Using Multilingual Topic Models for Improved Alignment in English-Hindi MT

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

Abstract

Parallel corpora are often injected with bilingual dictionaries for improved Indian language machine translation (MT). In absence of such dictionaries, a coarse dictionary may be required. This paper demonstrates the use of a multilingual topic model for creating coarse dictionaries for English-Hindi MT. We compare our approaches with: (a) a baseline with no additional dictionary injection, and (b) a corpus with a good quality dictionary. Our results show that the existing Cartesian product approach which is used to create the pseudo-parallel data results in a degradation on tourism and health datasets, for English-Hindi MT. Our paper points to the fact that existing Cartesian approach using multilingual topics (devised for European languages) may be detrimental for Indian language MT.

On the other hand, we present an alternate 'sentential' approach that leads to a slight improvement. However, our sentential approach (using a parallel corpus injected with a coarse dictionary) outperforms a system trained using parallel corpus and a good quality dictionary.

1 Introduction

Word Alignment is defined as the process of mapping synonymous words, using a parallel corpora. It is considered to be a valuable resource, and can be used for various applications such as word sense disambiguation, statistical machine translation (SMT), and automatic construction of bilingual text.

Statistical Machine Translation (SMT) is a technology for the automatic translation of text in one natural language into another. In a country like India where more than 22 official languages are spoken across 29 states, the task of translation becomes immensely important. A 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.

IBM models (Brown et al., 1993) are among the most widely used models for statistical word alignment. For these models, having a large parallel dataset can result in good alignment, and hence, facilitate a good quality SMT system. However, there is not a lot of parallel data available for English to Indian Languages, or for one Indian Language to another. Without sufficient amount of parallel corpus, it is very difficult to learn the correct correspondences between words that infrequently occur in the training data. Hence, a need for specialized techniques that improve alignment quality has been felt (Sanchis and Snchez, 2008; Lee et al., 2006; Koehn et al., 2007).

Mimno et al. (2009) present a multilingual topic model called PolyLDA, and apply it for Machine Translation for European and other languages such as Danish, German, Greek, English, Spanish, etc. Since multilingual topic models generate parallel topics: parallel clusters of words that are likely to be about the same theme, these topics provide coarse alignment that a Moses-like translation system can leverage on. The idea is to not rely on any external ontology such as WordNet and to rely purely on a parallel corpus to create such coarse alignments. The focus of our paper is to improve word alignments using multilingual Topic Models approach for English-Hindi MT. The key question that this paper attempts to answer is:

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 308–315,
Trivandrum, India. December 2015.

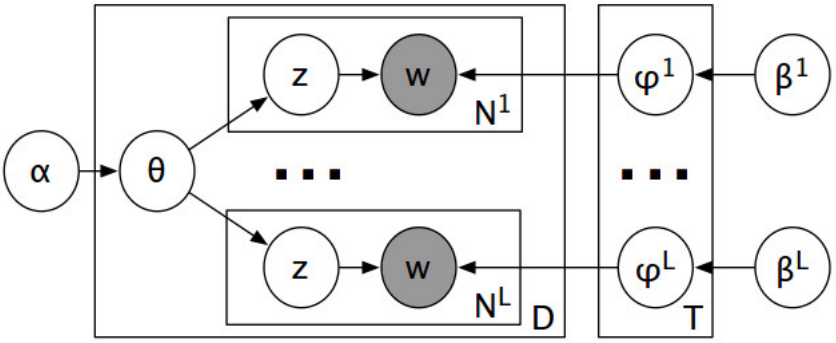

Figure 1: PolyLDA: Plate Diagram

'*Can the information gained from multilingual topic models help in improving the quality of SMT for English - Hindi Machine Translation?*'

The novelty of the paper lies in the following ways:

- Implementation of a multilingual topic model by Mimno et al. (2009) and applying it to English-Hindi MT using the Cartesian product approach.
- A novel approach to employ multilingual topics extracted by the multilingual topic model above. The approach called the sentential approach that performs better than the Cartesian product approach.

