Exploring System Combination approaches for Indo-Aryan MT Systems

Karan Singla[1], **Nishkarsh Shastri**[2], **Megha Jhunjhunwala**[2], **Anupam Singh**[3], **Srinivas Bangalore**[4], **Dipti Misra Sharma**[1]

[1]LTRC IIIT Hyderabad, [2]IIT-Kharagpur, [3]NIT-Durgapur,[4]AT&T Labs-Research

Abstract

Statistical Machine Translation (SMT) systems are heavily dependent on the quality of parallel corpora used to train translation models. Translation quality between certain Indian languages is often poor due to the lack of training data of good quality. We used triangulation as a technique to improve the quality of translations in cases where the direct translation model did not perform satisfactorily. Triangulation uses a third language as a pivot between the source and target languages to achieve an improved and more efficient translation model in most cases. We also combined multi-pivot models using linear mixture and obtained significant improvement in BLEU scores compared to the direct source-target models.

1 Introduction

Current SMT systems rely heavily on large quantities of training data in order to produce good quality translations. In spite of several initiatives taken by numerous organizations to generate parallel corpora for different language pairs, training data for many language pairs is either not yet available or is insufficient for producing good SMT systems. Indian Languages Corpora Initiative (ILCI) (Choudhary and Jha, 2011) is currently the only reliable source for multilingual parallel corpora for Indian languages however the number of parallel sentences is still not sufficient to create high quality SMT systems.

This paper aims at improving SMT systems trained on small parallel corpora using various recently developed techniques in the field of SMTs. Triangulation is a technique which has been found to be very useful in improving the translations when multilingual parallel corpora are present. Triangulation is the process of using an intermediate language as a pivot to translate a source language to a target language. We have used phrase table triangulation instead of sentence based triangulation as it gives better translations (Utiyama and Isahara, 2007). As triangulation technique explores additional multi parallel data, it provides us with separately estimated phrase-tables which could be further smoothed using smoothing methods (Koehn et al. 2003). Our subsequent approach will explore the various system combination techniques through which these triangulated systems can be utilized to improve the translations.

The rest of the paper is organized as follows. We will first talk about the some of the related works and then we will discuss the facts about the data and also the scores obtained for the baseline translation model. Section 3 covers the triangulation approach and also discusses the possibility of using combination approaches for combining triangulated and direct models. Section 4 shows results for the experiments described in previous section and also describes some interesting observations from the results. Section 5 explains the conclusions we reached based on our experiments. We conclude the paper with a section about our future work.

2 Related Works

There are various works on combining the triangulated models obtained from different pivots with the direct model resulting in increased confidence score for translations and increased coverage by (Razmara and Sarkar, 2013; Ghannay et al., 2014; Cohn and Lapata, 2007). Among these techniques we explored two of the them. The first one is the technique based on the confusion matrix (dynamic) (Ghannay et al., 2014) and the other one is based on mixing the models as explored by (Cohn and Lapata, 2007). The paper also discusses the better choice of combination technique

Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), pages 85–91,
October 29, 2014, Doha, Qatar.

among these two when we have limitations on training data which in our case was small and restricted to a small domain (Health & Tourism).

As suggested in (Razmara and Sarkar, 2013), we have shown that there is an increase in phrase coverage when combining the different systems. Conversely we can say that out of vocabulary words (OOV) always decrease in the combined systems.

3 Baseline Translation Model

In our experiment, the baseline translation model used was the direct system between the source and target languages which was trained on the same amount of data as the triangulated models. The parallel corpora for 4 Indian languages namely Hindi (hn), Marathi (mt), Gujarati (gj) and Bangla (bn) was taken from Indian Languages Corpora Initiative (ILCI) (Choudhary and Jha, 2011) . The parallel corpus used in our experiments belonged to two domains - health and tourism and the training set consisted of 28000 sentences. The development and evaluation set contained 500 sentences each. We used MOSES (Koehn et al., 2007) to train the baseline Phrase-based SMT system for all the language pairs on the above mentioned parallel corpus as training, development and evaluation data. Trigram language models were trained using SRILM (Stolcke and others, 2002). Table 1 below shows the BLEU score for all the trained pairs.

