FBK @ IWSLT-2008

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

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

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

Abstract

This paper reports on the participation of FBK at the IWSLT 2008 Evaluation. Main effort has been spent on the Chinese-Spanish Pivot task. We implemented four methods to perform pivot translation. The results on the IWSLT 2008 test data show that our original method for generating training data through random sampling outperforms the best methods based on coupling translation systems. FBK also participated in the Chinese-English Challenge task and the Chinese-English and Chinese-Spanish BTEC tasks, employing the standard state-of-the-art MT system Moses Toolkit.

1. Introduction

This paper reports on the participation of FBK at the IWSLT 2008 Evaluation. This year, we mainly focused on the Pivot task as defined by the organizers. The task consists in translating from Chinese to Spanish when parallel data are not available for this language pair; for training purpose, only *independent* Chinese-English and English-Spanish corpora are provided, in the sense that they do not derive from the same set of sentences.

A statistical machine translation (SMT) system relies on the availability of parallel corpus for the estimation of its models. The translation quality is affected by the size of such corpus and its closeness to the task domain. Unfortunately, for many relevant language pairs such parallel data are available only to a small extent, or they are out-of-domain.

To circumvent the data bottleneck for this low-resourced language pairs, research on SMT has been recently investigated the use of so-called *pivot* or *bridge* languages. An overview of research on pivot translation is given in our companion paper [1]. The assumptions underlying the adoption of a pivot language are simple to state: (i) there is lack of parallel texts between F and E, while (ii) there exists a language G for which (abundant) parallel texts between F and G and between G and E are available. These assumptions are fully matched by the specifications of the Pivot task, because the English parts of the Chinese-English and English-Spanish corpora do not overlap.

We analyzed the pivot translation task from a theoretical point of view providing a mathematically sound formulation of the various approaches presented in the literature, and introduced new variations related to training of translation models through pivot language. Hence, we implemented four different approaches to the problem and experimentally compared them on the Pivot task. Two of them couple Chinese-English and English-Spanish MT systems, the third approach creates a new translation model starting from them, and the fourth approach synthesizes Chinese-Spanish training data translating the target part of the available Chinese-English corpus and creates a MT system on these data. These approaches are briefly introduced in the following Section, while a detailed description can be found in [1].

To perform a fair comparison between these approaches we relied on the well-known open source MT system Moses [2] in its standard configuration, and we did not apply any specific enhancement like lexicalized reordering models or rescoring. For each approach specific training of the models were performed on the provided BTEC data only, without using any additional training data.

We also submitted runs for the Chinese-English and Chinese-Spanish BTEC tasks and for the Chinese-English Challenge task.

This paper is organized as follows. In next Section we introduce the four approaches we have taken into account to address the pivot translation. Section 3 describes the data and the systems we employed to participate in the 2008 IWSLT evaluation. Finally, in Section 4 official results of the competition are reported and commented.

2. SMT through Pivot Languages

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

2.1. Bridging at Translation Time

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

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

The statistical decision criterion can be derived by modeling the pivot text as an hidden variable and by assuming independence between the target and the source strings, given the pivot string. By assuming standard phrase-based models, we have to extend the search criterion with other two hidden variables $\mathbf{a}$ and $\mathbf{b}$, which model phrase segmentation and reordering for each considered translation direction.

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

The max approximation (instead of summation) is applied, to reduce the complexity of the search procedure.

2.1.1. Coupling Independent Alignments

By taking inspiration from approaches proposed to cope with the very similar optimization criterion of SLT [3], we can reduce the computational burden of (1) by limiting the pivot translations $\mathbf{g}$ to a subset $\mathcal{G}(\mathbf{f})$:

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

Natural candidates to represent such subsets of pivot translations are n-best produced by the source to pivot translation engine. The use of word-graphs of translations is an alternative option we will explore in the future.

The left-hand picture in Figure 1 shows the two level alignments for a simple example involving translations from Chinese to Spanish, through English. Horizontal segments show that the English string is segmented differently when it is generated from Chinese than when it is used to generated a Spanish translation. Coupling is hence performed only at sentence level.

2.1.2. Coupling Constrained Alignments

An other interesting alternative that has been proposed in the literature, is to constrain the alignments $\mathbf{a}$ and $\mathbf{b}$ such that they share exactly the same segmentation, and such that $\mathbf{b}$ is monotonic. Thus coupling is done at phrase level. An example showing the effect of these constraints is shown in the right side of Figure 1.

