The IIT Bombay English-Hindi Parallel Corpus

Anoop Kunchukuttan[1], **Pratik Mehta**[2], **Pushpak Bhattacharyya**[1]
[1]Center for Indian Language Technology,
Department of Computer Science and Technology,
Indian Institute of Technology Bombay.
[2] College of Information and Computer Sciences,
University of Massachusetts Amherst.
{anoopk,pb}@cse.iitb.ac.in, psmehta@cs.umass.edu

Abstract

We present the IIT Bombay English-Hindi Parallel Corpus. The corpus is a compilation of parallel corpora previously available in the public domain as well as new parallel corpora we collected. The corpus contains 1.49 million parallel segments, of which 694k segments were not previously available in the public domain. The corpus has been pre-processed for machine translation, and we report baseline phrase-based SMT and NMT translation results on this corpus. This corpus has been used in two editions of shared tasks at the Workshop on Asian Language Transation (2016 and 2017). The corpus is freely available for non-commercial research. To the best of our knowledge, this is the largest publicly available English-Hindi parallel corpus.

Keywords: machine translation, parallel corpus, Indian languages

1. Introduction

Hindi is one of the major languages of the world spoken primarily in the Indian subcontinent, and is a recognised regional language in Mauritius, Trinidad and Tobago, Guyana, and Suriname. In addition, it serves as a major *lingua franca* in India. According to the 2001 Census of India, Hindi has 422 million native speakers and more than 500 millions total speakers (Wikipedia, 2017). It is also an official language of the Union Government of India as well as major Indian states like Uttar Pradesh, Bihar, Rajasthan, *etc.* and is used for conducting business and administrative tasks. Many languages and dialects in the Gangetic plains are closely related to Hindi *e.g.* Bhojpuri, Awadhi, Maithili, *etc.* Hindi is the fourth-most spoken language in the world, and third-most spoken language along with Urdu (both are registers of the Hindustani language). In contrast, English is spoken by just around 125 million people in India, of which a very small fraction are native speakers.

Hence, there is a large requirement for digital communication in Hindi and interfacing with the rest of the word via English. Hence, there is immense potential for English-Hindi machine translation. However, the parallel corpus available in the public domain is quite limited. This work is an effort to consolidate all publicly available parallel corpora for English-Hindi as well as significantly add to the available parallel corpus through corpora collected in the course of this work.

2. Dataset

The parallel corpus has been compiled from a variety of existing sources (primarily OPUS (Tiedemann, 2012), HindEn (Bojar et al., 2014b) and TED (Abdelali et al., 2014)) as well as corpora developed at the *Center for Indian Language Technology*[1], IIT Bombay over the years. The training corpus consists of sentences, phrases as well as dictionary entries, spanning many applications and domains. The details of the training corpus are shown in Table 1. The subcorpora (in the corpus distribution that we make available) are in the same order as listed in the table, so they can be separately extracted, if required.

2.1. Corpus Details

We briefly describe the new sub-corpora we have added to the collection. For the corpora compiled from existing sources, please refer to the papers mentioned in Table 1.

Judicial domain corpus - I contains translations of legal judgements by in-house translators with many years, though not with a legal background.

Judicial domain corpus - II contains translation done by graduate students taking a graduate course on natural language processing as part of a course project. This was part of an exercise of collecting translations in complex domain by non-expert translators. The translations included in the corpus were determined to be of good quality by annotators.

Mahashabdkosh [2] is an online official terminology dictionary website which is hosted by Department of Official Language, India. It contains Hindi as well as English terms along with definitions and example usage which are translations. The translation pairs were crawled from the website.

Indian Government corpora has been manually collected by CFILT staff from various websites related to the Indian government like the National Portal of India, Reserve Bank of India, Ministry of Human Resource Development, NABARD,*etc.*

Hindi-English Linked Wordnet contains bilingual dictionary entries created from the linked Hindi and English wordnets.

