9,445
	dictionary	1,668	1,731	2,090

Table 1: Statistics of the available training data, dev and test sets: number of sentences, number of running words, and dictionary size.

Translation performance is reported in terms of case sensitive BLEU% score. Statistical significance tests are carried out by applying a paired t-test on a 50-fold partition of the test set.

5. System description

All experiments have been run with phrase-based statistical MT systems developed with the Moses open-source toolkit [17]. The employed decoder features a statistical log-linear model including a phrase-based translation model, a language model, a distortion model, and word and phrase penalties. The resulting eight weights of the log-linear combination are optimized by means of a minimum error training procedure [18].

The phrase-based translation model provides direct and inverted frequency-based and lexical-based probabilities for each phrase pair included in a given phrase table. Phrase pairs are extracted from symmetrized word alignments generated by GIZA++ [19]. This extraction method does not apply in the case of pivoting with constrained alignments (see Section 3.1.2) as the phrase table is obtained by taking the product of two existing phrase tables.

A 5-gram word-based LM is estimated on the target side of the parallel corpora using the improved Kneser-Ney smoothing [20]. The distortion model is a standard negative-exponential model.

Table 2 shows the BLEU scores achieved by the baseline systems (*Direct*) on the dev and test sets. It is worth noticing that the system trained on CE1 outperforms the one trained on CE2. The reason is that both dev and test sets seem much closer to the former corpus than to the latter, as shown by the out-of-vocabulary (OOV) rates on the source side.

task	data	dev		test	
		BLEU	OOV	BLEU	OOV
Chi-Eng	CE1	43.08	2.90	26.91	2.00
	CE2	33.22	4.14	19.09	3.80
Eng-Spa	ES1	54.39	1.97	49.13	2.01
Chi-Spa	CS1	31.94	2.90	23.67	2.00

Table 2: Results of *direct* translation systems trained on different language pair corpora.

In the following, we describe how we implemented the pivoting approaches introduced in Section 3.

6. Sentence-level Coupling

In system coupling with unconstrained alignments, we consider two methods for interfacing the CE and ES systems. The easiest method, called *Cascade*, uses only the 1-best English translation $\hat{\mathbf{g}}$ of the Chinese sentence $\mathbf{f}$. The second way, named *Nbest*, consists of generating m-best Spanish translations for each of the n-best English translations $\mathbf{g}_1 \dots \mathbf{g}_n$ generated by the CE system, and rescoring all $n \times m$ hypotheses using both CE and ES translation scores. In this case the subset $\mathcal{G}(\mathbf{f}) = \{\mathbf{g}_1 \dots \mathbf{g}_n\}$.

The CE system has been trained on the CE2 data while the ES system on ES1. Table 3 reports BLEU scores obtained with different settings of n and m. (Notice that for each setting a specific weight optimization was performed.)

We considered two ways of combining the translation scores during re-scoring of the $n \times m$ hypotheses: either a log-linear model of the two global scores of the systems, or a log-linear model of all their 16 features. The second strategy showed to be vastly superior ($\alpha = .02$), and was applied in all subsequent experiments.

Increasing n and m from 1 to 10 gives a statistically significant benefit ($\alpha = .02$). Unfortunately, further increases do not show to pay off on the test set, probably because weight optimization tends to overfit the dev data.

Our phrase-based SMT system can possibly generate identical (or duplicate) translation alternatives with different feature scores, as different phrase segmentations are taken into account. By comparing the generation of $n \times m$-best alternatives with or without duplicates, no definitive choice can be made because performance seems again suffering from the overfitting problem. Hence, we preferred to use duplicates because it results mathematically correct as shown in Eq 3.