The enforced one-to-one correspondence between phrases used in the two translation directions, suggest that the same space of translation options can be achieved by performing a single translation step, directly from $\mathbf{f}$ to $\mathbf{e}$, exploiting a phrase table obtained by taking the product of the two phrase tables. Phrase pairs of the new phrase table are scored as follows:

$$t(\tilde{e} \mid \tilde{f}) = \sum_{\tilde{g}} t(\tilde{f}, \tilde{g}) \times t(\tilde{g}, \tilde{e}) \quad (3)$$

where the summation is over all pivot phrases which can be translated both from $\mathbf{f}$ and to $\mathbf{e}$. $t(\cdot,\cdot)$ is the score of a phrase pair in the corresponding phrase table.

As at first sight, it seems difficult to combine the single distortion models in the way we do with the phrase tables; hence, a simple exponential distortion model is adopted.

2.2. Bridging at Training Time

A different approach to pivot translation is to directly estimate the parameters of a translation system from F to E exploiting the available corpora (F,G) and (G,E). The formal description of the method can be found in the companion paper [1]. As an efficient parameter estimation for most translation models is hard to achieve, some approximations are needed to make it manageable.

The resulting training procedure is therefore much easier and consists in three steps: i) create Ē by translating the G part of (F,G) by means of the translation engine trained on (G,E), ii) build a synthetic parallel corpus (F,Ē), and iii) train a translation system on (F,Ē).

This approach of synthesizing parallel data can be considered as an unsupervised training method.

3. Systems' Development

The first subsection reports on the available data for training and development, and the employed preprocessing. Then, the baseline system is described, which is used both for the BTEC and Challenge tasks and as building blocks for the Pivot task. Later, the systems specific for the Pivot task are presented with some details. Finally, the performance of the developed systems on a blind test are reported.

3.1. Data

Five monolingual corpora are employed for training our systems: namely two for Chinese (C1 and C2), two for English (E1 and E2) and one for Spanish (S1). All corpora are officially provided by the organizers, and are extracted from BTEC [4]; each of them contains 20K sentences.

According to the evaluation specification, the parallel corpora CE1 and CS1 are exploited for the CE- and CS-btec tasks, respectively; CE1 for the CE-challenge task; CE2 and ES1 are used to train the systems for the CES-pivot task. We stress that the parallel corpus CE1 are not considered at all for this task.

Six development sets are provided consisting of about 500 sentences each and a number of references ranging from 6 to 16 for the CE-btec task. Only one of them is available

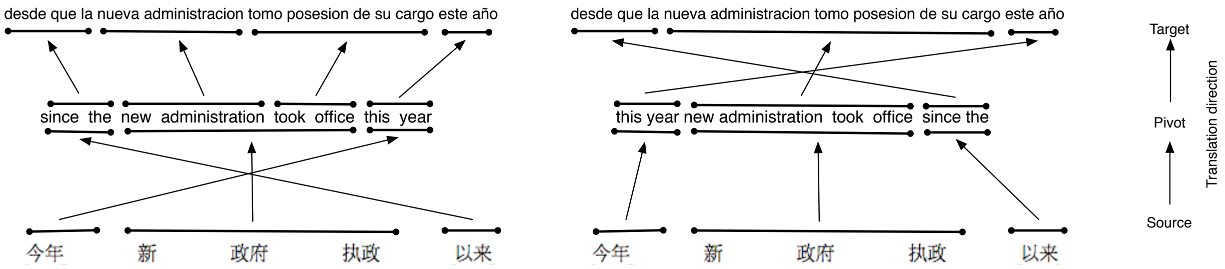
desde que la nueva administracion tomo posesion de su cargo este año
since the new administration took office this year
今年 新 政府 执政 以来
desde que la nueva administracion tomo posesion de su cargo este año
this year new administration took office since the
今年 新 政府 执政 以来
Target
Pivot
Source
Translation direction

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

for the other two tasks, namely the CS-btec and CES-pivot tasks. A further dev set of 250 sentences and one reference is provided for the CE-challenge task.

For the sake of systems' development, only one development set had been provided for several task. Hence, we randomly extracted about 1K sentences from the training data and used them as a blind test. We exploited the reduced data for training and the official development set (dev3) for tuning.

The training of the final systems had been performed exploiting the whole corpora of 20K sentences, i.e. including the blind-test, and adding the available development data. Multiple references (with their source input) are considered as distinct sentence pairs.

No additional data are employed.

Table 1 reports statistics of the parallel corpora actually exploited for training for all tasks.

Task	# sent	source		target	
		words	dict	words	dict
CE-btec	54,021	439K	8,847	499K	10,765
CS-btec	28,068	229K	8,284	250K	11,734
CE-chal	55,743	447K	8,864	507K	11,051
CE-pivot	28,095	217K	8,987	248K	8,951
ES-pivot	19,972	182K	8,385	177K	11,019

Table 1: Statistics of the parallel data used to train the final systems of different tasks.

A simple preprocessing was performed for all languages consisting in tokenizing text, and transforming numbers into digits. Chinese text is segmented into words on the basis of the word frequencies obtained from the training data. Both training, dev and test sets are actually re-segmented.

