Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages

Raj Dabre
Graduate School of Informatics
Kyoto University
Kyoto 606-8501
prajdabre@gmail.com

Fabien Cromieres
Japan Science and Technology Agency
Kawaguchi-shi
Saitama 332-0012
fabien@pa.jst.jp

Sadao Kurohashi
Graduate School of Informatics
Kyoto University
Kyoto 606-8501
kuro@i.kyoto-u.ac.jp

Pushpak Bhattacharyya
CFILT
IIT Bombay, Powai
India 400076
pushpakbh@gmail.com

Abstract

We present our work on leveraging multilingual parallel corpora of small sizes for Statistical Machine Translation between Japanese and Hindi using multiple pivot languages. In our setting, the source and target part of the corpus remains the same, but we show that using several different pivot to extract phrase pairs from these source and target parts lead to large BLEU improvements. We focus on a variety of ways to exploit phrase tables generated using multiple pivots to support a direct source-target phrase table. Our main method uses the Multiple Decoding Paths (MDP) feature of Moses, which we empirically verify as the best compared to the other methods we used. We compare and contrast our various results to show that one can overcome the limitations of small corpora by using as many pivot languages as possible in a multilingual setting. Most importantly, we show that such pivoting aids in learning of additional phrase pairs which are not learned when the direct source-target corpus is small. We obtained improvements of up to 3 BLEU points using multiple pivots for Japanese to Hindi translation compared to when only one pivot is used. To the best of our knowledge, this work is also the first of its kind to attempt the simultaneous utilization of 7 pivot languages at decoding time.

1 Introduction

With the increasing size of parallel corpora it has become possible to achieve very high quality translation. However, not all language pairs are blessed with the availability of large parallel corpora in the sizes of millions of lines. With the exception of the major European languages and a few Asian languages like Chinese and Japanese, other languages have parallel corpora in the sizes of a few thousands of lines. Since translation quality is related to the size of the parallel corpus, it is impossible to achieve the same level of translation quality as that in the case of resource rich languages. To remedy this scenario, an intermediate resource rich language can be exploited. Although, finding a direct parallel corpus between source and target languages might be difficult, there are higher odds of finding a pair of parallel corpora: one between the source language and an intermediate resource rich language (henceforth called pivot[1]) and one between that pivot and the target language.

Using the methods developed for Pivot Based SMT (Wu and Wang, 2007) (Utiyama and Isahara, 2007) one can use the source-pivot and pivot-target parallel corpora to develop a source-target translation system (henceforth called as pivot based system [2]) . Moreover, if there exists a small source-target parallel corpus then the resulting system (henceforth called as direct system[3]) can be supported by the pivot based source-target system to significantly improve the translation quality. Note that in this paper we use the terms "translation system" and "phrase table" interchangeably since the phrase table is the

[1] In most cases this is English.

[2] The phrase table will be known as the pivot phrase table.

[3] The phrase table will be called as direct phrase table and the corpus will be the direct parallel corpus.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1192–1202,
Denver, Colorado, May 31 – June 5, 2015.

main component of the translation system. Reordering tables are supplementary and can usually be replaced by a simple distortion model.

Major problems arise when source-pivot and pivot-target corpora belong to different domains leading to rather poor quality translations. Even if the individual corpora are large, one will run into domain adaptation problems. In such a scenario the availability of a small size multilingual corpus of a few thousand lines belonging to a single domain can be beneficial. The setting of this paper is:

1. We suppose the existence of a multilingual corpus with sentences aligned across N[4] different languages.

2. We show using the other languages as additional pivots leads to the construction of better phrase tables and better translation results.

Note that this setting is realistic and differs from the majority of existing work on pivot languages, in which the source-pivot and pivot-target corpora are unrelated (or at least do not have equivalent sentences). In addition to the well-known Europarl corpus, many other similar multilingual corpora exist. For example, a multilingual parallel corpus for 9 major Indian Languages belonging to the Health and Tourism domain of approximately 50000 lines was used to develop basic SMT systems (Kunchukuttan et al., 2014). For our experiments we will use a recently released Bible domain multilingual parallel corpus (Resnik et al., 1999) for a large number (over 25) of languages (other than Indian) including Japanese and Hindi (Japanese to Hindi translation being our focus) of approximately 30000 lines. We chose this setting because we feel that this multilingual approach is especially important for low-resource language pairs.

