That'll do *Fine*!: A *Coarse* Lexical Resource for English-Hindi MT, using Polylingual Topic Models

Diptesh Kanojia[2], **Aditya Joshi**[1,2,3], **Pushpak Bhattacharyya**[2], **Mark James Carman**[3]
[1] IITB-Monash Research Academy, Mumbai, India,
[2] IIT Bombay, Mumbai, India,
[3] Monash University, Melbourne, Australia
diptesh@cse.iitb.ac.in, adityaj@cse.iitb.ac.in, pb@cse.iitb.ac.in, mark.carman@monash.edu

Abstract

Parallel corpora are often injected with bilingual lexical resources for improved Indian language machine translation (MT). In absence of such lexical resources, multilingual topic models have been used to create coarse lexical resources in the past, using a Cartesian product approach. Our results show that for morphologically rich languages like Hindi, the Cartesian product approach is detrimental for MT. We then present a novel 'sentential' approach to use this coarse lexical resource from a multilingual topic model. Our coarse lexical resource when injected with a parallel corpus outperforms a system trained using parallel corpus and a good quality lexical resource. As demonstrated by the quality of our coarse lexical resource and its benefit to MT, we believe that our sentential approach to create such a resource will help MT for resource-constrained languages.

Keywords: Topic models, Machine Translation, Statistical Machine Translation, Coarse Dictionary

1. Introduction

Parallel corpora are often injected with bilingual lexical resources for improved machine translation (MT). These lexical resources are often manually created dictionaries. Creation of such a resource is a time and effort intensive task. This paper presents a novel approach to use a coarse lexical resource using parallel topics obtained from multilingual topic models. We observe that for a machine translation system for English-Hindi, these *coarse* alignments do *fine*!

In a country like India where more than 22 official languages are spoken across 29 states, the task of translation becomes immensely important. A statistical machine translation (SMT) system typically uses two modules: alignment and reordering. The quality of an SMT system is dependent on the alignments discovered. The initial quality of word alignment is known to impact the quality of SMT (Och and Ney, 2003; Ganchev et al., 2008). Many SMT based systems are evaluated in terms of the information gained from the word alignment results. However, there is not a lot of parallel data available for these languages making it necessary for specialized techniques that improve alignment quality has been felt (Sanchis and Sánchez, 2008; Lee et al., 2006; Koehn et al., 2007).

The existing baseline approach is called **Cartesian product Approach**. This approach was used by Mimno et al. (2009). In their work, they analyzed the characteristics of MLTM in comparison to monolingual LDA, and demonstrated that it is possible to discover aligned topics. They also demonstrated that relatively small numbers of topically comparable document tuples are sufficient to align topics between languages in non-comparable corpora. They then use MLTM to create bilingual lexicons for low resource language pairs, and provided candidate translations for more computationally intense alignment processes without the sentence-aligned translations. They conduct experiments for Spanish, English, German, French, and Italian. Figure 2 summarizes the approach. For parallel topic i with top 3 words, we add 9 pseudo-parallel sentences with one-to-one word alignment, as shown. Thus for T topics, and K top words, Cartesian product approach results in pseudo-parallel data of $T * K$ sentences of length 1 each. This is appended to the parallel corpus.

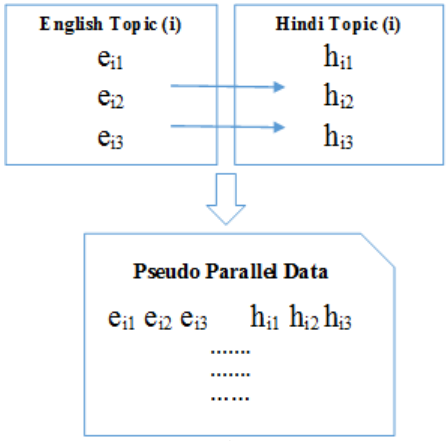

Figure 1: Our Sentential Approach to create pseudo-parallel data

The Cartesian product approach adds the word sets for every topic to a set of candidate translations. While it provides with a lot more pseudo parallel data to be injected, it also injects one to one aligned non synonymous words to the parallel data. On the other hand, sentential approach only provides fewer pairs. Thus, intuitively, sentential approach performs better in this regard, while injecting less noisy data to the parallel corpus.

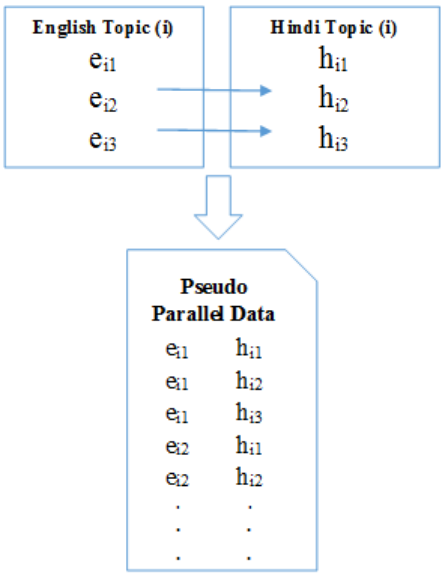

Figure 2: Existing Cartesian Product Approach to generate pseudo-parallel data

2. Related Work

Our work covers two broad areas of research: Multilingual topic models and improvement of alignment in MT. We now describe the two in this section.