For all tasks, we were required to translate both the correct recognition result transcripts (CRR) and the ASR output; we chose to feed the systems only the 1-best transcriptions (ASR.1). Nevertheless, no particular development for the ASR condition was done, but the estimation of specific weights.

3.2. Baseline System

The baseline system *Direct* is built upon the open-source MT toolkit Moses [2]. The 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 8 weights of the log-linear combination are optimized by means of a minimum error training procedure [5].

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++ [6]. This extraction method does not apply in the case of pivoting with constrained alignments (see Section 2.1.2): phrase pairs and their scores are obtained by the product of two existing phrase tables (from source to pivot and from pivot to target). A 5-gram word-based LM is estimated on the target side of the parallel corpora using the improved Kneser-Ney smoothing [7]. The distortion model is a standard negative-exponential model.

The *Direct* systems have been used in the BTEC and Challenge tasks, and they have been exploited as constituents of the systems employed in the Pivot task.

3.3. Pivot Systems

3.3.1. Sentence-level Coupling

The first approach taken into account consists in coupling unconstrained alignments as proposed in Section 2.1.1.

Practically, we consider the CE and ES systems as black boxes, and we feed the latter the output of the former. We compare two methods for interfacing the systems. The easiest method, called *Cascade*, uses only the best English translation $\hat{\mathbf{g}}$ of the Chinese sentence $\mathbf{f}$ as an interface. In this case the subset $\mathcal{G}(\mathbf{f}) = \{\hat{\mathbf{g}}\}$ in Eq. 2.

The second way, named *Nbest*, consists in i) generating m-best Spanish translations for each of the n-best English translations $\mathbf{g}_1 \dots \mathbf{g}_n$ generated by the Chinese-English system, and ii) rescoring all nxm hypotheses using both CE and ES translation scores, 16 scores in total. In this case the subset $\mathcal{G}(\mathbf{f}) = \{\mathbf{g}_1 \dots \mathbf{g}_n\}$. Notice that *Cascade* is trivially a simplified *Nbest*.

The CE system has been trained on the CE2 parallel corpus, the ES system on ES1. In the development phase we found that $n = m = 100$ is the best configuration for *Nbest*; duplicate translation alternatives are kept.

3.3.2. Phrase-level Coupling

As remarked in Section 2.1.2 coupling constrained alignments corresponds to taking the product of the CE and ES phrase tables. We called this approach *PhraseTable*.

	CE2	ES1	product
src phr	76K	277K	21K
trg phr	82K	284K	32K
phr pairs	133K	333K	592K
avg trans	1.77	1.20	28.17
common	-	-	59K

Table 2: Statistics about the original and the product phrase tables.

Table 2 reports statistics of the original CE2 and ES1 phrase tables and 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 table 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.

Only 59K of the 133K phrase pairs (44%) in the CE2 table have a match in the ES1 table, and the common pivot phrases are mainly of length 1 (65%). These figures show that the two English corpora (E1 and E2) significantly differ, although they are in the same domain.

Henceforth, it is hard to find Spanish translations of Chinese phrases which has common correspondents into E1 and E2. In fact, less than 30% (21K over 76K) of the original Chinese phrases can be translated into Spanish through English; instead, the average number of translations hugely increases. This suggests that this approach is not recommended because the coverage on the source side is low, while the ambiguity on the target is high, at least with respect to this training data.

3.3.3. Synthesis of Training Data

The last approach we implemented, called *Synthesis*, consists in generating CS synthetic parallel data and using it as a training corpus to realize a CS translation engine (see Section 2.2). We propose to exploit the ES system trained on ES1 to translate E2 into a synthetic corpus $\bar{S}2$. The parallel corpus C$\bar{S}$2 is then used to directly train a CS system.

During the development phase we found that exploiting more translation alternatives is more beneficial than just taking the best translation provided by the ES system, and that the most effective method to select such alternatives is a random sampling according to the scores provided by the ES system. Practically, we generate n-best Spanish translations, properly normalize their scores, and sample (with replacement) m alternatives. The Chinese sentences are replicated in order to match the number of sampled translations. Experiments on the blind test show that the sampling method configured with $m = 100$ and $n = 100$ achieves the best results.

Once generated, the synthetic corpus is used to also train the target LM. Employing synthetic data $\bar{S}2$ significantly improved the scores with respect to using the supplied data S1 only; using both sets gives the best results.

More details and intermediate results can be found in the companion paper [1].

3.4. Development Results

Table 3 reports the BLEU scores of the systems we implemented for each task during the development phase. These results are given on the blind test we introduced before, which has only one reference per input. Notice that these results were obtained using the reduced training corpora without any development set. Systems in **bold** are chosen as primary submissions for the official evaluation. The Table also reports the performance of the two CE and ES systems trained on Pivot data and used as building blocks for the systems developed in the Pivot task. No performance are reported for the system developed for the CE-challenge task: actually for this condition we used exactly the same system developed for the CE-btec task, but for the feature weights which were optimized on the provided development set of spontaneous speech utterances.