7. Phrase-level Coupling

In system coupling with constrained alignments, we compared performance under two different training conditions: namely, the use of two disjoint bilingual corpora or the ex-

n,m	rescoring features	not distinct dev	not distinct test	distinct dev	distinct test
1	-	25.13	16.44	25.13	16.44
10	2	25.28	16.60	24.98	16.75
	16	26.65	17.59	27.00	17.96
20	16	27.18	17.03	27.54	17.51
50	16	27.78	16.96	27.87	17.21
100	16	27.89	17.64	28.55	16.93

Table 3: Results of the *Cascade* and *Nbest* approaches. Either distinct or duplicate $n \times m$-best alternatives are compared.

ploitation of one trilingual parallel corpus. In the former case, we took the product of phrase tables estimated from CE2 and ES1, in the latter the product of phrase tables estimated on CE1 and ES1.

7.1. Description of the Algorithm

The source-pivot and pivot-target phrase tables are sorted lexicographically on the pivot phrases. The algorithm reads in parallel from both tables and matches lines with equal pivot phrases. This is linear in terms of numbers of phrase pairs. The matched lines are then combined by multiplying scores as explained in section 3.1.2 into a preliminary table. As the resulting table could contain duplicates of $(\tilde{f}, \tilde{e})$, we sort it on the phrase pairs and collapse duplicate entries either by summing or maximizing equal entries.

Scores of phrase pairs have been computed according to either the policies proposed. Results (BLEU%) on the test set reported in Table 4 show that integrating scores is significantly better choosing the maximum of them ($\alpha = .01$).

	disjoint	overlap
integration	16.65	23.50
maximization	15.88	22.82

Table 4: Results on the test set achieved by the *PhraseTable* approach with two different policies for generating phrase pair scores.

	CE2	CE1	ES1	product disj	product over
src phr	76K	128K	277K	21K	94K
trg phr	82K	134K	284K	32K	108K
phr pairs	133K	185K	333K	592K	696K
avg trans	1.8	1.4	1.2	28.2	7.4
common	-	-	-	59K	143K

Table 5: Statistics about the original and the product phrase tables when pivot data are either disjoint or overlapped.

Table 5 reports statistics of the original CE and ES phrase tables and of the phrase table generated by multiplication: the number of source phrases, target phrases and phrase pairs, and the average number of translations for each source phrase. Furthermore, for the derived phrases tables the amount of common pivot (English) phrases in both original phrase tables is reported: this figure gives a rough estimate of the overlap between the two original phrase tables, and hence it indirectly measures how much Chinese content can be conveyed into Spanish through English,

In the disjoint training condition only 1/3 of the original Chinese phrase can be translated into Spanish through English. Instead, the average number of translations hugely grows, by significantly increasing the ambiguity of translations and the number of wrong translations. Furthermore, only 59K of the 133K phrase-pairs (44%) in the CE2 table have a match in the ES1 phrase table. In fact, the common pivot phrases are mainly of length 1 (65%) as shown in Figure 2.

When the overlapped training condition is applied, the percentage of common pivot phrases increases to 77%. In this case, missing phrase matches are due to different phrase segmentations on the English side, that result after training separately CE and ES systems.

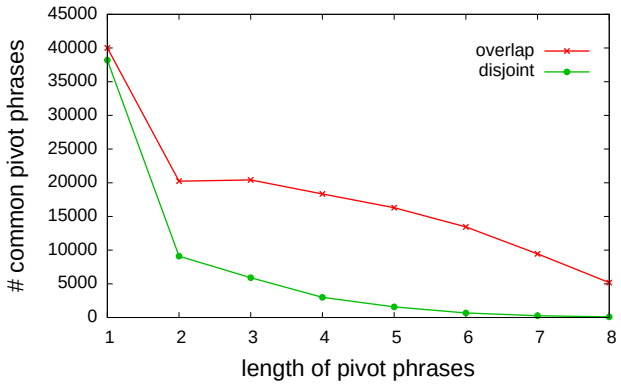

Figure 2: Number of common pivot phrases in the two original phrase tables.

8. Synthesis of Training Data

In order to automatically generate Chinese-Spanish parallel data, we used the system trained on ES1 translate data E2 into a synthetic corpus $\bar{S}2$. In this way we obtain the parallel corpus $C\bar{S}2$ that can be used to directly train a Chinese-Spanish SMT system.