We implement the algorithm by Mimno et al. (2009) called PolyLDA. This model discovers topics for English - Hindi Parallel text, and use it to create pseudo-parallel data. They proposed Cartesian approach to inject the pseudo parallel data in the training corpora. They evaluate their topics for machine translation. Such multilingual topic models have been applied to a variety of tasks. Ni et al. (2009) extract topics from wikipedia, and use the top terms for a text classification task. They observe that parallel topics perform better than topic words that are translated into the target language. Approaches that do not rely on parallel corpus have also been reported. Jagarlamudi and Daumé III (2010) use a bilingual lexical resource, and a comparable corpora to estimate a model called JointLDA. Boyd-Graber and Blei (2009) use unaligned corpus and extract multilingual topics using a multilingual topic model called MuTo.

The second area that our work is related to is improvement of alignment between words/phrases for machine translation. Och and Ney (2000) describe improved alignment models for statistical machine translation. They use both the phrase based and word based approaches to extend the baseline alignment models. Their results show that this method improved precision without loss of recall in English to German alignments. However, if the same unit is aligned to two different target units, this method is unlikely to make a selection. Cherry and Lin (2003) model the alignments directly given the sentence pairs whereas some researchers use similarity and association measures to build alignment links (Ahrenberg et al., 1998; Tufiş and Barbu, 2002). In addition, Wu (1997) use a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments. Some researchers use preprocessing steps to identity multi-word units for word alignment (Ahrenberg et al., 1998; Tiedemann, 1999; Melamed, 2000). These methods obtain multi-word candidates, but are unable to handle separated phrases and multiwords in low frequencies. Hua and Haifeng (2004) use a rule based translation system to improve the results of statistical machine translation. It can translate multiword alignments with higher accuracy, and can perform word sense disambiguation and select appropriate translations while a translation lexical resource can only list all translations for each word or phrase. Some researchers use Part-of-speeches (POS), which represent morphological classes of words, tagging on bilingual training data (Sanchis and Sánchez, 2008; Lee et al., 2006) give valuable information about words and their neighbors, thus identifying a class to which the word may belong. This helps in disambiguation and thus selecting word correspondences but can also give rise to increased vocabulary thus making the training data more sparse. Finally, Koehn et al. (2007) propose a factored translation model that can incorporate any linguistic factors including POS information in phrase-based SMT.

It provides a generalized representation of a translation model, because it can map multiple source and target factors. It may help to effectively handle out-of-vocabulary (OOV) by incorporating many linguistic factors, but it still crucially relies on the initial quality of word alignment that will dominate the translation probabilities. In this way, our paper attempts to verify the claim that multilingual topics can be used to address the problem of improved alignment generation. We use a baseline that contains no bilingual lexical resource, and an approach that contains a good quality bilingual lexical resource. This is similar to the approach in Och and Ney (2000).

3. Our Coarse Lexical Resource

Our coarse lexical resource is a set of pseudo-parallel sentences where each word on the source side has at least one target alignment on the other side. This resource is created from the parallel topics as obtained from the polylingual topic model, using a sentential approach. This is a novel approach which concatenates words belonging to the same topic as a pseudo-sentence. The approach is shown in Figure 1. We use the words aligned in topic models and put them in a sentence to create parallel sentences for the training corpora to be used in creating the MT system. Thus, the pseudo-parallel data generated in this case consists of 1 sentence per topic: $e_{i1}e_{i2}e_{i3}$ in parallel with $h_{i1}h_{i2}h_{i3}$. We use the sentential approach for the English - Hindi where the sentences constructed may not be word aligned but, unlike so many one to one Cartesian product alignments, our approach keeps them in the same sentence, thus reducing the chances of the system learning non synonymous candidate translations.

TOPIC 1		TOPIC 2		TOPIC 3		TOPIC 4	
vitamin	मात्रा	clean	साफ	disease	रोग	cancer	कैंसर
quantity	विटामिन	acid	पथरी	blood	रक्त	nose	नाक
amount	महीने	ulcer	अल्सर	heart	हृदय	breast	शिकायत
large	बड़ी	stones	एसिड	diabetes	बीमारी	complaint	गर्भाशय
months	नाम	asthma	पड़ती	increases	बढ़	uterus	भूख

Figure 3: Parallel English-Hindi topics as generated by the topic model for the health dataset

	Hindi (%)	English (%)	Kappa
A1	69.6	70.4	0.838
A2	65.6	68.4	

Table 1: Quantitative Evaluation of our resource; Hindi (%) for A1 indicates the proportion of Hindi words that had a corresponding English translation, according to annotator A1

Thus, for T topics, and K top words, sentential approach results in a coarse lexical resource of T X K pseudo-parallel sentences. The coarse lexical resource for varying values of T is available freely for download.

