Augmenting Pivot based SMT with word segmentation

Rohit More[†], **Anoop Kunchukuttan**[†], **Pushpak Bhattacharyya**[†]
[†] Department of Computer Science And Engineering
IIT Bombay, India
{rohit,anoop,pb}@cse.iitb.ac.in,

Raj Dabre[‡],
[‡] Graduate School of Informatics
Kyoto University
Japan
prajdabre@gmail.com

Abstract

This paper is an attempt to bridge two well known performance degraders in SMT, *viz.*, (i) difference in morphological characteristics of the two languages, and (ii) scarcity of parallel corpora. We address these two problems using "word segmentation" and through "pivots" on the morphologically complex language. Our case study is Malayalam to Hindi SMT. Malayalam belongs to the Dravidian family of languages and is heavily agglutinative. Hindi is a representative of the Indo-Aryan language family and is morphologically simpler. We use triangulation as pivoting strategy in combination with morphological pre-processing. We observe that (i) significant improvement in translation quality over direct SMT occurs when a pivot is used in combination with direct SMT, (ii) the more the number of pivots, the better the performance and (iii)word segmentation is a must. We achieved an *improvement* of 9.4 BLEU points which is over 58% compared to the baseline direct system. Our work paves way for SMT of languages that face resource scarcity and have widely divergent morphological characteristics.

1 Introduction

Hindi (hin) and Malayalam (mal) are two important languages from Indian sub-continent. Hindi is a language belonging to Indo-Aryan family with 300 million native speakers. Malayalam belongs to the Dravidian language family. It is spoken by over 38 million people.

The task of translation between Hindi and Malayalam proves to be a difficult one. This is due to scarcity of available parallel corpus and high agglutinative nature of Malayalam. Malayalam is a morphologically rich, agglutinative language in which complex words are formed by concatenating morphemes together. For example, "अगर बादल नहीं बरसे तो भी" (*if cloud not rain_verb then also*) in Hindi (5 words) would translate to "മഴാ പെയ്യുതിലെങ്കിലും" (*rain_noun rain_verb+not+even_if+then_also*) in Malayalam (2 words).

In this paper, we present a case of translation from Malayalam to Hindi. Our approach is based on combined use of pivot strategies for Statistical Machine Translation (SMT) and word segmentation techniques. We show that word segmentation of source language as well as pivot language helps to improve the translation quality. Section 2 contains details about relevant work done in the field. Section 3 explains the design of our system in detail. Section 4 describes the experimental setup. Results of the experiments are discussed in Section 5. Section 6 includes concluding remarks on the mal-hin translation task.

2 Related Work

There is substantial amount on pivot-based SMT. De Gispert and Marino (2006) discuss translation tasks between Catalan and English with the use of Spanish as a pivot language. Pivoting is done using two techniques: pipelining of source-pivot and pivot-target SMT systems and direct translation using a synthesized Catalan-English. In Utiyama and Isahara (2007), the authors propose the use of pivot language through - phrase translation (phrase table creation) and sentence translation. Wu and Wang (2007) compare three pivot strategies *viz.* - phrase translation (*i.e.* triangulation), transfer method and synthetic method. Nakov and Ng (2012) try to exploit the similarity between resource-poor languages and resource-rich languages for the translation task. Dabre et al. (2014) used multiple decoding paths (MDP) to overcome

the limitation of small sized corpora.Paul et al. (2013) discusses criteria to be considered for selection of good pivot language. Use of source-side segmentation as pre-processing technique has been demonstrated by (Kunchukuttan et al., 2014). Goldwater and McClosky (2005) investigates several methods for incorporating morphological information to achieve better translation from Czech to English.

Most of the pivot strategies mentioned above focus on the situation of resource-poor languages where direct translation is either very poor or not available. Our approach, like Dabre et al. (2014), tries to employ pivot strategy to help improve the performance of existing SMT systems. To the best of our knowledge, our work is the first attempt to integrate word segmentation with pivot-based SMT.

3 Our System

We propose a system which integrates word segmentation with triangulation and combines more than one SMT systems. The required concepts are explained as follows.

