Appendix

MT Experimental Setup

We conduct the two sets of experiments, similar to the NER experiments:

1. Experiments where word embeddings are transferred from Uyghur and Bengali, using Turkish and Hindi as the respective high resource languages. Both the proposed models, CT-Joint and CT-FineTune, are used for experimentation with select subword units.

2. Monolingual experiments on the two low resource languages: Uyghur and Bengali, with select subword combinations.

Data Processing

We use data, comprised of unlabeled corpora, training translation pairs (if provided), from the same sources as used for the NER experiments. The training data comprises of translation pairs between the source language and the target language, English. We create our own train-dev-test splits for the experiments as there are no official splits provided. The data splits can be seen in Table 1. The Uyghur corpus has 31 million tokens (extracted from a different set than the one used for NER) and the Turkish corpus about 40 million tokens. The Bengali corpus has 125 million tokens and we downsampled the original Hindi corpus to a comparable size.

Lang.	**Train**	**Dev**	**Test**
Uyghur	99379	994	994
Bengali	101523	1989	1988

Table 1: Sentences in train/dev/test set for MT.

Model Setup

We train the model using 512-dimensional word embeddings, pre-trained using strategies described in the main text. We run each MT system 3 times and report the median score, as a way to control for variance in training.