Persian–Spanish Low-Resource Statistical Machine Translation
Through English as Pivot Language

Benyamin Ahmadnia\textsuperscript{1}, Javier Serrano\textsuperscript{1}, Gholamreza Haffari\textsuperscript{2}
\textsuperscript{1}Autonomous University of Barcelona, Cerdanyola del Valles, Spain
\textsuperscript{2}Monash University, Clayton, VIC, Australia
\{benyamin.ahmadnia,javier.serrano\}@uab.cat
gholamreza.haffari@monash.edu

Abstract
This paper is an attempt to exclusively focus on investigating the pivot language technique in which a bridging language is utilized to increase the quality of the Persian–Spanish low-resource Statistical Machine Translation (SMT). In this case, English is used as the bridging language, and the Persian–English SMT is combined with the English–Spanish one, where the relatively large corpora of each may be used in support of the Persian–Spanish pairing. Our results indicate that the pivot language technique outperforms the direct SMT processes currently in use between Persian and Spanish. Furthermore, we investigate the sentence translation pivot strategy and the phrase translation in turn, and demonstrate that, in the context of the Persian–Spanish SMT system, the phrase-level pivoting outperforms the sentence-level pivoting. Finally we suggest a method called combination model in which the standard direct model and the best triangulation pivoting model are blended in order to reach a high-quality translation.

1 Introduction
The goal of Statistical Machine Translation (SMT) is to translate a source language sequences into a target language by assessing the plausibility of the source and the target sequences in relation to existing bodies of translation between the two languages. The presence of sizable bodies of aligned parallel corpora affects the SMT systems function and performance. On the other hand, gathering parallel data in practice becomes an issue due to the high costs and the limitation in scope which as a result may constrain the related research and its applications. Therefore, the scarcity of parallel data for many language pairs is amongst the main issues in SMT. (Babych et al., 2007).

Corpora of this type are usually rare, especially for under-resource pairs such as Persian and Spanish. Even for well-resource languages, such as those included in Europarl (Koehn, 2005), which covers the language of debates in the European Parliament, SMT performance degrades significantly while being applied to a slightly different domain. Therefore, with a change in the domain, the performance loses its efficiency. A common solution to the lack of parallel data is using pivot language technique (El-Kholy and Habash, 2013). This technique is used to generate a systematic SMT when a proper bilingual corpus is lacking or the existing ones are weak. This issue becomes significant when there are languages with inefficient NLP (Natural Language Processing) resources to be able to provide an SMT system. However, there are sufficient resources between them and some other languages. Though it is claimed that, the intermediary languages do not lead to an improvement in general case, this idea can be employed as a simple method to enrich the translation performance even for existing systems (Matusov et al., 2008).

In this paper we display how this idea acts effectively concerning the low-resource Persian-Spanish language pair. Besides, we examine a selective combination approach to efficiently blend a pivot and a direct model developed by a given parallel corpora to achieve better coverage and overall translation quality. We increase the obtained i-
formation through picking the relevant portions of
the pivot model that do not interfere with the more
trusted direct model.

2 Language Issues

SMT has proven to be successful for a number of
language pairs. However, as soon as the Persian language is involved with any sort of machine translation, a number of difficulties are encountered. Of other common languages, English seems to be the best language to pair with Persian, since it is best supported by resources such as large corpora, language processing tools, and syntactic tree banks. Persian is the complete opposite, with a significant shortage of digitally available text, both parallel and monolingual. Other language pairs make use of parallel corpora of many millions of sentences, giving any applied system a huge database to work from, and thus output much more accurate results. Persian is morphologically rich, with many characteristics not shared by other languages. It makes no use of articles (“a”, “an”, “the”), there is no distinction between capital and lower-case letters, and symbols and abbreviations are rarely used. Sentence structure is also different, Persian placing parts of speech such as nouns, subjects, adverbs and verbs in different locations in the sentence, and sometime even omitting them altogether. Some Persian words have many different versions of spelling, and it is not uncommon for translators to invent new words. This can result in an Out-Of-Vocabulary (OOV) output.