3.1 Pivoting by Triangulation

Wu and Wang (2007) discuss triangulation as a pivoting strategy. In this method, the source-pivot models and pivot-target models are trained using source(L_s)-pivot(L_p) and pivot(L_p)-target(L_t) corpora respectively. Using these two models, we induce a source-target model. The two important components to be calculated are - 1) phrase translation probability and 2) lexical weight.

The **Phrase translation probability** is estimated by marginalizing over all possible pivot phrase, along with the assumption that the target phrases are independent of the source phrase given the pivot phrase. The phrase translation probability can be calculated as shown below:

$$\phi\left(\vec{s}\|\vec{t}\right) = \sum_{\vec{p}} \phi\left(\vec{s}\|\vec{p}\right) \phi\left(\vec{p}\|\vec{t}\right) \quad (1)$$

Where, $\vec{s}$, $\vec{p}$, $\vec{t}$ are phrases in languages L_s, L_p, L_t respectively.

The **Lexical Weight**, according to Koehn et al. (2003), depends on - 1) word alignment information a in a phrase pair (s, t) and 2) lexical translation probability $w(s|t)$.

Lexical weight can be modeled using following equation,

$$p_w\left(\vec{f}\|\vec{e}, a\right) = \prod_{i=1}^{n} \frac{1}{\|j|\,(i,j) \in a\|} \sum_{\forall(i,j)\in a} w\left(f_i\|e_j\right) \quad (2)$$

Wu and Wang (2009) discuss in detail about alignments information and lexical translation probability.

3.2 Word segmentation

We use unsupervised word segmentation as pre-processing technique. For this purpose, Morfessor (Virpioja et al., 2013) is used. It performs morphological segmentation of words of a natural language, based solely on raw text data. Morfessor uses probabilistic machine learning methods to do the task of segmentation. The trained models for word segmentation of Indian languages are available to use[1].

3.3 Integrating word segmentation with Triangulation

In our system, we use both phrase table triangulation and word-segmentation. The words in the source and pivot language training corpora are segmented before training the SMT systems. The target language is left unchanged. The phrase tables are then triangulated. This process is shown in Figure 1.

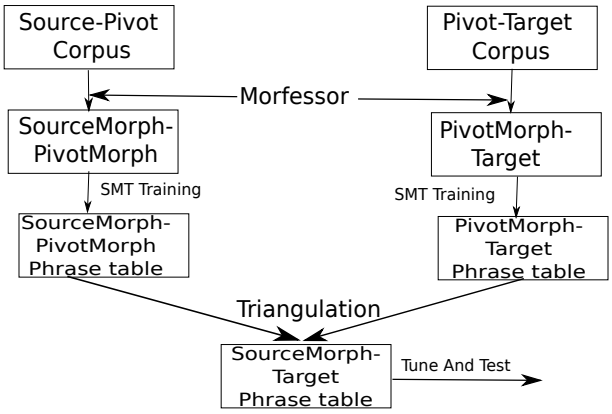

Figure 1: Integration of word segmentation with triangulation

3.4 Multiple Decoding Paths

We use the triangulated phrase table to supplement the direct phrase table. In order to integrate these two phrase tables, we use the multiple decoding paths (MDP) feature provided by the

[1] https://github.com/anoopkunchukuttan/indic_nlp_library

Moses decoder. Multiple decoding paths (Koehn and Hoang, 2007) allows us to lookup multiple translation models for hypothesis at decoding time, and the choice of best hypothesis at decoding time based on available evidence. We use MDP to combine one or more pivot-based MT systems with the direct MT system. This constitutes our **final decoding system**. We preferred this option over offline linear interpolation of phrase tables since the framework can dynamically consider phrases from multiple phrase tables and wouldn't need any hyperparameter tuning.

4 Experiments

The aim of experiments is to study impact of pivot strategies and word segmentation, separately and together.