The paper is organized as follows. In section 2, we describe our related work. In section 3, we introduce multilingual Topic Models. In section 4 and 5, we describe the architecture of our work, and the experiment setup. We show the results obtained and our error analysis in section 6, and conclude in section 7.

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 dictionary, 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; Tufi 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 mul-

tiword alignments with higher accuracy, and can perform word sense disambiguation and select appropriate translations while a translation dictionary 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 Snchez, 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. Joshi et al. (2013) use in domain parallel data to inject additional alignment mappings for the news headline domain. 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 dictionary, and an approach that contains a good quality bilingual dictionary. This is similar to the approach in Och and Ney (2000).

3 Multilingual Topic Models: An Introduction

A topic model takes as input a set of documents, and generates clusters of words called 'topics'. These topics help to understand themes underlying a dataset. A popular topic model by Blei et al. (2003) called Latent Dirichlet Allocation (LDA) is an unsupervised model. This model takes as input the value of T as the number of topics, and a set of Dirichlet hyperparameters. LDA takes as input a dataset and models two kinds of distributions: (a) a document-topic distribution that determines the distribution of topics within a document, and (b) a word-topic distribution that determines the distribution of words across topics. The two distributions can be estimated using an Expectation-Maximization approach, or Gibbs sampling, among other approaches.

Multilingual LDA (PolyLDA) introduced by Mimno et al. (2009) is based on the idea of LDA, and extends it for extracting equivalent topics across languages. This topic model takes as input parallel documents (or sentences) across two or more languages, and discovers T topics in each of the languages. For example, a PolyLDA may discover an English topic '*book, books*' and the corresponding topic '*libre, libres*' in French, and '*libro, libri*' in Italian. Figure 1 shows the plate diagram for our implementation of PolyLDA. We use a simplified version of PolyLDA for two languages: *i.e.*, in our case $L = 2$.

Each document is modeled as a set of N^l words w where l ranges from 1 to L for L parallel languages. Each document has a distribution θ over all topics, over a total of D documents. Since L is 2 for us, There are two word-topic distributions ϕ_1 and ϕ_2 for English and Hindi, respectively. Thus, based on which language the word belongs to, it is picked from the appropriate distribution. α and β are the Dirichlet hyperparameters.

The output of PolyLDA is the estimation for ϕ for the two languages. The top n words for each topic are considered as candidates for generation of pseudo-parallel data.

4 Architecture

Indian languages are morphologically complex, and sometimes very agglutinative. English to Hindi poses a separate challenge of word ordering as well. We take help of multilingual topic models (MLTM) to obtain coarse alignments. In this section, we describe our architecture.

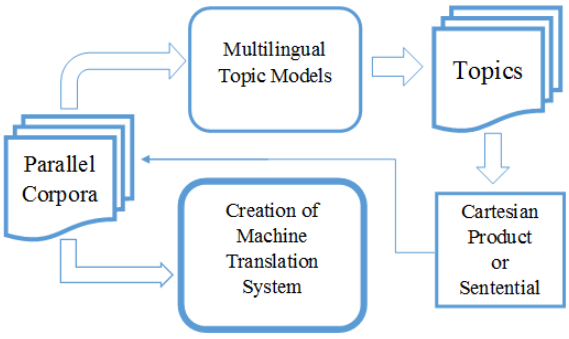

Figure 2: Our Architecture

The basic architecture is shown in Figure 2. We use multilingual topic models with the parameter of number of topics Z set. We thus obtain Z topics: each consisting of top k English and

top k Hindi words. Using these topics, 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).

The key step in the architecture is the approach used to create pseudo-parallel data. We consider two approaches to do so:

1. **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 3 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.
2. **Sentential Approach**: It is a novel approach which concatenates words belonging to the same topic as a pseudo-sentence. The approach is shown in Figure 4. 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.