Task	Data	System	BLEU
CE-btec	CE1	*Direct*	**26.91**
CS-btec	CS1	*Direct*	**23.67**
CS-pivot	CE2+ES1	*Cascade*	16.44
		Nbest	17.64
		PhraseTable	16.65
		Synthesis	**17.68**
CE-pivot	CE2	*Direct*	19.09
ES-pivot	ES1	*Direct*	49.13

Table 3: Results (BLEU) on a blind test set achieved by different systems implemented during the development.

Note that from a computational point of view *Nbest* is expensive at run-time; it actually translates $n + 1$ times (1 for CE and n for ES) and rescores and reranks nxm alternatives per input sentence. Instead, *Synthesis* requires much more time for training because of the translation of the whole English corpus, but it is fast at run time, because it translates each input sentence only once. Furthermore, we found that *Synthesis* significantly outperforms *Nbest* in preliminary experiments carried out on Chinese-Spanish-English pivot translation task we created using the available BTEC data. A reasonable explanation for this behavior is that *Synthesis* completely skips one of the translation steps and fully exploits the other one. Skipping completely the most difficult step – i.e. translating from Chinese into English or Spanish – is a rewarding strategy.

For these reasons, we preferred consider *Synthesis* rather than *Nbest* as primary system.

4. Evaluation Results

We submitted two runs for each of the CE- and CS-btec tasks and CE-challenge task. The contrastive runs for the ASR.1 input conditions were obtained using the optimal weights tuned on the CRR development input; notice that this condition would not be allowed by the evaluation specification.

For the CES-pivot task we submitted several contrastive runs to compare different approaches. For the contrastive run 1 (and the corresponding 3), the *Synthesis* system was trained with the CS development set as supplied. Although we supposed in advance that such data would have improved the performance, we decided not to use this system as primary because such data in our opinion violate the pivot assumption – that is, unavailability of parallel CS data. In the primary *Synthesis* system, only the Chinese and English component of the development data were employed, while the Spanish was synthesized by translating and sampling as previously described.

Table 4 reports the official BLEU% scores of our submitted runs provided by the organizers.

Task	System	Run	BLEU	
			ASR.1	CRR
CE-btec	*Direct*	**prim**	**36.91**	**40.18**
		contr	36.45	”
CS-btec	*Direct*	**prim**	**26.67**	**30.29**
		contr	27.05	”
CE-chal	*Direct*	**prim**	**23.84**	**27.00**
		contr	23.88	”
CES-pivot	*Cascade*	contr6	29.20	33.52
	Nbest	contr7	32.69	37.41
	PhraseTable	contr4	28.52	33.13
		contr5	30.09	”
	Synthesis	**prim**	**33.11**	**39.69**
		contr2	35.94	’
		contr1	34.14	39.93
		contr3	35.98	”

Table 4: Results (BLEU) on the official IWSLT08 test set.

Figures about pivot systems confirm what we found in the development phase: *Synthesis* outperforms *Direct* and *PhraseTable*, which achieve very close performance. In the CRR input condition *Synthesis* is significantly better than *Nbest*. Interestingly, the CRR-based optimal weights give better results than the ASR-based, at least in the Pivot task.

The comparison against the IWSLT08 top performing system shows a large gap (40.18 vs. 50.85) in the CE-btec task, which halves (30.29 vs 35.82) in the CS-btec task, where we rank second.

In the CES-pivot task, where we mostly focused our efforts, the gap further reduces to less than 2 BLEU% points (39.69 vs. 41.57), ranking again second. Instead the gap from our *Cascade* system is still large (33.52 vs. 41.57). This confirms our assumption that avoiding the CE translation, which poorly performs, is a winning strategy.

Furthermore, comparison between primary and contr1 runs of the *Synthesis* corroborate the straightforward intuition that using correct Spanish translations is better than using synthesized ones.

Results achieved with the ASR input essentially confirm rankings and gaps of the CRR condition. Instead, the poor performance achieved in the CE-challenge task are explained by the lack of effort on the specific domain and genre condition.

Finally, we want to stress again that we exploited only the allowed BTEC data: neither bilingual nor monolingual training corpora are added.

5. References

[1] N. Bertoldi, et al., “Phrase-based statistical machine translation with pivot languages,” in *Proc. of the International Workshop on Spoken Language Translation - IWSLT*, Honolulu, Hawaii, USA, 2008.

[2] 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.

[3] F. Casacuberta, et al., “Recent efforts in spoken language processing,” *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, 2008.

[4] T. Takezawa, et al., “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.

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

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

[7] 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.