Semantic merge

Dataset

I used the MaCoCu dataset, since it scored well in the manual translation quality review (brief manual overview using an LLM to judge correctness of translation).

The merge rate was quite low; 11.7% for Croatian and Slovenian only even though the quality of the dataset should be quite high.

Method

1) Split the keys into hr_only (1.7 million keys) and sl_only (1.4 million keys).
2) Embed the english side of every key (model used = all-MiniLM-L6-v2) a simple and more importantly fast embedding model.
3) Index and search with ChromaDB: The sl_only embeddings were loaded into a chroma collection; every hr_only string when then queried against these keys for it's nearest neighbour by cosine similarity.

Result

0-0.5	2.9%
0.5-0.6	23.9%
0.6-0.7	42.7%
0.7-0.8	21.9%
0.8-0.9	5.5%
0.9-1	3%

The pairs from 0.6-0.8 don't seem 'ideal' but the sentences are clearly related but definitely also not a 1-1 translation, anything lower is essentially noise from my brief look at the sampled sentences. I will include the .txt so you can judge yourself. Real quality seems to start around 0.8 and up.

Conclusion

It's hard to say based on 1 corpus. Other corpus could have higher or lower merge rates and the result is definitely not bad but it did take a very long time to run (+- 6 hours) and, originally there are 500 million english keys so this would obviously take an enormous runtime.