Several ways were investigated to generate the synthetic corpus $C\bar{S}2$. The simplest method is to use the 1-best Spanish translation of each English sentence [9]. The second method, exploits instead the n-best Spanish translations. Chinese sentences are replicated in order to match the number of generated translations. A more theoretically sound method – de-

scribed in Section 3.2.2 is to generate a random sample of size m from the n-best Spanish translations after properly normalizing the translation scores.

All these methods can be seen as a way to perform unsupervised training. Once generated, synthetic parallel corpora can be used to train phrase tables and LMs.

	n,m	lm	dev	test
1-best	1	S1	22.05	14.56
1-best	1	$\bar{\text{S}}2$	23.58	15.38
1-best	1	S1+$\bar{\text{S}}2$	24.57	16.13
n-best	100	S1+$\bar{\text{S}}2$	26.04	17.03
sampling	100	S1+$\bar{\text{S}}2$	26.02	17.68

Table 6: Results of the Synthesis approach using CE2 and ES1 training corpora.

Table 6 reports results of the *Synthesis* approaches using the data available for the pivot condition, the parallel corpora CE2 and ES1. For each setting a specific weight optimization is performed. Concerning the estimation of the target LM, the table shows that using synthetic data $\bar{\text{S}}2$ significantly improves the scores with respect to using the supplied data S1 only ($\alpha = .03$); using both sets gives the best results. Regarding the different methods to generate the parallel corpus, `sampling` shows to outperform the other two methods ($\alpha = .03$). The choice of randomly selecting 100 translations from the 100-best alternatives resulted as a good compromise for the task at hand, which features a rather limited vocabulary and short input sentences.

An explanation of the difference in performance of the various synthesis methods concerns the way data reflect the confidence of the system that generated them. When data consists of 1-best translations, no information is conveyed about the level of confidence of each single translation. In the n-best case, more information about the confidence of the translations is supplied, implicitly. Typically all translations in the n-best list are very similar to each other, hence the most stable portions of them occur more frequently than those for which the system was more uncertain. However, the drawback of the approach is that very low scoring translations receive the same status of the top scoring one. The random sampling approach improves the n-best approach by penalizing the selection of low scoring translations and by generating data which better reflects the confidence of the system.

9. Discussion

For the sake of comparison, main results of the previously presented approaches are summarized in Table 7.

Nbest and *Synthesis* approaches seem to achieve comparable performance and both outperform the *Cascade* and *PhraseTable* methods. Differences are statistically significant at a level $alpha = .05$.

From a computational point of view, the *Nbest* approach is expensive at run-time: it actually translates $n + 1$ times (1 for Chinese-to-English and n for English-to-Spanish) and re-scores and re-ranks $n \times m$ alternatives per input sentence. Instead, the *Synthesis* approach requires more resources for training due to cost of translating the whole English corpus and to compute word alignments over a potentially very large synthetic corpus.

	CS task		CE task
training	disjoint	overlap	overlap
Direct	–	23.67	26.91
Cascade	16.44	24.04	22.36
Nbest	17.64	25.16	23.39
PhraseTable	16.65	23.50	24.01
Synthesis	17.68	25.19	27.58

Table 7: Results on the test set achieved by different pivot approaches to Chinese-Spanish and Chinese-English translation. The *Direct* system is also reported.

As a contrastive condition we also used the CE1 parallel corpus to train the all systems –practically CE1 replaces the corpus CE2. Notice that this training condition assumes that a trilingual parallel corpus CES1 is available.

All systems achieved significantly larger BLEU scores in this contrastive condition. This also confirms that the quality of CE1 is much better than CE2 as already stated in Section 5.

We also run the *Direct* system trained on the CS1 corpus. Interestingly, its score is comparable with that of the *Cascade* and *PhraseTable* systems, but clearly below the score of the *Nbest* and *Synthesis* systems ($\alpha = .02$).