Typically system combination methods like linear interpolation are used to combine the direct and pivot phrase tables by modifying the probabilities of phrase pairs leading to the modification of the underlying distribution which affects the resultant translation quality. The Multiple Decoding Paths (Birch and Osborne, 2007) (MDP) feature has been used to combine two source-target phrase tables of different domains for domain adaptation (Koehn and Schroeder, 2007) but not so extensively in a pivot language scenario, especially when multiple pivots are involved (7 in our case). Our work is different from other previous works in the following ways:

- We work on a realistic low resource setting for translation between Japanese and Hindi in which we use small sized multilingual corpora containing translations of a sentence in multiple languages.

- We focus on the impact of using a relatively large number of pivot languages (7 to be precise) to improve the translation quality and compare this to when only one pivot language is used.

- Most works focus on obtaining pivot based phrase tables on relatively larger corpora than the ones used for the direct phrase table. We use the same corpora sizes for the pivot as well as direct tables.

- We verify that Multiple Decoding Paths (MDP) feature of Moses is much more effective than plain linear interpolation, especially when more pivot languages are used together.

- We show that simply varying the pivot language leads to additional phrase pairs being acquired that impact translation quality.

Section 2 contains the related work. Section 3 begins with a basic description about the languages involved, followed by the corpora details and the experimental methodology. Section 4 consists of results, observations and discussions. The paper ends with conclusions and future work.

2 Related Work

Utiyama and Isahara (2007) developed a method (sentence translation strategy) for cascading a source-pivot and a pivot-target system to translate from source to target using a pivot language. Since this results in multiplicative error propagation Wu and Wang (2009) developed a method (triangulation) in which they combined the source-pivot and pivot-target phrase tables to get a source-target

[4]The construction of a multilingual corpus has already the benefit that each new language added to it will allow direct translation with a SMT system for N new language pairs.

phrase table. They then combine the pivoted and direct tables by linear interpolation whose weights were manually specified. There is a method to automatically learn the weights (Sennrich, 2012) but it requires reference phrase pairs not easily available in resource constrained scenarios like ours. Work on translation from Indonesian to English using Malay and Spanish to English using Portuguese (Nakov and Ng, 2009) as pivot languages worked well since the pivots had substantial similarity to the source languages. This is one of the first works to use MDP in the pivot based SMT scenario.

(Paul et al., 2013) and (Paul et al., 2009) showed that English is not the best pivot language for many language pairs, including Japanese and Hindi. This was reason enough for us to not consider English as a pivot in our experiments. None of the above works focus on the utilization and impact of more than 2 pivots in their experiments which was one of our main objectives. Related to multilingual translation are works by Habash and Hu (2009), El Kholy et al. (2013), Salloum et al. (2014) and Koehn et al. (2009). Work on multi source translation (Och and Ney, 2001) which is complementary to our work must also be noted.

In the related field of information retrieval, pivot languages were employed to translate queries in cross-language information retrieval (CLIR) (Gollins and Sanderson, 2001) (Kishida and Kando, 2003). Chinnakotla et al. (2010) retrieved feedback terms from documents written in the pivot languages (after translating back from the pivot), and augmented source queries leading to improvements in information retrieval. We now talk about the languages, corpora and experiments conducted.

3 Description of Languages, Corpora and Experiments

We first describe the pivot languages and the corpora we use. We follow this with a description of the triangulation method which we use to construct phrase tables using the pivot languages, the methods used to combine the constructed tables and then the experiments that use them.

3.1 Languages involved

We performed experiments on translation between Japanese and Hindi which do not belong to the same language group but exhibit many similarities: Japanese (J) and Hindi (H) both have SOV order and are morphologically rich. For pivots we considered languages like Chinese, Korean (East-Asian languages of which Korean is closer to source), Marathi, Kannada, Telugu (Indian languages closer to target), Paite (Sino-Tibetian) and Esperanto (relatively distant from both source and target). Increasing the number of languages reduced the size of multilingual parallel translations available[5]. Our choice of languages was initially random but led to interesting observations as will be seen later.

