A Appendix

A.1 Language Centroids

We select 5k monolingual sentences from Wikipedia for 19 languages (each with at least 20 characters). Then, we normalize them by removing all punctuation, and use them to estimate language centroid vectors for each language. To do so, we first obtain their sentence embeddings by executing the mean pooling operation for the last layer of m-BERT (or XLM-R) contextualized word embeddings without [CLS] and [SEP] tokens involved. Then, we average these sentence embeddings to obtain language-specific centroid vectors.

A.2 Our Modifications

Re-mapping We fine-tune m-BERT ($L = 12$, $H = 768$, $110M$ params) and XLM-R ($L = 12$, $H = 768$, $70M$ params) on the concatenated mutual word translations of 18 languages paired with English, using the loss function obtained as Eq. 3. The mutual word translations are extracted with FastAlign (Dyer et al., 2013) on parallel text from the combination of following publicly available parallel corpora.

- Europarl (Koehn, 2005): We select 9 languages (German, Spanish, French, Italian, Dutch, Finnish, Hungarian, Portuguese, Estonian) out of 21 languages from Europarl. The size varies from 400k to 2M sentences depending on the language pair. We extract 100k parallel text for each language paired with English.
- JW300 (Agić and Vulić, 2019): We select the remaining languages (Tagalog, Bengali, Javanese, Marathi, Hindi, Urdu, Afrikaans, Malay, Indonesian) out of 380 languages from JW300. The average size is 100K parallel sentences per language pair. We extract 100k parallel text based on sampling for each language paired with English.

Input Normalization In the RFEval setup, we do not modify system translations (in English), and instead manipulate source language texts. For XNLI, we manipulate both premise and hypothesis texts. To examine the impact of linguistic changes to cross-lingual transfer, we remove all punctuation from input texts. We extract word and lemma forms, universal part-of-speech (POS) tags, morphological features and universal dependency relations from input texts using UDPipe, which is a pipeline trained on UD treebank 2.5. Each orthographic token is split into several tokens that can be directly obtained from the corresponding word forms. To reverse noun-adjective and object-verb ordering, we use a simple rule-based strategy based on universal POS tags and universal dependency relations.

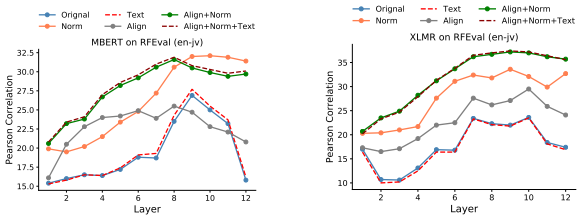

Figure 10: Results in one low-resource language pair (en-jv).