Language Pair	BLEU Score
bn-mt	18.13
mt-bn	21.83
bn-gj	22.45
gj-mt	23.02
gj-bn	24.26
mt-gj	25.5
hn-mt	30.01
hn-bn	32.92
bn-hn	34.99
mt-hn	36.82
hn-gj	40.06
gj-hn	43.48

Table 1: BLEU scores of baseline models

4 Triangulation: Methodology and Experiment

We first define the term *triangulation* in our context. Each source phrase `s` is first translated to an intermediate (pivot) language `i`, and then to a target language `t`. This two stage translation process is termed as triangulation.

Our basic approach involved making triangulated models by triangulating through different pivots and then interpolating triangulated models with the direct source-target model to make our combined model.

In line with various previous works, we will be using multiple translation models to overcome the problems faced due to data sparseness and increase translational coverage. Rather than using sentence translation (Utiyama and Isahara, 2007) from source to pivot and then pivot to target, a phrase based translation model is built.

Hence the main focus of our approach is on phrases rather than on sentences. Instead of using combination techniques on the output of several translation systems, we constructed a combined phrase table to be used by the decoder thus avoiding the additional inefficiencies observed while merging the output of various translation systems. Our method focuses on exploiting the availability of multi-parallel data, albeit small in size, to improve the phrase coverage and quality of our SMT system.

Our approach can be divided into different steps which are presented in the following sections.

4.1 Phrase-table triangulation

Our emphasis is on building an enhanced phrase table that incorporates the translation phrase tables of different models. This combined phrase table will be used by the decoder during translation.

Phrase table triangulation depends mainly on phrase level combination of the two different phrase based systems mainly source (src) - pivot (pvt) and pivot (pvt) - target (tgt) using pivot language as a basis for combination. Before stating the mathematical approach for triangulation, we present an example.

4.1.1 Basic methodology

Suppose we have a Bengali-Hindi phrase-table (T_{BH}) and a Hindi-Marathi phrase-table (T_{HM}). From these tables, we have to construct a Bengali-Marathi phrase-table (T_{BM}). For that we need

Triangulated System	Full-Triangulation (phrase-table length)	Triangulation with top 40 (Length of phrase table)	Full Triangulation (BLEU Score)	Triangulation with top 40 (BLEU SCORE)
gj - hn - mt	3,585,450	1,086,528	24.70	24.66
gj - bn - mt	7,916,661	1,968,383	20.55	20.04

Table 2: Comparison between triangulated systems in systems with full phrase table and the other having top 40 phrase-table entries

to estimate four feature functions: phrase translation probabilities for both directions $\phi(\bar{b}|\bar{m})$ and $\phi(\bar{m}|\bar{b})$, and lexical translation probabilities for both directions $lex(\bar{b}|\bar{m})$ and $lex(\bar{m}|\bar{b})$ where $\bar{b}$ and $\bar{m}$ are Bengali and Marathi phrases that will appear in our triangulated Bengali-Marathi phrase-table T_{BM}.

$$\phi(\bar{b}|\bar{m}) = \sum_{\bar{h} \in T_{BH} \cap T_{HM}} \phi(\bar{b}|\bar{h})\phi(\bar{h}|\bar{m}) \quad (1)$$

$$\phi(\bar{m}|\bar{b}) = \sum_{\bar{h} \in T_{BH} \cap T_{HM}} \phi(\bar{m}|\bar{h})\phi(\bar{h}|\bar{b}) \quad (2)$$

$$lex(\bar{b}|\bar{m}) = \sum_{\bar{h} \in T_{BH} \cap T_{HM}} lex(\bar{b}|\bar{h})lex(\bar{h}|\bar{m}) \quad (3)$$

$$lex(\bar{m}|\bar{b}) = \sum_{\bar{h} \in T_{BH} \cap T_{HM}} lex(\bar{m}|\bar{h})lex(\bar{h}|\bar{b}) \quad (4)$$

In these equations a conditional independence assumption has been made that source phrase $\bar{b}$ and target phrase $\bar{m}$ are independent given their corresponding pivot phrase(s) $\bar{h}$. Thus, we can derive $\phi(\bar{b}|\bar{m}), \phi(\bar{m}|\bar{b}), lex(\bar{b}|\bar{m}), lex(\bar{m}|\bar{b})$ by assuming that these probabilities are mutually independent given a Hindi phrase $\bar{h}$.

The equation given requires that all phrases in the *Hindi-Marathi* bitext must also be present in the *Bengali-Hindi* bitext. Clearly there would be many phrases not following the above requirement. For this paper we completely discarded the missing phrases. One important point to note is that although the problem of missing contextual phrases is uncommon in multi-parallel corpora, as it is in our case, it becomes more evident when the bitexts are taken out from different sources.