3.2 Corpora Details

The corpora used comes from the freely available multilingual Bible corpus[6] stored in XML files. After sentence aligning all 9 languages we got 29780 sentence tuples. A tuple contains 9 sentences: 1 for each language. This we divided into 29000 training tuples, 280 tuning tuples and 500 testing tuples. The Japanese sentences were segmented using JUMAN (Kurohashi et al., 1994). The Chinese and Korean (Hangul blocks were space separated) sentences were directly available in their character segmented form. The corpora of the other languages were left morphologically and syntactically unprocessed.

3.3 Phrase Table Triangulation

We implemented the phrase table triangulation method (Wu and Wang, 2007) using JAVA as the programming language. The phrase table has 4 main scores: forward and inverse phrase translation probabilities (equations 1 and 2) accompanied by forward and inverse lexical translation probabilities (equations 3 and 4). The formulae for generating them using pivots are:

$$\Theta(f|e) = \sum_{p_i} \Theta(f|p_i) * \Theta(p_i|e) \qquad (1)$$

[5]It must be noted that Hebrew and Greek are most likely the languages from which the Bible sentences were translated into the other languages.

[6]http://homepages.inf.ed.ac.uk/s0787820/bible/

$$\Theta(e|f) = \sum_{p_i} \Theta(e|p_i) * \Theta(p_i|f) \quad (2)$$

$$P_w(f|e,a) = \sum_{p_i} P_w(f|p_i,a_1) * P_w(p_i|e,a_2) \quad (3)$$

$$P_w(e|f,a) = \sum_{p_i} P_w(e|p_i,a_2) * P_w(p_i|f,a_1) \quad (4)$$

Here a_1 is the alignment between phrases f (source) and p_i (pivot), a_2, the alignment between p_i and *e* (target) and *a* the alignment between *e* and *f*. Note that the lexical translation probabilities are calculated in the same way as the phrase probabilities. Our results might improve even more if we used more sophisticated approaches like crosslanguage similarity method or the method which uses pivot induced alignments (Wu and Wang, 2007).

3.4 Phrase Table Combination

There are 3 ways to combine phrase tables: linear interpolation, fillup interpolation and multiple decoding paths. Linear interpolation is performed by merging the tables and computing a weighted sum of phrase pair probabilities from each phrase table giving a final single table. Typically, the direct phrase table is given a significantly higher weight than the pivot based table.

$$\Theta(f|e) = \alpha_0 * \Theta_{direct}(f|e) + \sum_{l_i} \alpha_{l_i} * \Theta_{l_i}(f|e)$$
$$subject\ to\ \alpha_0 + \sum_{l_i} \alpha_{l_i} = 1 \quad (5)$$

Typically α_0 is 0.9 (Wu and Wang, 2009) and the pivot languages are collectively given a weight of 0.1. $\Theta_{l_i}(f|e)$ is the inverse translation probability for language l_i. In our experiments we set the α's according to the ratio of the BLEU scores, on the test set, of the translations using the individual phrase tables. It is possible to learn optimal weights but this requires a collection of reference phrase pairs which would not be readily available in a resource constrained scenario.

Fillup interpolation does not modify phrase probabilities but selects phrase pair entries from the next table if they are not present in the current table. The priority of the phrase tables should be specified which we do by ranking them according to the BLEU scores on a test set.

Multiple Decoding Paths (MDP) method of Moses which uses all the tables simultaneously while decoding ensures that each pivot table is kept separate and translation options are collected from all the tables. Increasing the number of pivot languages slows the decoding process drastically but the existence of powerful machines negates this limitation. For the sake of completeness we also experimented with a combination of both, linear and MDP, methods by: Firstly, combining the pivot based phrase tables into a single table using equation 5 (using the ratio of BLEU scores as interpolation weights) followed by using this table to support the direct phrase table by MDP. Note that the right way would be to use the BLEU scores on the tuning set but our objective was to show that even in the best case scenario (also called Oracle[7] scenario) this method is still inferior compared to only using the MDP method.

