No Language Left Behind:

Low resource languages and how we can improve it.

(Flores 200)

↳ focus on low resource languages

it's a way to way to ensure cross-checking

english → chinese
english → french

document-level (not sentence level)

↳ uses document-level context to clear ambiguities

a major challenge is the collection of multiple scripts & support for local variations of languages

NLLB seed → sentences sampled from Wikimedia's list of articles for every language in Wikipedia.

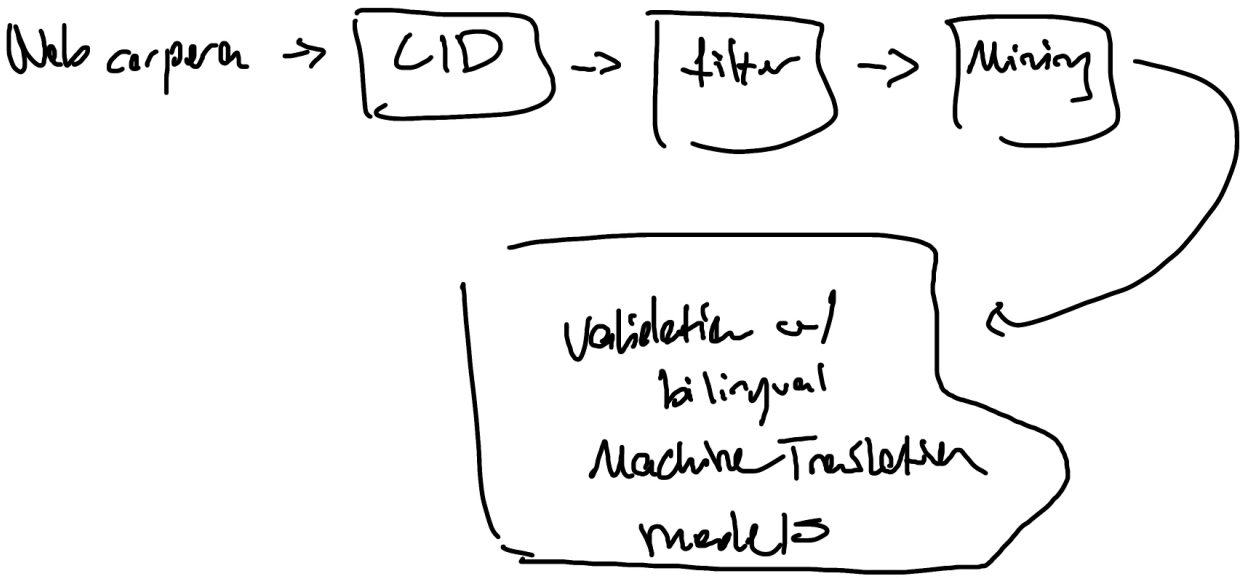
data mining:

sentence alignment: a technique for finding correspondence between source sentences and equivalent translations.

Steps Language Pipeline

Language identification classifier (LID)

(which is evaluated by human)



Mixture of Experts:

training multiple models to divide problem space into regions