In general, wider range of possible translations are found for any source phrase through triangulation. We found that in the direct model, a source phrase is aligned to three phrases then there is high possibility of it being aligned to three phrases in intermediate language. The intermediate language phrases are further aligned to three or more phrases in target language. This results in increase in number of translations of each source phrase.

4.1.2 Reducing the size of phrase-table

While triangulation is intuitively appealing, it suffers from a few problems. First, the phrasal translation estimates are based on noisy automatic word alignments. This leads to many errors and omissions in the phrase-table. With a standard source-target phrase-table these errors are only encountered once, however with triangulation they are encountered twice, and therefore the errors are compounded. This leads to much noisier estimates than in the source-target phrase-table. Secondly, the increased exposure to noise means that triangulation will omit a greater proportion of large or rare phrases than the standard method. An alignment error in either of the source-intermediate bitext or intermediate-target bitext can prevent the extraction of a source-target phrase pair.

As will be explained in the next section, the second kind of problem can be ameliorated by using the triangulated phrase-based table in conjunction with the standard phrase based table referred to as direct *src-to-pvt* phrase table in our case.

For the first kind of problem, not only the compounding of errors leads to increased complexity but also results in an absurdly large triangulated phrase based table. To tackle the problem of unwanted phrase-translation, we followed a novel approach.

A general observation is that while triangulating between *src-pvt* and *pvt-tgt* systems, the resultant src-tgt phrase table formed will be very large since for a translation $\bar{s}$ to $\bar{i}$ in the *src-to-pvt* table there may be many translations from $\bar{i}$ to $\bar{t}1, \bar{t}2...\bar{t}n$. For example, the Bengali-Hindi phrase-table(T_{BH}) consisted of 846,106 translations and Hindi-Marathi phrase-table(T_{HM}) consisted of 680,415 translations and after triangulating these two tables our new Bengali-Marathi triangulated table(T_{BM}) consisted of 3,585,450 translations as shown in Table 2. Tuning with such a large phrase-table is complex and time-consuming. To reduce the complexity of the phrase-table, we used only the top-40 transla-

tions (translation with 40 maximum values of $P(\bar{f}|\bar{e})$ for every source phrase in our triangulated phrase-table(T_{BM}) which reduced the phrase table to 1,086,528 translations.

We relied on $P(\bar{f}|\bar{e})$(inverse phrase translation probability) to choose 40 phrase translations for each phrase, since in the direct model, MERT training assigned the most weight to this parameter.

It is clearly evident from Table 2 that we have got a massive reduction in the length of the phrase-table after taking in our phrase table and still the results have no significant difference in our output models.

4.2 Combining different triangulated models and the direct model

Combining Machine translation (MT) systems has become an important part of Statistical MT in the past few years. There have been several works by (Rosti et al., 2007; Karakos et al., 2008; Leusch and Ney, 2010);
We followed two approaches

1. A system combination based on confusion network using open-source tool kit **MANY** (Barrault, 2010), which can work dynamically in combining the systems

2. Combine the models by linearly interpolating them and then using MERT to tune the combined system.

4.2.1 Combination based on confusion matrix

MANY tool was used for this and initially it was configured to work with TERp evaluation matrix, but we modified it to work using METEOR-Hindi (Gupta et al., 2010), as it has been shown by (Kalyani et al., 2014), that METEOR evaluation metric is closer to human evaluation for morphologically rich Indian Languages.

4.2.2 Linearly Interpolated Models

We used two different approaches while merging the different triangulated models and direct src-tgt model and we observed that both produced comparable results in most cases. We implemented the linear mixture approach, since linear mixtures often outperform log-linear ones (Cohn and Lapata, 2007). Note that in our combination approaches the reordering tables were left intact.

1. Our first approach was to use linear interpolation to combine all the three models (Bangla-Hin-Marathi, Bangla-Guj-Marathi and direct Bangla-Marathi models) with uniform weights, i.e 0.3 each in our case.

2. In the next approach, the triangulated phrase tables are combined first into a single triangulated phrase-table using uniform weights. The combined triangulated phrase-table and direct src-tgt phrase table is then combined using uniform weights. In other words, we combined all the three systems, Ban-Mar, Ban-Hin-Mar, and Ban-Guj-Mar with 0.5, 0.25 and 0.25 weights respectively. This weight distribution reflects the intuition that the direct model is less noisy than the triangulated models.