3.5 Descriptions of Experiments

Our experiments were centered around Phrase Based SMT (PBSMT). We used the open source Moses decoder (Hoang et al., 2007) package (including Giza++) for word alignment, phrase table extraction and decoding for sentence translation. We also used the Moses scripts for linear and fillup interpolation along with the multiple decoding paths (MDP) setting (by modifying the moses.ini files). We performed MERT (Och, 2003) based tuning using the MIRA algorithm. We used BLEU (Papineni et al., 2002) as our evaluation criteria and the bootstrapping method (Koehn, 2004) for significance testing. For the sake of comparison with previous methods, we experimented with sentence translation strategy (Utiyama and Isahara, 2007) using 10 as the n-best list size for intermediate and target language translations. The experiments we performed are given below. Each experiment involves either the creation of a phrase tables or combination of phrase tables. We tune, test and evaluate these tables or combinations.

1. A src (source) to tgt (target) direct phrase table.
2. For piv in Pivot_Languages_Set; the set of pivot languages to be used (Tables 1 and 2):

[7] By Oracle scenario we mean that we already know the performance on the test sets and exploit this information to "unfairly" boost the translation scores.

(a) src to piv and piv to tgt phrase tables. Translate the src test sentences to tgt using the sentence translation strategy and evaluate. (Column 2)

(b) Triangulate the src-piv and piv-tgt phrase tables to get the src-piv-tgt phrase table. (Column 3)

(c) Perform linear interpolation of the src-tgt and src-piv-tgt table using 9:1 weight ratio in equation 5 to get a combined table. (Column 4)

(d) Perform linear interpolation of the src-tgt and src-piv-tgt table using the ratio of their BLEU scores as weights in equation 5 to get a combined table. (Column 5)

(e) Perform fillup interpolation of the src-tgt (main) and src-piv-tgt table (secondary) to get a combined table. (Column 6)

(f) Combine the src-tgt and src-piv-tgt phrase table using MDP (2 paths, 1 for direct and 1 for pivot). (Column 7)

3. Combine **all** the src-piv-tgt tables into a single table using linear (weights are ratios of BLEU scores) and fillup interpolation independently, giving the phrase tables: linear_interp_all and fill_interp_all respectively. Table 3, rows 4 and 5.

4. Perform linear interpolation of the src-tgt and linear_interp_all tables using 9:1 weight ratio in equation 5 to get a combined table. Table 3, row 6.

5. Perform linear interpolation of the src-tgt and **all** src-piv-tgt phrase tables using the ratio of their BLEU scores as weights in equation 5 to get a combined table. Table 3, row 7.

6. Perform fillup interpolation of the src-tgt and **all** src-piv-tgt phrase tables. The priority of the tables is given by the descending order of BLEU scores. Table 3, row 8.

7. Combine the linear_interp_all with the src-tgt phrase table using MDP. Repeat this for fill_interp_all. Table 3, rows 9 and 10.

8. Combine **all** the src-piv-tgt phrase tables with the src-tgt phrase table using MDP (8 paths, 1 for direct and 1 for each of the 7 pivots). Table 3, row 11.

9. Combine the **top 3** pivot phrase tables with the src-piv-tgt phrase tables with the src-tgt phrase table using MDP (4 paths, 1 for direct and 1 for each of the 3 pivots). The pivot tables with the 3 highest[8] standalone BLEU scores are selected. Table 3, row 12.

4 Results and Discussions

BLEU scores obtained after testing the tuned tables are reported. Scores in bold are statistically significant ($p<0.05$) over the baseline which is the system trained using a direct src-tgt parallel corpus.

4.1 Results

The Japanese-Hindi direct translation system gave a BLEU of 33.86 whereas the Hindi-Japanese one gave 37.47. For the rest of the paper these will be the baselines, unless mentioned otherwise.

The evaluation scores are split into 3 tables. Table 1 contains the scores for Japanese to Hindi (Table 2 for Hindi to Japanese) translation using each pivot separately and has 7 columns whose details are given in section 3.5 from 2.a to 2.f. Table 3 contains the scores for Japanese to Hindi (and vice versa) translation using all 7 pivots together in various ways. Each row is self explanatory. In row 6, we mean that the direct phrase table has a weight of 0.9 and the remainder 0.1 is distributed amongst the pivot phrase tables in the ratio of their standalone BLEU scores which can be seen in column 3 of tables 1 and 2. It is quite clear that sentence translation strategy is the most inferior technique.