The Spanish language utilizes the Latin alphabet, with a few special letters; Vowels with an acute accent (á, ú, é, í, ó), (u) with an umlaut (ü), and an (n) with a tilde (ñ). The Spanish language spelling system, due to a substantial number of reforms, is almost perfectly phonemic and, therefore, easier to learn than the majority of languages. The Spanish language is pronounced phonetically. However, beware of the trilled (r) which is somewhat complex to reproduce. The letters (b) and (v) are almost indistinguishable. The letter (h) is silent. Spanish language punctuation is very close to, but not the same as English. There are a few significant differences. For example, in Spanish, exclamation and interrogative sentences are preceded by inverted question and exclamation marks. Also, in a Spanish conversation, a change in speakers is indicated by a dash, while in English, each speaker’s remark is placed in separate paragraphs. Formal and informal translations address several different characteristics. Inflection, declination and grammatical gender are important features of the Spanish language.

In both Persian and Spanish languages, the word-order is different from English in two ways; First, the modifier comes before the word it modifies. Second, the sentences follow a "Subject", "Object", "Verb" (SOV) order.

3 Baseline Translation System

The SMT paradigm has, as its most important elements, the idea; That probabilities of source and target sentences can find the best translations. Frequently used paradigms of SMT on the log-linear model are the phrase-based SMT, the hierarchical phrase-based SMT, and the ngram-based SMT. In our experiments we use the phrase-based SMT system with the maximum entropy framework (Berger et al., 1996).

\[
\hat{y}_1^t = \arg \max_{y_1^t} P(x|y) \tag{1}
\]

The phrase-based SMT model is an example of the noisy-channel approach, where we can present the translation hypothesis \( t \) as the target sentence (given \( s \) as a source sentence), maximizing a log-linear combination of feature functions:

\[
\hat{y}_1^t = \arg \max_{y_1^t} \left\{ \sum_{m=1}^{M} \lambda_m h_m(x_1^t, y_1^t) \right\} \tag{2}
\]

This equation called the log-linear model, where \( \lambda_m \) corresponds to the weighting coefficients of the log-linear combination, and the feature functions \( h_m(x,y) \) to a logarithmic scaling of the probabilities of each model. The translation process involves segmenting the source sentence into source phrases \( s \), translating each source phrase into a target phrase \( t \), and reordering these target phrases to yield the target sentence \( t \).

4 Pivoting Strategy for SMT

High-quality data set is not always available for training the SMT systems. One of the possible ways to solve this impasse is to using a third language as a bridge one for which there exist high-quality source–pivot and pivot–target bilingual resources. Pivot-based strategies which are employed for SMT systems can be classified into three categories (Wu and Wang, 2007);
**Transfer method:** This method, which is also recognized as cascade or sentence translation pivot strategy, translates the text in the source language to the pivot through employing a source–pivot translation model, and subsequently translate it to a target language utilizing a pivot–target translation model.

**Triangulation method:** This method is known as phrase-table multiplication or phrase translation pivot strategy, which combines the corresponding translation probabilities of the translation models for the source–pivot and the pivot–target languages, thus generating a novel model for the source–target translation.

**Synthetic corpus method:** This method attempts to develop a synthetic source–target corpus by translating the pivot part in the source–pivot corpus, into the target language by means of a pivot–target model, and translating the pivot part in the target–pivot corpus into the source language with a pivot–source model. Eventually, it combines the source sentences with the translated target sentences or combines the target sentences with the translated source sentences. However, it is complicated to create a high-quality translation system with a corpus compiled merely by an MT system.

In the present paper, we will rely on the first and the second methods for doing our SMT pivoting experiments.

### 4.1 Transfer Pivoting Method

In the sentence translation pivot strategy, first the Persian sentences are translated into the English ones, followed by translation of these English sentences into the Spanish ones separately. We choose the highest scoring sentence amongst the Spanish sentences.