In the experiments below, both weight settings produced comparable results. Since we performed triangulation only through two languages, we could not determine which approach would perform better. An ideal approach will be to train the weights for each system for each language pair using standard tuning algorithms such as MERT (Zaidan, 2009).

4.2.3 Choosing Combination Approach

In order to compare the approaches on our data, we performed experiments on Hindi-Marathi pair following both approaches discussed in Section 4.2.1 and 4.2.2. We also generated triangulated models through Bengali and Gujarati as pivot languages.

Also, the approach presented in section 4.2.1 depends heavily on LM (Language Model).In order to study the impact of size, we worked on training Phrase-based SMT systems with subsets of data in sets of 5000, 10000, 150000 sentences and LM was trained for 28000 sentences for comparing these approaches. The combination results were compared following the approach mentioned in 4.2.1 and 4.2.2.

Table 3, shows that the approach discussed in 4.2.1 works better if there is more data for LM but we suffer from the limitation that there is no other in-domain data available for these languages. From the Table, it can also be seen that combining systems with the approach explained in 4.2.2 can also give similar or better results if there is scarcity of data for LM. Therefore we followed the

#Training	#LM Data	Comb-1	Comb-2
5000	28000	21.09	20.27
10000	28000	24.02	24.27
15000	28000	27.10	27.63

Table 3: BLEU scores for Hindi-Marathi Model comparing approaches described in 3.2.1(Comb-1) and 3.2.2(Comb-2)

approach from Section 4.2.2 for our experiments on other language pairs.

5 Observation and Resuslts

Table 4, shows the BLEU scores of triangulated models when using the two languages out of the 4 Indian languages Hin, Guj, Mar, Ban as source and target and the remaining two as the pivot language. The first row mentions the BLEU score of the direct src-tgt model for all the language pairs. The second and third rows provide the triangulated model scores through pivots which have been listed. The fourth and fifth rows show the BLEU scores for the combined models (triangulated+direct) with the combination done using the first and second approach respectively that have been elucidated in the Section 4.2.2

As expected, both the combined models have performed better than the direct models in all cases.

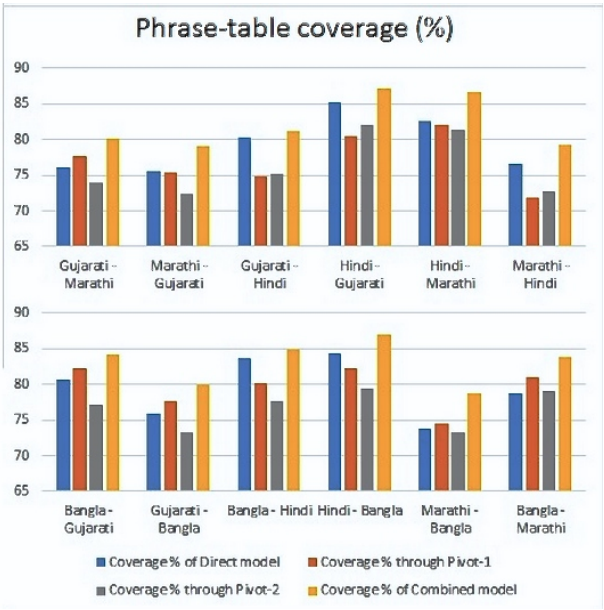

Figure 1: Phrase-table coverage of the evaluation set for all the language pairs

Figure 1, shows the phrase-table coverage of the evaluation set for all the language pairs. Phrase-table coverage is defined as the percentage of unigrams in the evaluation set for which translations are present in the phrase-table. The first bar corresponds to the direct model for each language pair, the second and third bars show the coverage for triangulated models through the 2 pivots, while the fourth bar is the coverage for the combined model (direct+triangulated). The graph clearly shows that even though the phrase table coverage may increase or decrease by triangulation through a single pivot the combined model (direct+triangulated) always gives a higher coverage than the direct model.

Moreover, there exists some triangulation models whose coverage and subsequent BLEU scores for translation is found to be better than that of the direct model. This is a particularly interesting observation as it increases the probability of obtaining better or at least comparable translation models even when direct source-target parallel corpus is absent.