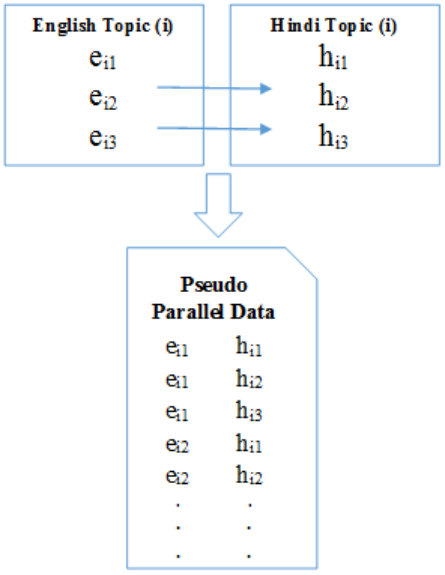

Figure 3: Existing Cartesian Product Approach to generate pseudo-parallel data

Thus for T topics, and K top words, sentential approach results in pseudo-parallel data of T sentences of length K each. This is appended to the parallel corpus.

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.

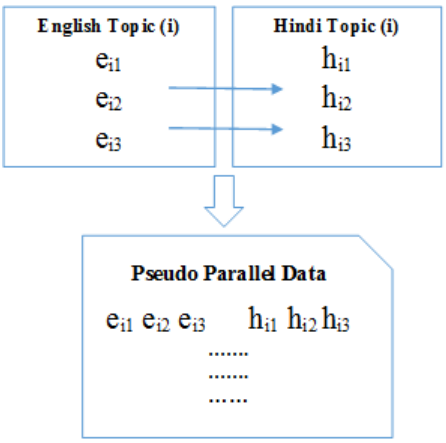

Figure 4: Our Sentential Approach to create pseudo-parallel data

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

TOPIC 1		TOPIC 2		TOPIC 3		TOPIC 4	
lake	झील	famous	प्रसिद्ध	worth	नाम	road	मार्ग
station	स्टेशन	country	देश	place	देखने	india	भारत
railway	रेलवे	state	प्रमुख	named	दर्शनीय	route	दिल्ली
nearest	धीरे	main	रूप	temples	नगर	path	सड़क
munnar	मुन्नार	centre	राज्य	city	मंदिरों	kms	रोड

Figure 6: Parallel English-Hindi topics as generated by the topic model for the tourism dataset

5 Experiment Setup

In this section, we describe the dataset, and setup for the experiments conducted.

5.1 Dataset

For our experiments, 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 use these for both the creation of pseudo parallel data, and training Machine translation systems. We separate 500 sentences each for testing and tuning purposes. We ensure that they are not present in the training corpus.

5.2 Setup

We implemented the multilingual topic model in Java. Our implementation uses Gibbs sampling as described in the original paper. We used two configurations for our experiment. They vary in the pseudo parallel data which was included along with the training data used for MT system. We use MOSES Toolkit for all our experiments. We perform experiments and obtained results for 3 different data sets as indicated in Figure 7. We set K, the number of words in a topic model, to be 11 for our experiments. For the initial experiments, we use number of topics as 50.

1. **No dictionary (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 earlier, and added to the training data before training the MT systems. We added the pseudo parallel data to training data using the approach indicated in Figure 3. Thus, for 50 topics and 11 top words, we add 550 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 4. Thus, for 50 topics and 11 top words, we add 50 pseudo-parallel sentences, each of length 11.

4. **Full dictionary**: While the baseline uses no dictionary, this approach considered uses a good quality bilingual dictionary from `http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php`. The dictionary consists of more than 100,000 mappings between English and Hindi words.

6 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.

6.1 Quantitative evaluation of Multilingual topics

Figures 5 and 6 show top 5 words for sample parallel English-Hindi topics for the health and tourism datasets respectively. The total number of topics, as stated before, is 50. Figure 5 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 6. Consider topic 2. The word temples on English side aligns with two words on the Hindi side: one in the root form, and one in the inflected form. Thus, our model is able to discover **inflected forms** of words. Consider topic 4. The synonyms 'road', 'route' and 'path' occur on the English side. Three synonyms 'maarg', 'sadak' and 'road' also appear on the Hindi side. 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.

