The University of Edinburgh's Bengali↔Hindi Submissions to the WMT21 News Translation Task

THE UNIVERSITY of EDINBURGH

Proyag Pal **Alham Fikri Aji** **Pinzhen Chen** **Sukanta Sen**

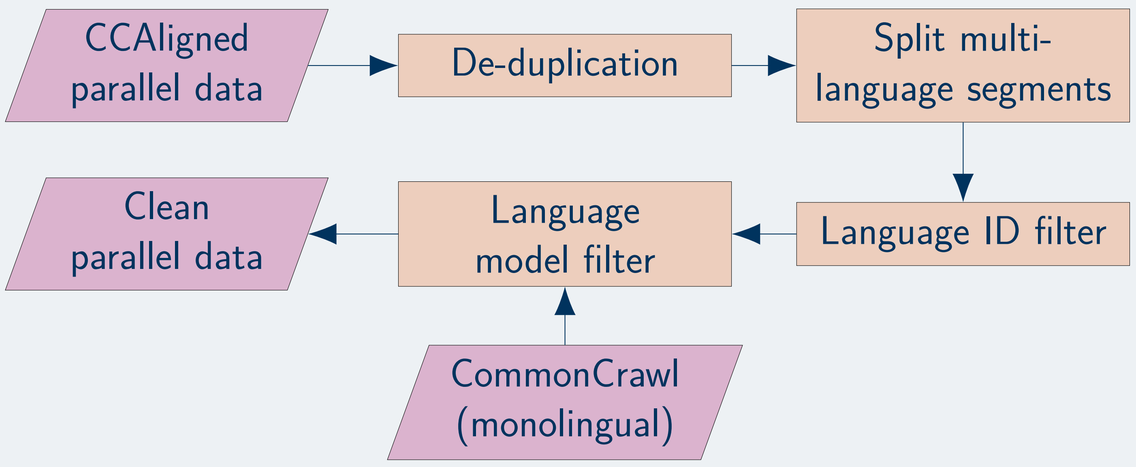
Overview

- **UEdin**'s Bengali→Hindi (bn→hi) and Hindi→Bengali (hi→bn) systems submitted to the News Translation task at WMT21.
- **Top (tied)** systems among all **constrained** submissions for both directions, according to human evaluation.
- Our approach mainly focuses on **cleaning**, **back-translation**, and **fine-tuning to the target domain**.
- All models are trained with **parallel and synthetic data**, **fine-tuned on retrieved in-domain data**, further **fine-tuned on dev set**. Models fine-tuned in different ways are **ensembled**.

Data and Cleaning

Constrained condition:

- 3.3M parallel sentence pairs from CCAligned
- NewsCrawl monolingual: 10.1M lines of bn, 46.1M lines of hi
- CommonCrawl monolingual: 49.6M lines of bn, 202M lines of hi



Training with Synthetic Data

Use back-translation and forward translation using models trained only on parallel data to generate synthetic data. Use this synthetic data in different ways:

- Tagged back-translation
- Train models on all back-translated data, then continue training with parallel data only
- Train on parallel, back-translated, and forward translated data, then continue training with parallel data only