In this methodology for assigning the best Spanish candidate sentence \( s \) to the input Persian sentence \( p \), we maximize the probability \( P(s|p) \) by defining hidden variable \( e \), which stands for the pivot language sentences, we gain:

\[
\arg \max_p P(s|p) = \arg \max_p \sum_e P(s,e|p)
= \arg \max_p \sum_e P(s|e,p) P(e|p)
\]

Assuming that, \( p \) and \( s \) are independent given \( e \):

\[
\approx \arg \max_p \sum_e P(s|e) P(e|p)
\]

In Equation (4) summation on all \( e \) sentences is difficult, so we replace it by maximization, and Equation (5) is an estimate of Equation (4):

\[
\approx \arg \max_p \max_e P(s|e) P(e|p)
\]

Instead of searching all the space of \( e \) sentences, we can just search a subspace of it. For simplicity we limit the search space in Equation (6). A good choice is \( e \) subspace produced by the \( k\)-best list output of the first SMT system (source-pivot):

\[
\approx \arg \max_p \max_{e \in k\text{-best}(s)} P(s|e) P(e|p)
\]

In fact each sentence \( p \) of the Persian test set is mapped to a subspace of total \( e \) space and search is done in this subspace for the best candidate sentence \( s \) of the second SMT system (pivot–target).

### 4.2 Triangulation Pivoting Method

Concerning the phrase translation pivot strategy, we directly create a Persian–Spanish phrase translation table from a Persian–English and an English–Spanish phrase-table.

In this technique, phrase \( p \) in the source–pivot phrase-table is connected to \( e \), and this phrase \( e \) is associated with phrase \( s \) in the pivot–target phrase-table. We link the phrases \( p \) and \( s \) in the new phrase-table for the source–target. For scoring the pair phrases of the new phrase-table, assuming \( P(e|p) \) as the score of the Persian–English phrases and \( P(s|e) \) as the score of the English–Spanish phrases, then the score of the new pair phrases \( p \) and \( s \), \( P(s|p) \), in Persian–Spanish phrase-table is counted:

\[
P(s|p) = \sum_e P(s,e|p)
\]

\( e \) is a hidden variable and actually stands for the phrases of pivot language:

\[
P(s|p) = \sum_e P(s|e,p) P(e|p)
\]

Assume that, \( p \) and \( s \) are independent, given \( e \):

\[
P(s|p) \approx \sum_e P(s|e) P(e|p)
\]
For simplicity the summation on all the $e$ phrases is replaced by maximization, then Equation (9) is approximated by:

$$P(s|p) \approx \max_e P(s|e) P(e|p) \quad (10)$$

Applying a translation model on a small bilingual corpus alone will result in a poor translation system performance. Therefore, the cause of such a poor performance is sparse data. Aiming at improving this performance, we can utilize additional source–pivot and pivot–target parallel corpora. Furthermore, more than one pivot languages can be utilized in order to enrich the quality of the translation performance. Different pivot language may catch different language phenomenon and can improve translation quality by adding quality source–target phrase pairs.

If we include $k$ pivot languages, $k$ pivot models can be estimated. Linear interpolation is employed for combining all these generated models with the standard model trained with the source-target corpus. Equations (11) and (12) demonstrate the estimation of the phrase translation probability and the lexical weight respectively.

$$P(s|t) = \sum_{i=1}^{k} \alpha_i P_i(s|t) \quad (11)$$

$$P(s|t, \alpha) = \sum_{i=1}^{k} \beta_i P_i(s|t, \alpha) \quad (12)$$

Where $P(s|t)$ and $P(s|t, \alpha)$ denote the phrase translation probability and the lexical weight trained with the source–target corpus estimated by using pivot languages. Both $\alpha_i$ and $\beta_i$ are interpolation coefficients. Meanwhile $\sum_{i=1}^{k} \alpha_i = 1$, and $\sum_{i=1}^{k} \beta_i = 1$.

5 Experimental Framework

The data is gathered from in-domain **Tanzil** parallel corpus\(^1\) (Tiedemann, 2012). In this corpus, the Persian–Spanish part encompasses more than (68K) parallel sentences, nearly (2.06M) words in the Persian side, and more than (1.45M) words in the Spanish side. Besides, the Persian–English part includes more than (1M) parallel sentences, around (30.88M) Persian words, and more than (26.14M) English words. The English–Spanish part contains more than (138K) parallel sentences, approximately (2.37M) words in the English side, and over (1.94M) words in the Spanish side. Table below presents the corpus statistics, which have been used in our experiments, including the source and the target languages information in each direction.