work done at IIT Bombay

[1] www.cfilt.iitb.ac.in

[2] e-mahashabdkosh.rb-aai.in

Corpus Id	Source	Number of segments
1	GNOME (OPUS) (Tiedemann, 2012))	145,706
2	KDE4 (OPUS)	97,227
3	Tanzil (OPUS)	187,080
4	Tatoeba (OPUS)	4,698
5	OpenSubs2013 (OPUS)	4,222
6	HindEnCorp (Bojar et al., 2014b)	273,885
7	Hindi-English Linked Wordnets (Bhattacharyya, 2010)	175,175
8	Mahashabdkosh: Administrative Domain Dictionary* (Kunchukuttan et al., 2013)	66,474
9	Mahashabdkosh: Administrative Domain Examples*	46,825
10	Mahashabdkosh: Administrative Domain Definitions*	46,523
11	TED talks (Abdelali et al., 2014)	42,583
12	Indic Multi-parallel corpus (Alexandra Birch and Post, 2011)	10,349
13	Judicial domain corpus - I* (Kunchukuttan et al., 2013)	5,007
14	Judicial domain corpus - II* (Kunchukuttan et al., 2012)	3,727
15	Indian Government corpora*	123,360
16	Wiki Headlines (Provided by CMU: `www.statmt.org/wmt14/wiki-titles.tgz`)	32,863
17	Gyaan-Nidhi Corpus * (`tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=281`)	227,123
Total		**1,492,827**

Table 1: Details of the IITB English-Hindi Parallel Corpus (training set). * indicates new corpora not in the public domain previously.

	Language	Train	Test	Dev
#Sentences		1,492,827	2,507	520
#Tokens	eng	20,667,259	57,803	10,656
	hin	22,171,543	63,853	10,174
#Types	eng	250,782	8,957	2,569
	hin	343,601	8,489	2,625

Table 2: Statistics of data sets

Gyaan-Nidhi Corpus is a multilingual parallel corpus between English and multiple Indian languages. The data is available in HTML format, hence it is not sentence aligned. We used the sentence alignment technique proposed by Moore (2002) to extract parallel corpora from this comparable corpus. This method combines sentence-length models and word-correspondence based models, and requires no language or corpus specific knowledge. We manually checked a small sample of 300 sentences from the parallel sentences extracted. We found that the precision of extraction of parallel sentences was 88.6%.

2.2. Corpus Statistics

The test and dev corpora are newswire sentences, which are the same ones as used in the WMT 2014 English-Hindi shared task (Bojar et al., 2014a). The training, dev and test corpora consist of 1,492,827 and 520 and 2507 segments respectively. Detailed Statistics are shown in Table 2. The Hindi and English OOV rate (for word types) is 11.4% and 6.7%.

3. Baseline Systems

We trained baseline machine translation models using the parallel corpus with popular off-the-shelf machine translation toolkits to provide benchmark translation accuracies for comparison. We trained phrase-based Statistical Machine Translation (PBSMT) systems as well as Neural Machine Translation systems for English-Hindi and Hindi-English translation.

3.1. Data Preparation

Text Normalization: For Hindi, characters with *nukta* can have two Unicode representations. In one case, the character and nukta are represented as two Unicode characters. In the other case, a single Unicode character represents the composite character. We choose the former representation. The normalization script is part of the *IndicNLP*[3] library .

For English, we used true-cased representation for our experiments. However, the parallel corpus being distributed is available in the original case.

Tokenization: We use the *Moses* tokenizer for English and the *IndicNLP* tokenizer for Hindi.

3.2. SMT Setup

We trained PBSMT systems with *Moses*[4] (Koehn et al., 2007). We used the *grow-diag-final-and* heuristic for extracting phrases, lexicalised reordering and Batch MIRA (Cherry and Foster, 2012) for tuning (default parameters). We trained 5-gram language models with Kneser-Ney smoothing using *KenLM* (Heafield, 2011). We used the *HindMono* (Bojar et al., 2014b) corpus for Hindi and the WMT NEWS Crawl 2015 corpus for English as additional monolingual corpora to train language models. These contain roughly 44 million and 23 million sentence for Hindi and English respectively.

[3] `anoopkunchukuttan.github.io/indic_nlp_library`

[4] `www.statmt.org/moses`

System	eng-hin		hin-eng	
	BLEU	*METEOR*	*BLEU*	*METEOR*
SMT	11.75	0.313	14.49	0.266
NMT	12.23	0.308	12.83	0.219

Table 3: Results for Baseline Systems

3.3. NMT Setup

We trained a subword-level encoder-decoder architecture based NMT system with attention (Bahdanau et al., 2015). We used *Nematus* [5] (Sennrich et al., 2017) for training our NMT systems.

Vocabulary: We used Byte Pair Encoding (BPE) to learn the vocabulary (with 15500 merge operations) (Sennrich et al., 2016b). We used the *subword-nmt* [6] tool for learning the BPE vocabulary. Since the writing systems and vocabularies of English and Hindi are separate, BPE models are trained separately.

