CVIT's submissions to WAT-2019

Jerin Philip, Shashank Siripragada, Upendra Kumar, Vinay. P. Namboodiri, C.V. Jawahar

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Overview

We present CVIT's submissions to WAT-2019.

- **Tasks**: Hindi-English and Tamil-English.
- **Models**: Transformer-Base
- **Directions**:
  - Multiway shared encoder decoder (7 languages)
  - Unidirectional (English-Tamil)
- **Tokenization**: SentencePiece

Leaderboard best in English-Tamil task (both directions) and Hindi to English task.

Datasets

For multilingual model (ilmulti), we use a compilation of parallel corpora across several available datasets.

Source	#pairs	type
IITB-hi-en[1]	1.5M	en-hi
Backtranslated-Hindi	2.5M	en-hi
WAT-ILMPC[2]	188K	xx-en
ILCI[3]	50K	xx-yy
Backtranslated-wiki	10.4M	mono

Table 1: Training dataset used for ilmulti model. xx-yy indicates parallel sentences aligned across multiple languages. xx-en indicates bilingual corpora with English in one direction.

In case of English-Tamil task, we use the UFAL English-Tamil dataset provided[4] along with some additional monolingual data obtained from the web.

Source	#pairs	type
UFAL EnTam	160K	en-ta
Leipzig Newscrawl	300K	ta mono
Indian Politics News	300K	en mono

Table 2: Training dataset used for UFAL English-Tamil task.

Components

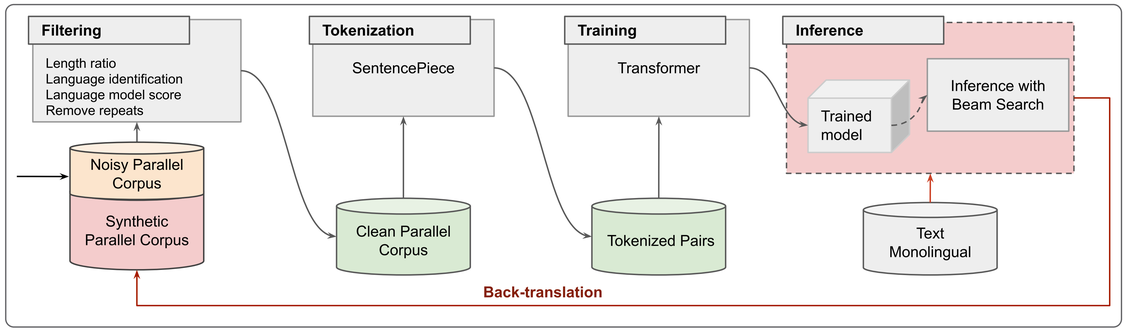

Figure 1: Training and Inference Pipeline including Back-translation

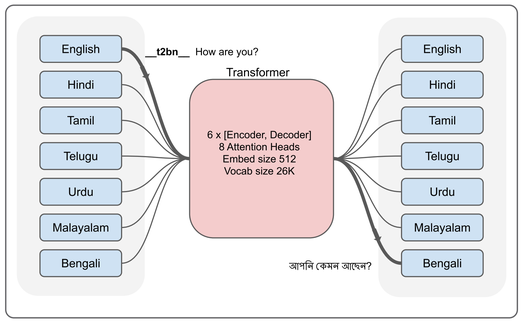

Figure 2: Our ilmulti model formulation. The system is capable of translating between 7 languages, in addition to performing well on IIT-Bombay Hindi-English test-sets.

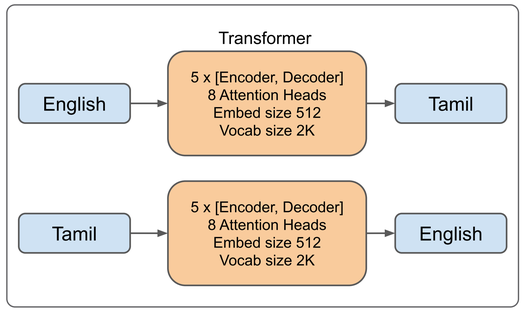

Figure 3: We use separate models for both directions to get the best numbers in UFAL English-Tamil test sets.

Results

Table 3: Results on IIT-Bombay Hindi-English and UFAL English-Tamil test sets. 3 and 4 indicate BLEU obtained during ilmulti inference out-of-box and warm-start respectively. Bold indicates best values among all submissions at the time of writing this paper.

No	Model	BLEU	
		en-hi	hi-en
1	ilmulti	20.17	22.62
2	1 + backtranslation	20.46	**22.91**
		en-ta	ta-en
3	2	0.80	4.68
4	2 + UFAL warm-start	10.91	27.14
5	UFAL cold-start	**13.05**	**30.04**

Table 4: Automated evaluation scores on the UFAL EnTam v2.0 test set. This table demonstrates incremental improvements which got us to the final submission in Table 3. † indicates numbers from the submission site, others were computed locally and have minor differences.

Id	Model	BLEU	
		en-ta	ta-en
U-1	Transformer-base	11.59	27.31
U-2	+ filtered	11.73	27.58
U-3	+ ensemble	11.96	28.05
U-4	+ backtranslation	12.63	29.21
U-5	+ ensemble	12.87	29.75
U-6	+ length penalty	13.14	30.10
U-6	(submission site)	**13.05**†	**30.04**†

Training

Software: `fairseq`, `sentencepiece`

Multiway model: ilmulti

- multi-node training with 4 x [4 x 1080Ti] NVIDIA GPUs.
- cross entropy criterion.
- Inference: Beam size 10.

English-Tamil unidirectional models

- 1 node with [4 x 1080Ti] NVIDIA GPUs.
- cross entropy criterion with label smoothing.
- Inference: beam size 5, length penalty 2.0

References

[1] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi Parallel Corpus. In *LREC*, 2018.

[2] Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. Overview of the 5th workshop on asian translation. In *WAT*, 2018.

[3] Girish Nath Jha. The TDIL Program and the Indian Language Corpora Intitiative (ILCI). In *LREC*, 2010.

[4] Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*, pages 113–122, 2012.

Project Information

Find interactive demo of models and more project information in the below website.

bhasha.iiit.ac.in/mt