4.2 Observations

Below, we give an explanation of the observed scores from various points of views.

4.2.1 On the Pivots Used

It is logical to consider that the closeness of a pivot language to the source or target is an important factor in the improvement of translation quality, since Korean helps Japanese-Hindi translation.

[8] We chose 3 since our evaluation showed that the BLEU scores for the 3 pivot languages were much larger than the remaining ones.

Pivot Language	Sentence Strategy	Standalone	Linear Interpolate (1) With Direct	Linear Interpolate (2) With Direct	Fill Interpolate With Direct	MDP With Direct
1. Direct	**33.86**					
2. Chinese	23.53	28.89	34.03	**34.61**	34.31	**35.66**
3. Korean	26.30	28.92	**34.65**	34.18	34.64	**35.60**
4. Esperanto	22.43	28.73	**34.63**	**34.55**	35.32	**35.74**
5. Paite	19.40	26.64	34.17	34.40	**34.66**	**35.22**
6. Marathi	15.68	21.80	33.88	33.80	33.83	34.03
7. Kannada	16.94	24.15	33.74	34.13	**34.87**	**35.52**
8. Telugu	14.15	21.31	33.81	33.85	34.04	**34.57**

Table 1: Japanese-Hindi Results Using Single Pivots

Pivot Language	Sentence Strategy	Standalone	Linear Interpolate (1) With Direct	Linear Interpolate (2) With Direct	Fill Interpolate With Direct	MDP With Direct
1. Direct	**37.47**					
2. Chinese	27.93	30.97	35.90	**38.47**	**38.41**	**39.49**
3. Korean	30.68	32.67	35.99	**38.72**	**38.55**	**39.49**
4. Esperanto	26.67	30.80	36.07	37.82	37.85	**39.14**
5. Paite	23.37	29.17	35.89	37.73	37.39	**38.19**
6. Marathi	20.59	26.21	35.89	37.57	37.72	**38.30**
7. Kannada	23.21	26.96	35.84	38.05	37.79	**38.05**
8. Telugu	19.01	25.22	37.25	36.98	37.11	37.04

Table 2: Hindi-Japanese Results Using Single Pivots

Of all the scores, the ones obtained using Korean and Chinese as pivots stand out as the best and it is known that Korean and Japanese share many similarities. Although this gives reason to believe that languages belonging to the same language group should act as good choices of pivots, the languages Kannada, Telugu and (especially) Marathi should have helped improve Hindi to Japanese translation. Moreover, languages like Paite and Esperanto which are relatively distant from both Hindi and Japanese gave better performance than the Indian Languages. Remember that the Chinese and Korean corpora were character segmented[9] and that Esperanto and Paite are not so morphologically rich. The Indian pivot languages have agglutinative features which is one of the main causes of poor quality SMT. This clearly indicates that morphological similarity to source and target is another equally important aspect that affects the translation quality. Had this not been the case, the Indian Languages would have acted as good pivots. This shows that experiments involving forcing the morpheme to morpheme ratio, of the source to pivot to target sentences, to be the same, must be conducted. Henceforth, it is to be expected that the most significant improvements will be obtained when Chinese, Korean and Esperanto (in a number of cases) are used as pivots.

[9]Hangul blocks were space separated in the Korean case.