Network parameters: The network contains a single hidden encoder and decoder RNN layer, containing 512 GRU units each. The dimension of input and output embedding layers is 256 units.

Training details: The model is trained with a batch size of 50 sentences and maximum sentence length of 100 using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0001. The output parameters were saved after every 10,000 iterations. We used early-stopping based on validation loss with *patience=10*.

Decoding: We used a beam size of 12. We decoded the test set with an ensemble of four models (best model and the last three saved models).

3.4. Results

We evaluated our system using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). We used a *METEOR-Indic*[7], a customized version of METEOR Indic, for evaluation of Hindi as target language. *METEOR-Indic* can perform synonym matches for Indian languages using synsets from IndoWordNet (Bhattacharyya, 2010). It can also perform stem matches for Indian languages using a trie-based stemmer (Bhattacharyya et al., 2014).

Table 3 shows the results of our experiments.

4. Availability

The homepage for the dataset can be accessed here: `http://www.cfilt.iitb.ac.in/iitb_parallel`.

The new corpora we release are available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike License [8]. The corpora we compiled from other sources are available under their respective licenses.

[5] `github.com/EdinburghNLP/nematus`

[6] `github.com/rsennrich/subword-nmt`

[7] `github.com/anoopkunchukuttan/meteor_indic`

[8] `http://creativecommons.org/licenses/by-nc-sa/4.0`

5. Conclusion and Future Work

We presented the IIT Bombay English-Hindi Parallel corpus version 1.0, and provided benchmark baseline SMT and NMT results on this corpus. This corpus has been used for the two shared tasks (Workshop on Asian Language Translation 2016 and 2017). The *HindiEn* component of the corpus has also been used for the WMT 2014 shared task. The corpus is available under a Creative Commons Licence.

In future, we plan to enhance the corpus from additional sources, most websites of the Government of India which is still a largely untapped source of parallel corpora. We also plan to build stronger baselines like pre-ordering with PBSMT (Ramanathan et al., 2008) for English-Hindi translation, and use of synthetic corpora generated via back-translation for NMT systems (Sennrich et al., 2016a).

6. Acknowledgements

We thank past and present members of CFILT for their efforts in creating various parts of the corpora over the years: Pallabh Bhattacharjee, Kashyap Popat, Rahul Sharnagat, Mitesh Khapra, Jaya Jha, Rajita Shukla, Laxmi Kashyap, Gajanan Rane and many members of the Hindi Wordnet team. We also thank the Technology Development for Indian Languages (TDIL) Programme and the Department of Electronics & Information Technology, Government of India for their support.

7. Bibliographical References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Alexandra Birch, Chris Callison-Burch, M. O. and Post, M. (2011). The indic multi-parallel corpus. `http://homepages.inf.ed.ac.uk/miles/babel.html`.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Bhattacharyya, P., Bahuguna, A., Talukdar, L., and Phukan, B. (2014). Facilitating multi-lingual sense annotation: Human mediated lemmatizer. In *Global WordNet Conference*.

Bhattacharyya, P. (2010). Indowordnet. In *Language Resources and Evaluation Conference*.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. (2014a). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Bojar, O., Diatka, V., Rychlỳ, P., Stranák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014b). Hindencorp-hindi-english and hindi-only corpus for machine translation. In *Language Resources and Evaluation Conference*.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., 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*.

Kunchukuttan, A., Roy, S., Patel, P., Ladha, K., Gupta, S., Khapra, M. M., and Bhattacharyya, P. (2012). Experiences in resource generation for machine translation through crowdsourcing. In *Language Resources and Evaluation Conference*.

Kunchukuttan, A., Chatterjee, R., Roy, S., Mishra, A., and Bhattacharyya, P. (2013). Transdoop: A map-reduce based crowdsourced translation for complex domain. In *Conference of the Association of Computational Linguistics (System Demonstrations)*.

Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Association for Computational Linguistics*.

Ramanathan, A., Hegde, J., Shah, R., Bhattacharyya, P., and Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *International Joint Conference on Natural Language Processing*.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Workshop on Machine Translation (Shared Task)*.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Conference of the Association of Computational Linguistics*.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Language Resources and Evaluation Conference*.

Wikipedia. (2017). Hindi. `https://en.wikipedia.org/w/index.php?title=Hindi&oldid=802224343`. [Online; accessed 02-October-2017].