Decoding and Post-processing

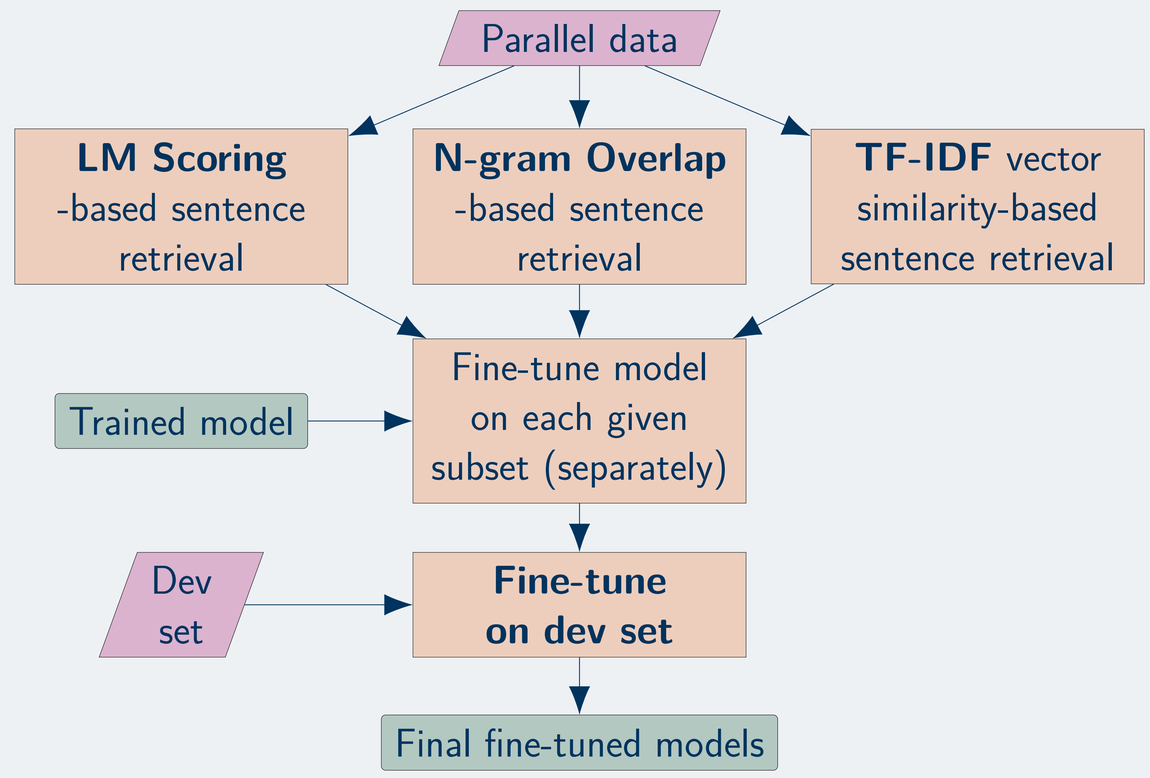
- Ensemble many fine-tuned models to decode.
- Run sentence splitter on test source; rejoin outputs.
- Transliterate all numerals to Latin script for consistency.

Model and Training Configuration

- **32k subword** SentencePiece vocabulary **shared** between source and target sides.
- **Transformer-Big** architecture - 6 encoder + 6 decoder layers, 16 heads, embedding size 1024, unit size 4096.
- 32GB dynamic batch size, Adam optimizer with learning rate 0.0003, optimizer delay 3, early stop when dev set BLEU doesn't improve for 20k updates.

Fine-tuning

Adapt to the target domain by retrieving sentences similar to the dev/test set and fine-tuning the models on those subsets of sentences. Finally, fine-tune to the dev set, since that's the most in-domain data available.



Human Evaluation

We produced the best constrained systems (tied) for both directions.

Ave.	Ave. z	System
82.1	0.202	GTCOM
79.1	0.163	Online-B
77.5	0.080	TRANSSION
78.0	0.076	MS-EgDC
78.0	**0.054**	**UEdin**
76.1	-0.015	Online-Y
75.7	-0.080	HuaweiTSC
75.7	-0.107	Online-A
70.8	-0.373	Online-G

(a) bn→hi

Ave.	Ave. z	System
95.0	0.245	HuaweiTSC
94.8	0.236	Online-A
94.5	0.233	GTCOM
94.6	**0.214**	**UEdin**
92.3	0.080	Online-Y
92.0	0.045	TRANSSION
91.3	0.029	Online-B
90.9	-0.008	MS-EgDC
73.5	-1.100	Online-G

(b) hi→bn

□ constrained ▒ unconstrained

Our submissions are in bold. Systems within a cluster are considered tied.