A possible explanation for this behavior is related to the nature of the three involved languages. Translating from Chinese to Spanish requires introducing significant morphology information and word re-ordering. In some sense, pivoting through English results in a nice factorization of the issues: Chinese-English translation copes with most of the word-reordering but little morphology, while English-Spanish translation implies little word re-ordering but more morphology. Probably this factorization has a positive impact in terms of less data sparseness in the training data and results in better statistical models. To provide an evidence to our claim, we performed corresponding experiments in the Chinese-English task pivoting through Spanish, with systems trained on CSE1. In this condition *Direct* significantly outperforms *Cascade* and *PhraseTable* ($\alpha = .01$). Hence, in this case the pivot language does not seem to factor out the complexity of the translation task.

Moreover, on the Chinese-English translation via Spanish *Nbest* performs significantly better than *Cascade* ($\alpha = .02$), but *PhraseTable* outperforms *Cascade* ($\alpha = .01$). It is worth noticing that *Synthesis* significantly outperforms all other approaches, *Direct* system included. A reasonable ex-

planation for this behavior is that *Synthesis* completely skips the most difficult step (translating from Chinese to Spanish) and fully exploits the easiest step (i.e. translating from Spanish to English). This property could be used for generating synthetic Chinese data as well, if parallel data with a language close to Chinese were available.

10. References

[1] K. Schubert, "Implicitness as a guiding principle in machine translation," in *Proc. of COLING*, Budapest, Hungary, 1988, pp. 599–601.

[2] G. S. Mann and D. Yarowsky, "Multipath translation lexicon induction via bridge languages," in *Proc. of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL)*. Morristown, NJ, USA, 2001, pp. 1–8.

[3] C. Schafer and D. Yarowsky, "Inducing translation lexicons via diverse similarity measures and bridge languages," in *Proc. of the Conference on Natural Language Learning (CoNLL)*, 2002.

[4] L. Borin, "You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment," in *Proc. of COLING*, vol. 1, Saarbrüken, 2000, pp. 97–103.

[5] H. Wang, H. Wu, and Z. Liu, "Word alignment for languages with scarce resources using bilingual corpora of other language pairs," in *Proc. of the COLING/ACL. Poster Sessions*. Sydney, Australia, July 2006, pp. 874–881.

[6] K. Filali and J. Bilmes, "Leveraging multiple languages to improve statistical mt word alignments," in *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU)*, 2005.

[7] T. Gollins and M. Sanderson, "Improving cross language retrieval with triangulated translation," in *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA, 2001, pp. 90–95.

[8] M. Kauers, S. Vogel, C. Fuegen, , and A. Waibel, "Interlingua based statistical machine translation," in *Proc. of INTERSPEECH - ICSLP*, Denver, Colorado, USA, 2002, pp. 1909–1912.

[9] A. de Gispert and J. B. Mario, "Catalan-english statistical machine translation without parallel corpus: bridging through spanish," in *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

[10] M. Utiyama and H. Isahara, "A comparison of pivot methods for phrase-based statistical machine translation," in *Proc. of Human Language Technologies*. Rochester, New York, USA 2007, pp. 484–491.

[11] S. Kumar, F. J. Och, and W. Macherey, "Improving word alignment with bridge languages," in *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, 2007, pp. 42–50.

[12] H. Wu and H. Wang, "Pivot language approach for phrase-based statistical machine translation," in *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, 2007, pp. 856–863.

[13] T. Cohn and M. Lapata, "Machine translation by triangulation: Making effective use of multi-parallel corpora," in *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, 2007, pp. 728–735.

[14] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, 2008.

[15] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–312, 1993.

[16] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," in *Proc. of 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, 2002, pp. 147–152.

[17] P. Koehn, et al., "Moses: Open source toolkit for statistical machine translation," in *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics. Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.

[18] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003, pp. 160–167.

[19] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[20] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.