4.1 Resource Details

We used the ILCI (Jha, 2010) multilingual corpora of around 50K sentences. The corpora belongs to Health and Tourism domain. Indian languages used in experiments are Bengali (ban), Gujarati (guj), Hindi (hin), Konkani (kok), Malayalam (mal), Marathi (mar), Panjabi (pan), Tamil (tam) and Telugu (tel).

Our data split was as follows: 46277 sentences are used for training, 500 sentences are used for tuning and 2000 sentences are used for testing.

For the experiments, we use phrase-based SMT training and 5-gram SRILM language model. Tuning is done using the MERT algorithm. The triangulated MT systems use default distance based reordering while direct systems use wbe-msd-bidirectional-fe-allff model

4.2 Experimental Setup

We trained various phrase based SMT systems by combining the basic systems mentioned in Section 3. We use a threshold of 0.001 for phrase translation probability to manage size of triangulated phrase table. The performance metric used is BLEU (Papineni et al., 2002). The following are the configurations we experimented with:

1. **DIR**: MT system trained on direct Source-Target corpus.
2. **DIR_Morph**: **DIR** system with source-text word-segmented.
3. **PIVOT**: MT system based on triangulated phrase table of Source-Target using a single Pivot language.
4. **PIVOT_Morph**: **PIVOT** system with both Source and Pivot texts segmented.
5. **PIVOT_SourceMorph**: **PIVOT** system with only Source text segmented.
6. **DIR+PIVOT**: MT system based on integration of **DIR** and **PIVOT** phrase tables using MDP.
7. **DIR_Morph+PIVOT_Morph**: MT system based on integration of **DIR_Morph** and **PIVOT_Morph** using MDP.
8. **DIR+All-PIVOT**: MT system based on integration of **DIR** and all 7 **PIVOT** systems using MDP.
9. **DIR_Morph+All-PIVOT_Morph**: MT system based on integration of **DIR_Morph** and all 7 **PIVOT_Morph** systems using MDP.

5 Results and Discussions

Table 1 shows BLEU scores for best performing MT systems for each experiment. Wherever a single pivot is used, it refers to the best supplementing pivot. In all the experiments, Punjabi was observed to be the best pivot.

Type of MT System	BLEU
DIR	16.11
DIR_Morph	23.35
PIVOT	15.72
PIVOT_Morph	23.02
PIVOT_SourceMorph	22.1
DIR+PIVOT	17.22
DIR_Morph+PIVOT_Morph	24.5
DIR+All-PIVOT	18.67
DIR_Morph+All-PIVOT_Morph	25.51

Table 1: BLEU Scores of MT systems

We can see that the pivot-only system gives translation accuracy comparable to the direct system for the best performing pivot language, Punjabi. Punjabi shares a large fraction of its vocabulary with Hindi. In addition, **pan** is morphologically simpler compared to other Indian languages. Hence, better word alignment can be achieved and during triangulation, more phrase pairs are obtained.

Though the **PIVOT** system does not better the performance of the **DIRECT** system, the combination of both the system *i.e.* **DIR+PIVOT** performs better than the **DIRECT** system.

DIR+PIVOT has relative improvement of 6.8% over **DIR** system. Addition of all **PIVOT** systems using multiple decoding paths shows that each pivot helps improve the BLEU score since each **PIVOT** system provides additional translation options. The all-pivot system *i.e.* **DIR+All-PIVOT** shows improvement of about 15% (2.56 BLEU points) over **DIR** system.

Pre-processing the corpus with word segmentation (**DIR_Morph** system) results in a substantial improvement of 44% over the **DIR** system. Use of word segmentation brings about a major improvement in the BLEU score. Since **mal** is morphologically rich, word segmentation reduces data sparsity, helps learn better alignment models and reduces the number of unknown words in test set.

We compared the use of word segmentation at source as well as pivot with word segmentation on the source language alone. Figure 2 compares these two system for various pivot languages.