4.2.2 On the Linear and Fill Interpolation Methods

Single pivots: All the interpolation methods (columns 4, 5 and 6 of Tables 1 and 2) gave small improvements in BLEU in most cases compared to the baselines for both language pairs. The results do not show drastic improvements, which is expected since the baseline and pivots based phrase tables are constructed from the same multilingual

Combination Type	Jap-Hin	Hin-Jap
1. Direct phrase table (baseline)	33.86	37.47
2. Best result using single pivot	35.74 (Esp.)	39.49 (Kor.)
3. Combine All Pivots using MDP	34.49	37.02
4. A - Linear Interpolate All Pivot tables (BLEU score ratio)	32.50	35.65
5. B - Fill Interpolate All Pivot tables (Priority according to BLEU score)	32.12	34.44
6. Linear Interpolate (9:1 ratio) Direct with All Pivot tables	**34.56**	**38.60**
7. Linear Interpolate (BLEU score ratio) Direct with All Pivots	**35.24**	**39.08**
8. Fill Interpolate Direct with All Pivots (Priority according to BLEU score)	**35.28**	**38.70**
9. Combine Direct and A using MDP	**36.40**	**39.85**
10. Combine Direct and B using MDP	**36.67**	**40.07**
11. Combine Direct and All Pivots tables using MDP	**38.42**	**40.19**
12. Combine Direct and Top 3 (BLEU) pivot tables using MDP (Oracle)	**38.22**	**41.09**

Table 3: Results Using Multiple Pivots With Different Combination Methods

training instances (29000 tuples - see section 3.2). Typically the interpolation methods are shown to give substantial performance boosts when the direct source-target phrase table is obtained using relatively smaller corpora sizes compared to those used for the source-pivot and pivot-target tables. In case of linear interpolation with a 9:1 weight ratio, the scores improve slightly in some cases for Japanese-Hindi but degrade in case of Hindi-Japanese. However, in the case of linear interpolation where the BLEU score ratio is used as the weight ratio, the improvements are much better[10].

Fill based interpolation also gives improvements in some cases, mostly when Chinese and Korean are used as pivots. An overall comparison shows that there is no consistency when a single pivot language is used and no conclusive comment can be made on the efficacy of these interpolation methods.

Multiple Pivots: However in Table 3, rows 6 to 8 show that using all the pivots together, result in a significant improvement over the direct phrase tables. Linear interpolation with BLEU score ratio gives 35.24 BLEU (33.86 for direct phrase table) for Japanese-Hindi and 39.08 BLEU (37.47 for direct phrase table). Rows 4 and 5 show the scores of the linear and fill interpolation of only the pivot based phrase tables. It is interesting to see that in case of Japanese-Hindi the BLEU scores rival that of the direct phrase table (32.50/32.12 v.s. 33.86). This is similar in the case of Hindi-Japanese: 35.65/34.44 v.s. 37.47. The following points must be noted:

a. Since the setting is multilingual and improvements, however slight, are observed in some cases it must be the case that, through pivoting, additional (and possibly improved) phrase pairs are induced which are not extracted using the direct source-target parallel corpus. This also gives reason to believe that every pivot induces a different set of phrase pairs thereby overcoming the limitations of poor alignment (and effectively phrase extraction) on small corpora. Even if there is no alignment error, pivoting still introduces new phrase pairs which improves MT performance.

b. The pivot based phrase tables already have an incomplete probability space with respect to the phrase pair distribution. Linear interpolation tends to violate the overall probability mass since the phrase pair distribution gets changed. Fill interpolation just adds additional phrase pairs from the next phrase table when not available in the current one which leads to poor mixing of different probability models giving poorer performance in-spite of additional phrase pairs being available.

c. Since some pivot languages are obviously bad, their probability scores would drastically affect the overall probability mass. They should be excluded or given low weights, which we do by considering the BLEU score ratio. However, this is not a good idea because the scores for Telugu, a bad pivot for Hindi-Japanese translation, degraded to a lesser ex-

[10]Expected as we use test set evaluation information.

tent when the Telugu based phrase table was linearly combined with the direct phrase table. Sennrich (2012) gave a method to learn these weights, but in a resource constrained scenario such a method is difficult to apply.

This motivated us to try the Multiple Decoding Paths (MDP) feature of Moses.

4.2.3 On using MDP

Single pivots: Since log linear combination does not modify the probability space it should lead to definitive increase in translation scores. This claim is validated by the last columns of Tables 1 and 2. For Japanese-Hindi: barring Marathi, the combination of the direct and pivot phrase table leads to significant improvement over the direct phrase tables. A similar situation occurs for Hindi-Japanese except that Telugu behaves as a bad pivot.

