The IIT Bombay English-Hindi Parallel Corpus

Anoop Kunchukuttan, Pratik Patel, Pushpak Bhattacharyya
Center for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{anoopk,pb}@cse.iitb.ac.in,pratikmehta1494@gmail.com

The IIT Bombay English-Hindi corpus contains parallel corpus for English-Hindi compiled from a variety of existing sources 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 sub-corpora in the download archive are in the same order as listed in the table, so they can be separately extracted if required.

We briefly describe some sub-copora which have not been described in previous literature. *Judicial domain corpus - I* consists of translations of legal judgements by expert translators, though not with a legal background. *Judicial domain corpus - II* contains translation done by students taking a graduate course on natural language processing as part of a course project. *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 *Indian Government corpora* has been manually collected by CFILT 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*.

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. The corpora can be downloaded from `http://www.cfilt.iitb.ac.in/iitb_parallel`.

We recommended the use of the following monolingual corpora for training language models - the corpora compiled by Bojar et al. (2014b) for Hindi, and the corpora provided by the WMT shared tasks [3] for English.

References

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

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014a. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics Baltimore, MD, USA.

Ondrej Bojar, Vojtech Diatka, Pavel Rychlỳ, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. 2014b. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*.

Anoop Kunchukuttan, Shourya Roy, Pratik Patel, Kushal Ladha, Somya Gupta, Mitesh M Khapra, and Pushpak Bhattacharyya. 2012. Experiences in resource generation for machine translation through crowdsourcing. In *LREC*.

[1] `http://www.cfilt.iitb.ac.in`
[2] `http://e-mahashabdkosh.rb-aai.in`
[3] `http://www.statmt.org/wmt14/translation-task.html`

Corpus Id	Source	Number of segments
1	KDE4 (Opus)	97,227
2	Tanzil (Opus)	187,080
3	Tatoeba (Opus)	4,698
4	OpenSubs2013 (Opus)	4,222
5	HindEnCorp (Bojar et al., 2014b)	273,885
6	Hindi-English Linked Wordnets (Bhattacharyya, 2010)	175,175
7	Mahashabdkosh: Administrative Domain Dictionary	66,474
8	Mahashabdkosh: Administrative Domain Examples	46,825
9	Mahashabdkosh: Administrative Domain Definitions	46,523
10	TED talks	42,583
11	Indic Multi-parallel corpus (Alexandra Birch and Post, 2011)	10,349
12	Judicial domain corpus - I	5,007
13	Judicial domain corpus - II (Kunchukuttan et al., 2012)	3,727
14	Indian Government corpora	123,360
15	Wiki Headlines	32,863
16	Book Translations (Gyaan-Nidhi Corpus)	227,123
Total		**1,492,827**

Table 1: Statistics of the IITB English-Hindi parallel corpus (training set)