6.2 Benefit to MT

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 1 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.

In the Cartesian product approach, we take parallel topics and map each topic word to every other parallel language topic word, and add them to training data. This may result in several incorrect translations being added to the parallel corpus.

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

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

Table 1: MT Results using no dictionary (baseline), good quality dictionary and coarse dictionary obtained through multilingual topic model (Cartesian product and Sentential approach)

This seems to be the likely cause for degradation from 26.14 to 25.98 in case of health dataset, and from 28.68 to 28.44 in case of tourism dataset. In case of health dataset, our sentential approach allows us to increase from baseline and move closer to using full dictionary.

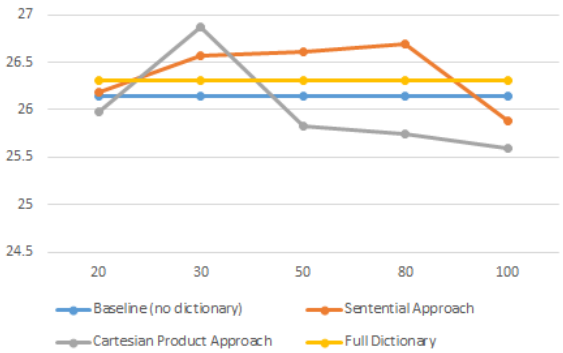

Figure 7: Change in BLEU scores for different value for Topics (T) for health domain

6.3 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 7). 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 sen-

tential 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 dictionary obtained through multilingual topics surpasses using a full, good quality dictionary.

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.

7 Conclusion and Future Work

We implemented a multilingual topic model based on a past work to extract parallel English-Hindi topics. These parallel topics can be used by Moses to improve the alignment quality. We discussed two approaches to generate pseudo-parallel data. 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. We compared against a baseline with no bilingual dictionary, and an approach that uses a good quality dictionary in order to understand how beneficial it is to add coarse mappings obtained from the topic models. The strength of our model lies in that our approach using a coarse dictionary performs better than using a good quality dictionary.

On experimentation with tourism and health datasets, we observed that the existing Cartesian product approach leads to a degradation in the BLEU score from 26.14 to 25.98. On the other hand, our sentential approach is able to restrict this degradation, but **results in a improvement for a range of number of topics**.

Our paper points to the fact that existing approaches using multilingual topics (devised for European languages) may be detrimental for Indian language MT. In addition, using additional observed variables for inflections and morphological characteristics, and latent variables for semantic classes, new multilingual topic models can be designed.

References

Lars Ahrenberg, Mikael Andersson, and Magnus Merkel. 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.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

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

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311, June.

Colin Cherry and Dekang Lin. 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.

Kuzman Ganchev, Joo V. Graa, and Ben Taskar. 2008. Better alignments = better translations. In *in Proc. of the ACL*.

Wu Hua and Wang Haifeng. 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.

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

Aditya Joshi, Kashyap Popat, Shubham Gautam, and Pushpak Bhattacharyya. 2013. Making headlines in hindi: Automatic english to hindi news headline translation. *Sixth International Joint Conference on Natural Language Processing*, page 21.

Mitesh M Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra

Constantin, and Evan Herbst. 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.

Jonghoon Lee, Donghyeon Lee, and Gary Geunbae Lee. 2006. Interspeech 2006 improving phrase-based korean-english statistical machine translation.

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

David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 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.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156. ACM.

Franz Josef Och and Hermann Ney. 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.

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

Germn Sanchis and Joan Andreu Snchez. 2008. Vocabulary extension via pos information for smt.

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

Dan Tufi and Anamaria Barbu. 2002. Lexical token alignment: Experiments, results and application. In *In Proc. of LREC-2002*, pages 458–465.

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