Multiple pivots: Row 3 of Table 3 indicates that the log linear combination of all the pivot tables using MDP for Japanese-Hindi gives a BLEU of 34.49, an improvement ($p<0.05$) over the direct table (BLEU 33.86). For Hindi-Japanese, although the equivalent BLEU score (37.02) is not an improvement over that of the direct table (37.47), it does show that multiple pivots can be used to achieve translation quality similar to the quality obtained by a direct table.

Since it was observed that the interpolation of all the pivot tables into a single one gave scores close to the direct tables we decided to try the combination of the all pivots interpolated table with the direct table using MDP. Rows 9 and 10 show that there is a significant improvement compared to the scores of the direct tables alone. But this method of linear + log linear combination would still suffer from the limitation of linear interpolation which led to the final 2 experiments which use only log linear combination.

Row 11 shows that the method of combining the direct and all the pivot tables using MDP (one for each table) outperforms all the methods so far. The reason is simple: Only good translation options are collected from all tables during hypothesis expansion, the bad ones are automatically pruned. For Japanese-Hindi the BLEU is 38.42 which is an improvement of 4.56 (13% relative) over the BLEU of the direct phrase table (33.86). For Hindi-Japanese the BLEU of 40.19 is an improvement of 2.72 (7.25% relative) over that of the direct table (37.47). The increment is lesser because of the premise we established in section 4.2.1. This points to an interesting observation that pivot languages induce better phrase pairs in a multilingual setting which are not present in the direct phrase table. This is quite beneficial when the corpora sizes are small which lead to poor quality phrase tables.

To test whether exclusion of bad performing pivots leads to improvements in BLEU we performed another oracle experiment in which we only included the pivot phrase tables having significant standalone BLEU difference compared to the others. Korean, Chinese and Esperanto were the ones that stood out. The last row shows that for Japanese-Hindi the BLEU (38.22) does not significantly increase over the situation when all pivots are used together (38.42). However for Hindi-Japanese the BLEU is 41.09 which is a significant ($p<0.05$) increase compared to when all the pivots are used together (40.19 - 2.2% relative). Note that this leads to an absolute BLEU difference of 3.62 (9.66% relative) compared to the BLEU of the direct phrase table. The improvements for Japanese-Hindi were already so large (13%) that more significant improvements would need deeper inspection and improved methods. We believe that further significant improvements are possible and advanced methods to effectively select multiple pivots need to be studied and implemented.

4.2.4 On the number of new phrase pairs induced

Based on the cutoff of 0.001 for the inverse translation probability, Table 4 contains the statistics of the unique phrase pairs in each pivot table (Columns 4 to 10) and the direct table (Column 3) along with the number of phrase pairs common (Column 2) to all. It is quite obvious that each pivot[11] induces its own set of unique phrase pairs.

4.2.5 On the improvement in translations

Table 5 gives the count of improved translations, out of 500 tested sentences, over the direct using sentence level BLEU difference at various cutoffs. On an average 50% of the sentences showed increase

[11] For each language we use their first 3 characters of their names as the shortened versions.

Direction	Common	Direct	Chi	Kor	Esp	Pai	Kan	Mar	Tel
1. Jap-Hin	0.032	1.404	20.74	18.65	16.06	23.85	26.56	30.92	26.84
2. Hin-Jap	0.034	1.528	26.20	20.26	18.06	28.83	29.90	36.98	31.23

Table 4: Unique phrase pairs in each table (in millions of pairs)

Direction	>0	>0.1	>0.2	>0.3	>0.4	>0.5	>0.6	>0.7
1. Jap-Hin	267	108	36	12	6	4	2	0
2. Hin-Jap	275	124	60	24	12	4	1	1

Table 5: Number of improved translations (out of 500) using sentence level BLEU difference at various cutoffs

in BLEU and the number of improved sentences decreases with increasing cutoff. We manually verified a random sample and found that the improvements were commensurate with the reported differences. Finally, consider an example of improvement in Japanese to Hindi translation.

Input: それから、ヨハネの弟子たちがきる、死体を引き取る葬る。そして、イエスのところに行く報告する。 (Sore kara, Yohane no deshi-tachi ga kiru, shitai o hikitoru homuru. Soshite, iesu no tokoro ni iku hokoku suru.)