<table>
<thead>
<tr>
<th>Direction</th>
<th>Pe-En</th>
<th>En-Es</th>
<th>Pe-Es</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>1,028,996</td>
<td>138,822</td>
<td>68,601</td>
</tr>
<tr>
<td>Src. Words</td>
<td>30,872,937</td>
<td>376,933</td>
<td>2,058,231</td>
</tr>
<tr>
<td>Trg. Words</td>
<td>26,143,026</td>
<td>1,932,696</td>
<td>1,454,778</td>
</tr>
</tbody>
</table>

Table 1: Corpus statistics

To examine the size factor on our SMT systems, data were compiled into two sets including System(1) and System(2). The `tokenize.perl` script has been employed for tokenizing all datasets. System(1) training part consists of (10K) sentences and it spreads almost (50K) sentences to System(2) with nearly (60K) sentences. In order to conduct the tuning and the testing steps, we gathered parallel texts from Tanzil corpus. (3K) sentences for the tuning, and (5K) sentences for the testing step were extracted.

**MOSES** package\(^2\) (Koehn et al., 2007), is employed for training our SMT systems. Through employing **MOSES** decoder, *fast-align* approach (Dyer et al., 2013), is applied for word alignment. We employ 3-grams language model for all SMT systems and they are developed by means of the **KenLM** toolkit (Heafield et al., 2013). In addition, for evaluating the systems performance, we use the **BLEU** metric. We set the beam-size to (100), and the distortion limit to (6). We restrain the maximum target phrases to (6) that are loaded for each source phrase, and we draw on the same other default features of **MOSES** toolkit.

For the translation systems we conduct two sets of experiments with different data sizes. The training data is collected from the beginning of the same parallel corpus, so the larger training set includes the smaller one. For instance, in the experiments with small dataset sizes, the two phrase-tables employed to shape a new table in the phrase pivoting method are extracted in turn from the Persian–English and the English–Spanish translation systems.

For conducting the first phase of our experiments, **English** was utilised by the transfer pivoting system as an interface between two separate phrase-based SMT systems, specifically a

\(^1\)http://opus.lingfil.uu.se/Tanzil.php

\(^2\)http://www.statmt.org/moses
**Persian–English** direct system and an **English–Spanish** direct system. Besides, while translating Persian to Spanish, the English top-1 output of the Persian–English system was forwarded as input to the English–Spanish system. The English language model which was used to train the Persian–English system is developed from the counterpart of the Spanish data used to build the Spanish language model in our considered parallel corpus.

For applying the triangulation method during the second portion of our experiments, we required to create a phrase-table to train the phrase-based SMT system. Therefore, a Persian–English phrase-table and an English–Spanish phrase-table were needed. Based on these tables, we formed a Persian–Spanish phrase-table. Furthermore, a matching algorithm that identifies parallel sentences pairs among the tables were utilized. After identifying candidate sentence pairs, we finally used a classifier to determine if the sentences in each pair are a good translation for each other and update our Persian–Spanish phrase-table with the selected pairs. Table below illustrates the results of both Persian–Spanish standard direct, and pivot-based translation systems through English as the intermediary language.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>System(1)</th>
<th>System(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>19.07</td>
<td>19.39</td>
</tr>
<tr>
<td>Transfer</td>
<td>20.33</td>
<td>20.78</td>
</tr>
<tr>
<td>Triangulation</td>
<td>21.02</td>
<td>21.55</td>
</tr>
</tbody>
</table>

Table 2: The BLEU scores comparing the performance of direct with pivoting Persian–Spanish SMT systems, through two training data sizes

As expected, by an increase in the dataset size, the **BLEU** score rises. For the large size of training data set, the best result of the direct translation system is (19.39) point in term of the BLEU. Also the system achieved a **BLEU** score of (21.55) for the phrase-level pivoting, while for the sentence-level pivoting the system achieved (20.78) **BLEU** point.

The results indicate that, the pivot-based translation method is suitable for the scenario that there exist large amounts of source–pivot and pivot–target bilingual corpora and only a little source–target bilingual data. Thus we selected (10K), and (60K) sentence pairs from the source–target bilingual corpora to simulate the lack of source–target bilingual data. As seen in Table 2, in phrase pivoting portion, the **Persian–English–Spanish** relative increase from System(1) to System(2) is approximately (10.25%), and in sentence pivoting portion, the **Persian–English–Spanish** relative increase from System(1) to System(2) is (10.22%). This suggests that, we are making better use of the available resources. The differences between pivot language method and direct translation approach are statistically significant confidence level.