6 Discussion

Dravidian languages are different from Indo-aryan languages but they are closely related amongst themselves. So we explored similar experiments with Malayalam-Telugu pair of languages with similar parallel data and with Hindi as pivot.

The hypothesis was that the direct model for Malayalam-Telegu would have performed better due to relatedness of the two languages. However the results via Hindi were better as can be seen in Table 5.

As Malayalam-Telegu are comparatively closer than compared to Hindi, so the results via Hindi should have been worse but it seems more like a biased property of training data which considers that all languages are closer to Hindi, as the translation data was created from Hindi.

7 Future Work

It becomes increasingly important for us to improve these techniques for such languages having rare corpora. The technique discussed in the paper is although efficient but still have scope for improvements.

As we have seen from our two approaches of combining the phrase tables and subsequent interpolation with direct one, the best combination among the two is also not fixed. If we can find the

BLEU scores	gj-mt		mt-gj		gj-hn		hn-gj		hn-mt		mt-hn	
Direct model	23.02		25.50		43.48		40.06		30.01		36.82	
Triangulated through pivots	**hn**	24.66	**hn**	27.09	**mt**	36.76	**mt**	33.69	**gj**	29.27	**gj**	33.86
	bn	20.04	**bn**	22.02	**bn**	35.07	**bn**	32.66	**bn**	26.72	**bn**	31.34
Mixture-1	26.12		**27.46**		43.23		39.99		33.09		**38.50**	
Mixture-2	**26.25**		27.32		**44.04**		**41.45**		**33.36**		38.44	

(a)

BLEU scores	bn-gj		gj-bn		bn-hn		hn-bn		mt-bn		bn-mt	
Direct model	22.45		24.26		34.99		32.92		21.83		18.13	
Triangulated through pivots	**hn**	23.97	**hn**	26.26	**gj**	31.69	**gj**	29.60	**hn**	23.80	**hn**	21.04
	mt	20.70	**mt**	22.32	**mt**	28.96	**mt**	27.95	**gj**	22.41	**gj**	18.15
Mixture-1	**25.80**		**27.45**		**35.14**		34.77		**24.99**		22.16	
Mixture-2	24.66		27.39		35.02		**34.85**		24.86		**22.75**	

(b)

Table 4: Table (a) & (b) show results for all language pairs after making triangulated models and then combining them with linear interpolation with the two approaches described in 3.2.2. In *Mixture-1*, uniform weights were given to all three models but in *Mixture-2*, direct model is given 0.5 weight relative to the other models (.25 weight to each)

System	Blue Score
Direct Model	4.63
Triangulated via Hindi	14.32

Table 5: Results for Malayalam-Telegu Pair for same data used for other languages

best possible weights to be assigned to each table, then we can see improvement in translation. This can be implemented by making the machine learn from various iterations of combining and adjusting the scores accordingly.(Nakov and Ng, 2012) have indeed shown that results show significant deviations associated with different weights assigned to the tables.

References

Loïc Barrault. 2010. Many: Open source machine translation system combination. *The Prague Bulletin of Mathematical Linguistics*, 93:147–155.

Narayan Choudhary and Girish Nath Jha. 2011. Creating multilingual parallel corpora in indian languages. In *Proceedings of Language and Technology Conference*.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 728. Citeseer.

Sahar Ghannay, France Le Mans, and Loıc Barrault. 2014. Using hypothesis selection based features for confusion network mt system combination. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 1–5.

Ankush Gupta, Sriram Venkatapathy, and Rajeev Sangal. 2010. Meteor-hindi: Automatic mt evaluation metric for hindi as a target language. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.

Aditi Kalyani, Hemant Kumud, Shashi Pal Singh, and Ajai Kumar. 2014. Assessing the quality of mt systems for hindi to english translation. *arXiv preprint arXiv:1404.3992*.

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, and Markus Dreyer. 2008. Machine translation system combination using itg-based alignments. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 81–84. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Gregor Leusch and Hermann Ney. 2010. The rwth system combination system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical*

Machine Translation and MetricsMATR, pages 315–320. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2012. Improving statistical machine translation for a resource-poor language using related resource-rich languages. *Journal of Artificial Intelligence Research*, 44(1):179–222.

Majid Razmara and Anoop Sarkar. 2013. Ensemble triangulation for statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 252–260.

Antti-Veikko I Rosti, Spyridon Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 312. Citeseer.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.

Omar Zaidan. 2009. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88.