English equivalent: After that, John's disciples came and took his corpse away, buried it and then went to Jesus to give him the news.

Direct translation: तब यूहन्ना के चेलों ने आकर उस की लोथ ले गया और उसके पास जाकर बता दिया (Tab yohanna ke chelo ne aakar uss ki lotha le gaya; aura uske paas jaakar bata diya)

Best translation using MDP: तब यूहन्ना के चेलों ने आकर उस की लोथ को ले जाकर गाढ दिया और जाकर यीशु को समाचार दिया (Tab yohanna ke chelo ne aakar usa ki lotha ko le jaakar gaad diya aura jakar yesu ko samachara diya)

Analysis: Note that in the direct translation the part about "burying the corpse" (gaad diya) and "Jesus" (yesu) is missing which is present in the MDP translation. Also the verb forms indicating the sequence of actions like "came and" (aakar) and "took his corpse away" (usa ki lotha ko le jaakar) are much better in the MDP translation. Instead of "samachara diya" (gave news) the preferred translation is "samachara di".

5 Conclusions and Future Work

We have presented our work on leveraging a small sized multilingual parallel corpus using 7 pivot languages for SMT between Japanese and Hindi. Our main objective was to augment a phrase table on direct parallel corpus using many pivot language based phrase tables constructed from the same multilingual corpus. We confirm that this induces additional and improved phrase pairs which, under the Multiple Decoding Paths setting (MDP), leads to substantial improvements over the direct phrase tables. More importantly, we show that using multiple pivot languages simultaneously lead to large improvements in BLEU compared to the when a single pivot is used; which is the novel aspect of our work. This opens up many further research directions like **a.** How can one choose a set of good pivot languages amongst available choices? **b.** Does this multilingual leveraging help in a situation where we have large size corpora like Europarl corpora? **c.** How much of an impact can treatment (morphological or syntactic) of the pivot language help in improving translation quality? **d.** Can good reordering information be extracted by pivoting? **e.** Can multi source and multi pivot setting further enhance quality? **f.** How can the noise induced by pivoting be controlled by methods other than probability cutoffs? and finally **g.** Can simpler but more effective methods compared to triangulation be exploited in a multilingual scenario? The last 4 questions have long been ignored and deserve to be answered. We will pursue these directions in the future and attempt to provide satisfactory answers.

References

Alexandra Birch and Miles Osborne. 2007. Ccg supertags in factored statistical machine translation. In *In ACL Workshop on Statistical Machine Translation*, pages 9–16.

Manoj K. Chinnakotla, Karthik Raman, and Pushpak Bhattacharyya. 2010. Multilingual pseudo-relevance feedback: Performance study of assisting languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1346–1356, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Selective combination of pivot and direct statistical machine translation models. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1174–1180. Asian Federation of Natural Language Processing.

Tim Gollins and Mark Sanderson. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 90–95, New York, NY, USA. ACM.

Nizar Habash and Jun Hu. 2009. Improving arabic-chinese statistical machine translation using english as pivot language.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation". acl-2007.

Kazuaki Kishida and Noriko Kando. 2003. Two-stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: An experiment at clef 2003. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *CLEF*, volume 3237 of *Lecture Notes in Computer Science*, pages 253–262. Springer.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger, 2009. *462 Machine Translation Systems for Europe*, pages 65–72. Association for Machine Translation in the Americas, AMTA.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multiway translation of indian languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1781–1787, Reykjavik, Iceland, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1355.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of japanese morphological analyzer juman. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.

Preslav Nakov and Hwee Tou Ng. 2009. Improved statistical machine translation for resource-poor languages using related resource-rich languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pages 1358–1367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *In MT Summit 2001*, pages 253–258.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 221–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Paul, Andrew Finch, and Eiichrio Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. 12(4):14:1–14:17, October.

Philip Resnik, MariBroman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the

'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778. Association for Computational Linguistics.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *in Proceedings of the conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL-HLT*, pages 484–491.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181, September.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 154–162, Stroudsburg, PA, USA. Association for Computational Linguistics.