### 6 Direct and Pivot Combination

We examine a combination approach so as to achieve a higher coverage and a better translation quality, aiming at efficiently merging both a phrase-based pivot and a direct translation models developed from a given parallel corpora.

In particular, this approach is an attempt to combine the direct and triangulation models in order to rise the amount of the gained information. We use **MOSES** toolkit as it lets employing the multiple translation tables for doing the combination experiments. In order to achieve this aim, several combination models are approachable and practical. In the current paper, we employ a combination model where the translation options are gathered from one table, and additional options are collected from other tables. Reaching similar translation options in multiple tables, we form separate translation options for each occurrence with different scores. Table 3 reveals the comparison between the findings of the basic combination technique with those of the best multiplication pivot translation and the direct translation models.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Direct + Triangulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>System(1)</td>
<td>21.88</td>
</tr>
<tr>
<td>System(2)</td>
<td>22.02</td>
</tr>
</tbody>
</table>

Table 3: The BLEU scores of the combination experiments between the best triangulated and the direct SMT models for Persian–Spanish languages

The findings indicate that, merging and combining these two models, results in an improvement in the performance.

### 7 Previous Work

The pivot language approach has been previously applied for diverse purposes. For instance developing a technique for mining the web to collect parallel corpora for low-density language pairs (Resnik and Smith, 2003), and running new SMT system for languages Catalan–English with no parallel corpus (Gispert and Mariño, 2006).
In a research conducted by Utiyama and Isahara (2007), the use of pivot language through phrase translation and sentence translation are investigated. Moreover, Wu and Wang (2007) discuss three methods for pivot strategies in their findings including phrase translation method (i.e. triangulation), transfer method, and synthetic method.

Some researchers investigated the SMT system with pivot language technique. For example Babych et al. (2008) used Russian language as a pivot for translating from Ukrainian to English. Their comparison revealed that it is possible to achieve better translation quality with pivot language approach.

Habash and Hu (2009) compared two approaches for Arabic–Chinese MT system with direct MT system through English as a pivot language. The findings of their study indicated that using English as a pivot language in either approach outperforms direct translation from Arabic to Chinese.

In another study, Bakhshaei et al. (2010) used English as a bridging language while translating from Persian to German and concluded that using the pivot technique in phrase-level combination outperforms direct translation system. Furthermore, Al-Hunaity et al. (2010) presented a comparison between two common pivot strategies: phrase translation and sentence translation in order to enhance Danish–Arabic SMT system. According to their findings, it is illustrated that sentence pivoting overtake phrase pivoting when common parallel corpora are not available.

Nakov and Ng (2012) try to exploit the similarity between resource-poor languages and resource-rich languages for the translation task.

Paul et al. (2013) debates over criteria to be considered for selection of good pivot language. Use of source-side segmentation as pre-processing technique is demonstrated by Kunchukuttan et al. (2014).

Dabre et al. (2015) used multiple decoding paths (MDP) to overcome the limitation of small sized corpora.

8 Conclusion

In this paper, we compared two common pivot language translation methods comprising phrase-level pivoting and sentence-level pivoting for low-resource Persian–Spanish SMT by employing English as an intermediary language.

Through conducting controlled experiments using the Tanzil corpus, we assessed the performances of these two methods against the performance of directly trained SMT system. The findings of our experiments revealed that utilising English as a bridging language in either approaches outperforms direct translation method from Persian to Spanish. Our best result is the phrase translation pivoting system scores higher than the best result of the sentence translation pivoting system by (0.77) BLEU points, and also higher than the best result of the Persian–Spanish direct translation system by (2.16) BLEU points.

Furthermore, the performance of a combination model between two different translation approaches on the translation quality is investigated in this paper. In order to apply this combination model, we employed the best pivoting translation model (phrase-level) along with the best standard direct translation model for attaining a high-quality translation. The results reveal that combining these two models cause an improvement in the performance quality. The BLEU score for this new combined translation system enhanced by (+0.33) point in comparison with the best triangulated, and (+2.49) point in comparison with the best standard direct translation system for Persian–Spanish language pair.

Acknowledgments

The authors would like to express their sincere gratitude to Vahe de Nosrati (PhD candidate at Monash University, Australia), for all her support. The authors have benefited from her erudition and thoughtful comments which truly enriched the present work.

References