3.1. Experiment Setup

To generate the topics, we use corpora from health and tourism domain by Khapra et al. (2010). These datasets contain approximately 25000 parallel sentences for English - Hindi language pair. We implement the multilingual topic model in Java. Our implementation uses Gibbs sampling as described in the original paper. Table **??** shows examples of the two domains, and the corresponding pseudo-parallel data additions.

3.2. Qualitative Evaluation

Figure 3 shows top 5 words for sample parallel English-Hindi topics for the health dataset. The total number of topics, as stated before, is 50. Figure 3 shows four topics which correspond to four thematic components of the health dataset. Topic 1 is about administration of medicines, Topic 2 and 3 are about two kinds of diseases, while Topic 4 is about different types of cancer. We also see that translations of the English words appear in the corresponding Hindi side for each of the topics. They may not appear in the same order, since these are dependent on the frequency of the word in the specific language. Thus, our model is able to discover **coarse topics** underlying the datasets. Similar trends are observed in case of the tourism dataset. Thus, our model is able to discover **parallel synonyms** across the two languages. Among 40 English words present in these figures, only 7 do not have a translation in the corresponding Hindi topic[1]. Similarly, for the 40 Hindi words, 6 do not have a translation in the corresponding English topic.

[1]These are 'nearest, centre, asthma, diabetes, place, kms, breast'

3.3. Quantitative Evaluation

Two human annotators evaluated the quality of the output obtained. Each word was marked as whether or not a translation in the other language was present in the same topic. The two annotators, A1 and A2, are native speakers of Hindi, and have had 15+ years of academic instruction in English. The inter-annotator agreement between them and their corresponding judgments are shown in Figure 1.

4. Application of our resource to improve Machine Translation

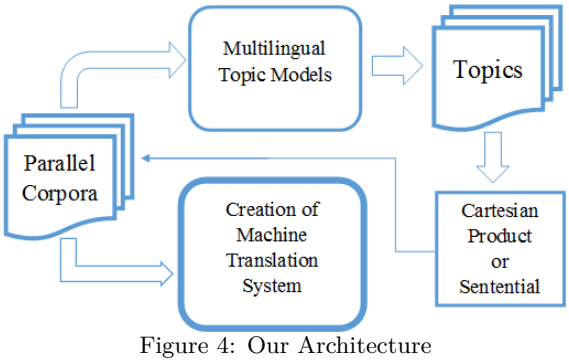

Figure 4: Our Architecture

The basic architecture for creation and use of our coarse lexical resource is shown in Figure 4. We create the coarse lexical resource as described above. Using this resource, we generate '*pseudo-parallel*' data - parallel words or groups of words that may be translations of each other. Finally, this data is appended to the parallel corpus used for training a Moses-based MT system (Koehn et al., 2007). We create pseudo-parallel data of T sentences of length K each, and is injected to the parallel corpus. For the parallel corpus, we use the same datasets as above. We separate 500 sentences each for testing and tuning purposes

The key step in the architecture is the approach used to create pseudo-parallel data from the lexical resource. The existing approach given by (Mimno et al., 2009) is called '**Cartesian approach**'. Our approach is called the '**sentential approach**' (as opposed to the original Cartesian product approach).

4.1. Setup

We used four configurations for our experiment. They vary in terms of the data which was injected into the

	Health	Tourism
No lexical resource (Baseline)	26.14	**28.68**
Cartesian product Approach (50 topics)	25.98	28.44
Sentential Approach (50 topics)	**26.25**	27.52
Full lexical resource	**26.31**	**29.30**

Table 2: MT Results using no lexical resource (baseline), good quality lexical resource and coarse lexical resource obtained through multilingual topic model (Cartesian product and Sentential approach)

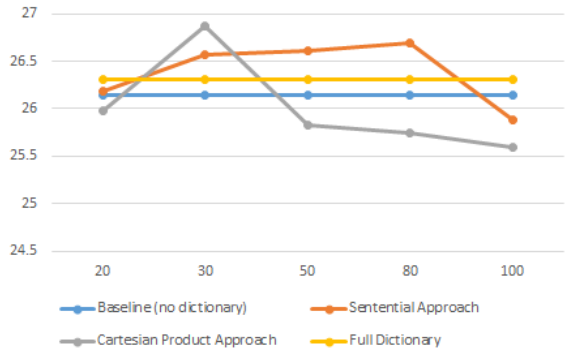

Figure 5: Change in BLEU scores for different value for Topics (T) for health domain

training data. We use MOSES Toolkit for all our experiments. We set K, the number of words in a topic model, to be 5.

