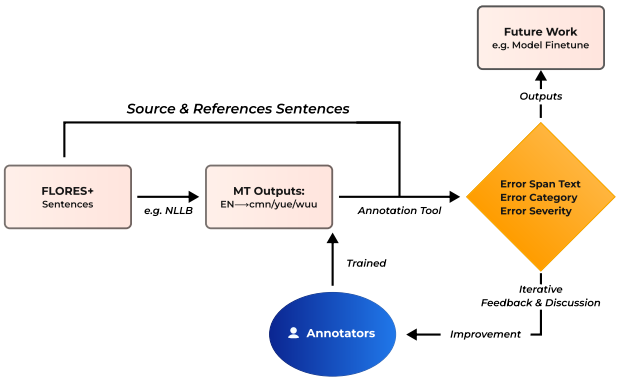
SiniticMTError: A Machine Translation Dataset with Error Annotations for Sinitic Languages

Despite major advances in machine translation (MT) in recent years, progress remains limited for many low-resource languages that lack large-scale training data and linguistic resources. Cantonese and Wu Chinese are two Sinitic examples, although each enjoys more than 80 million speakers around the world. Thus, we introduce SiniticMTError, a novel dataset that builds on existing parallel corpora to provide error span, error type, and error severity annotations in machine-translated examples from English to Mandarin, Cantonese, and Wu Chinese. SiniticMTError is a novel suite of datasets that build on `FLORES+` [1] to provide erroneous machine translation examples and detailed span-level error annotations including error type and severity for Mandarin, Cantonese, and Wu Chinese.

Our dataset serves as a resource for the MT community to utilize in fine-tuning models with error detection capabilities, supporting research on translation quality estimation, error-aware generation, and low-resource language evaluation. We also support the community with our annotation guidelines specifically tailored for Sinitic languages. We have now finished annotating across 2,000 Mandarin and 1,000 Cantonese sentences, with Wu Chinese annotations still in early start. We describe our annotation pipeline conducted by native speakers in Figure 1a.

We also report the error type distribution of the annotated Mandarin sentences in Figure 1b. Mistranslation errors appear the most (61.2%), which shows the difficulty of models accurately understanding Mandarin semantics. Omission (14.2%) and grammar errors (7.1%) are also frequently present. This suggests that MT systems often do not keep the full meaning or produce Mandarin sentences with proper structure. Less frequent cases, including unintelligible outputs, typography errors, and untranslated segments, still occur and show that many different kinds of errors exist in MT. These distributions show the need of span-level annotation in Sinitic MT, and provide concrete targets for future error detection modeling approaches.



(a) Annotation pipeline from FLORES+ input to span-level error labels.

Type	Count	Proportion
Mistranslation	2306	61.2%
Omission	534	14.2%
Grammar	267	7.1%
Unintelligible	198	5.3%
Typography	133	3.5%
Untranslated	128	3.4%
Spelling	114	3.0%
Addition	86	2.3%

(b) Error-type distribution in Mandarin (2009 sentences).

Figure 1: Overview of our dataset: (a) annotation pipeline; (b) error-type distribution.

References

[1] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846, 2024.