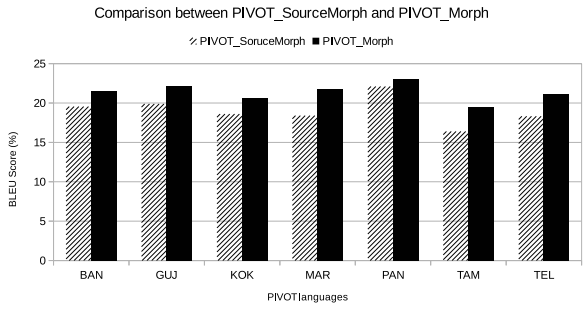

Figure 2: Comparison of PIVOT_Morph and PIVOT_SourceMorph MT systems

The additional use of word segmentation at pivot provides a 4% to 18% increase in the BLEU score.

The **DIR_Morph+All-PIVOT_Morph** system combines all pivot languages as well as source and pivot segmentation. This is the best performing system which achieves an improvement of 9.4 BLEU (58% improvement) over **DIR** system.

5.1 Examples

Below are few examples comparing baseline **DIR** and our final system *i.e.* **DIR_Morph+All-PIVOT_Morph**.

- **Example 1:**
  mal input: താമസിക്കുന്നതിനുള്ള സൌകര്യം സർക്കാർ വിശ്രമ കൃാംപുകൾ കാർലാ ഹോട്ടൽ എന്നിവയിലുണ്ട്.
  tAmasikkunnatinuLLa saukarya.n sarkkAr vishrama kyA.npukaL kArlA hoTTal .ennivayiluNT)
  Baseline: के ठहरने की व्यवस्था सरकारी विश्राम कैंप कारला होटल എന്നിവയിലുണ്ട് के नाम से जाना जाता है ।
  (ke Thaharane kI vyavasthA sarakArI vishrAma kai.npa kAralA hoTala .ennivayiluNT ke nAma se jAnA jAtA hai)
  Final system: ठहरने की व्यवस्था सरकारी विश्राम कैंप कारला होटल भी उपलब्ध हैं ।
  (Thaharane kI vyavasthA sarakArI vishrAma kai.npa kAralA hoTala bhI upalabdha hai.n)
  Better translations for words are generated, and the number of unknown words is reduced.
- **Example 2:**
  mal input: ചെവിയിലെ കുരുക്കൾ ആരെയും ഉറങ്ങാൻ അനുവദിക്കില്ല.
  (c.eviyil.e kurukkaL Ar.eyu.n uRa N NAn anuvadikkilla)
  Baseline: कान की फुंसी को सोने के ।
  (kAna kI phu.nsI ko sone ke)
  Final system: कान की फुंसी किसी को भी नींद नहीं होने दिया ।
  (kAna kI phu.nsI kisI ko bhI nI.nda nahI.n hone diyA)
  The direct system is not able to translate some words, whereas out final system translates most words correctly.

6 Conclusions

We have shown that pre-processing using word segmentation and augmentation of direct systems with pivot-based systems provides significant advancement in translation quality. This is an attempt to integrate word segmentation with pivot-strategies. We achieved compelling results with our approach. The approach is applicable to any resource-scarce language pair with morphologically rich source side. In future, we will focus on applying our approach to other challenging language pairs. We will also work on MT tasks which involve morphologically rich target language.

Acknowledgements

We would like to thank the Technology Development for Indian Languages (TDIL) Programme and the Department of Electronics & Information Technology, Govt. of India for providing the ILCI corpus.

References

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2014. Leveraging small multilingual corpora for smt using many pivot languages.

Adrià De Gispert and Jose B Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68. Citeseer.

Sharon Goldwater and David McClosky. 2005. Improving statistical mt through morphological analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 676–683. Association for Computational Linguistics.

Girish Nath Jha. 2010. The tdil program and the indian language corpora initiative (ilci). *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA)*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Anoop Kunchukuttan, Ratish Pudupully, Rajen Chatterjee, Abhijit Mishra, and Pushpak Bhattacharyya. 2014. The iit bombay smt system for icon 2014 tools contest. In *NLP Tools Contest at ICON 2014*.

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.

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*, pages 311–318. 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. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):14.

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

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

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

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*, pages 154–162. Association for Computational Linguistics.