1. **No lexical resource (Baseline)**: A basic setup for creating an MT system requires training, testing and tuning corpora which we obtained for HEALTH and TOURISM domains.
2. **Cartesian Product Approach**: In this approach the pseudo parallel data was created using MLTM approach described in the earlier work. Thus, for 50 topics and 5 top words, we add 250 pseudo-parallel sentences, each of length 1.
3. **Sentential Approach**: We added the pseudo parallel data created using MLTM approach to the training data using the approach indicated in Figure 1. Thus, for 50 topics and 5 top words, we add 50 pseudo-parallel sentences, each of length 5.
4. **Full lexical resource**: While the baseline uses no lexical resource, this approach considered uses a good quality bilingual lexical resource from `http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php`. The lexical resource consists of more than 100,000 mappings between English and Hindi words.

4.2. Results

This section evaluates our implementation of the multilingual topic model for its impact on machine translation. We first present sample topics that are generated by the model. In the next subsection, we discuss the impact on machine translation.

We now compare the baseline against the Cartesian product and sentential approaches that use multilingual topics. The total number of topics, as stated before, is 50. Table 2 shows the BLEU scores before and after injecting the multilingual topic modeling data, for the two datasets. We observe that BLEU score obtained on multilingual topic modeled data set using the Cartesian product approach for HEALTH domain is 25.98.

4.3. Impact of Number of Topics on MT

The degradation above is a parameter of number of topics; to ascertain that there is indeed degradation, we vary the number of topics. Hence, we conduct a separate run of our topic model for number of topics 20, 30, 50, 80 and 100. We then use different approaches as shown above, and show the results for health domain in the graph above (Figure 5). The x - axis represent the number of topics (T) varying from 20 to 100. The results of two topic modeled approaches namely Cartesian product approach and Sentence formation approach are shown above.

The baseline MT output is shown as a horizontal line as no topic model data is being added to it. The line representing the Cartesian product approach clearly shows the degradation of MT output for English - Hindi. On the other hand, the sentential approach shown minor improvements for a varied number of topic models. As more topics are added, sentential approach improves over the baseline. However, beyond 100, we observe a substantial degradation. This is because data sparsity along with too many topics introduces non-synonymous words in parallel topics.

For topics 30, 50 and 80, our approach of using a coarse lexical resource obtained through multilingual topics surpasses using a full, good quality lexical resource.

In summary, we see that existing Cartesian product approach using multilingual topics (devised for European languages) is detrimental for Indian language MT. A modified sentential approach results in marginal improvement.

5. Conclusion & Future Work

We discussed two approaches to generate a coarse lexical resource derived from topics obtained as a result of a multilingual topic model. We used the Cartesian product approach that adds all combinations of top words in a topic. On the other hand, we introduced the sentential approach which adds all words together as a single sentence. Our approach using a coarse lexical resource performs better than using a good quality lexical resource, over a range of topic-count values.

Our approach to create a coarse lexical resource paves the way for similar adaptations in resource-constrained languages, where a dictionary may not be present but some parallel data may be available.

6. Bibliographical References

Ahrenberg, L., Andersson, M., and Merkel, M. (1998). A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 1*, COLING '98, pages 29–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Boyd-Graber, J. and Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 75–82. AUAI Press.

Cherry, C. and Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 88–95, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ganchev, K., Graça, J. V., and Taskar, B. (2008). Better alignments = better translations. In *in Proc. of the ACL*.

Hua, W. and Haifeng, W. (2004). Improving statistical word alignment with a rule-based machine translation system. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jagarlamudi, J. and Daumé III, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval*, pages 444–456. Springer.

Khapra, M. M., Kulkarni, A., Sohoney, S., and Bhattacharyya, P. (2010). All words domain adapted wsd: Finding a middle ground between supervision and unsupervision. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1532–1541. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lee, J., Lee, D., and Lee, G. G. (2006). Interspeech 2006 improving phrase-based korean-english statistical machine translation.

Melamed, I. D. (2000). Models of translational equivalence among words. *Comput. Linguist.*, 26(2):221–249, June.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

Ni, X., Sun, J.-T., Hu, J., and Chen, Z. (2009). Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156. ACM.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sanchis, G. and Sánchez, J. A. (2008). Vocabulary extension via pos information for smt.

Tiedemann, J. (1999). Word alignment – step by step. In *Proceedings of the 12th Nordic Conf. on Computational Linguistics*, pages 216–227.

Tufiş, D. and Barbu, A. (2002). Lexical token alignment: Experiments, results and application. In *In Proc. of LREC-2002*, pages 458–465.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403, September.