

Translation Memories as Baselines for Low-Resource Machine Translation

Overview

- We propose using translation memories (TMs) as baselines and as tools to analyze low-resource machine translation datasets – complementary to machine translation baselines, detailed human analysis, and automatic or human evaluation.

What is a translation memory?

- A collection of source and target language sentence pairs, along with a way of measuring similarity between sentences
- When translating a new sentence in a computer aided translation interface, the TM can provide the translator with translations of similar source sentences to edit rather than translating from scratch
- *If you have parallel text, you have a TM* (though it might not be a highly-curated one)

Why use a TM baseline?

- Low-resource MT is often of too low quality to use without translator correction; this is a baseline comparable to the TM assistance a translator might have access to
- Can provide insights about data
- Can be used alongside automatic metrics and deeper human analysis to do evaluation

Implementation

Given a set of parallel sentences (the TM) and a test set and a function M that measures the similarity of two strings

- For every source sentence s in the test set:
 - Find s' in the TM that maximizes $M(s',s)$
 - Return t' (the translation of s') as the hypothesis translation of s

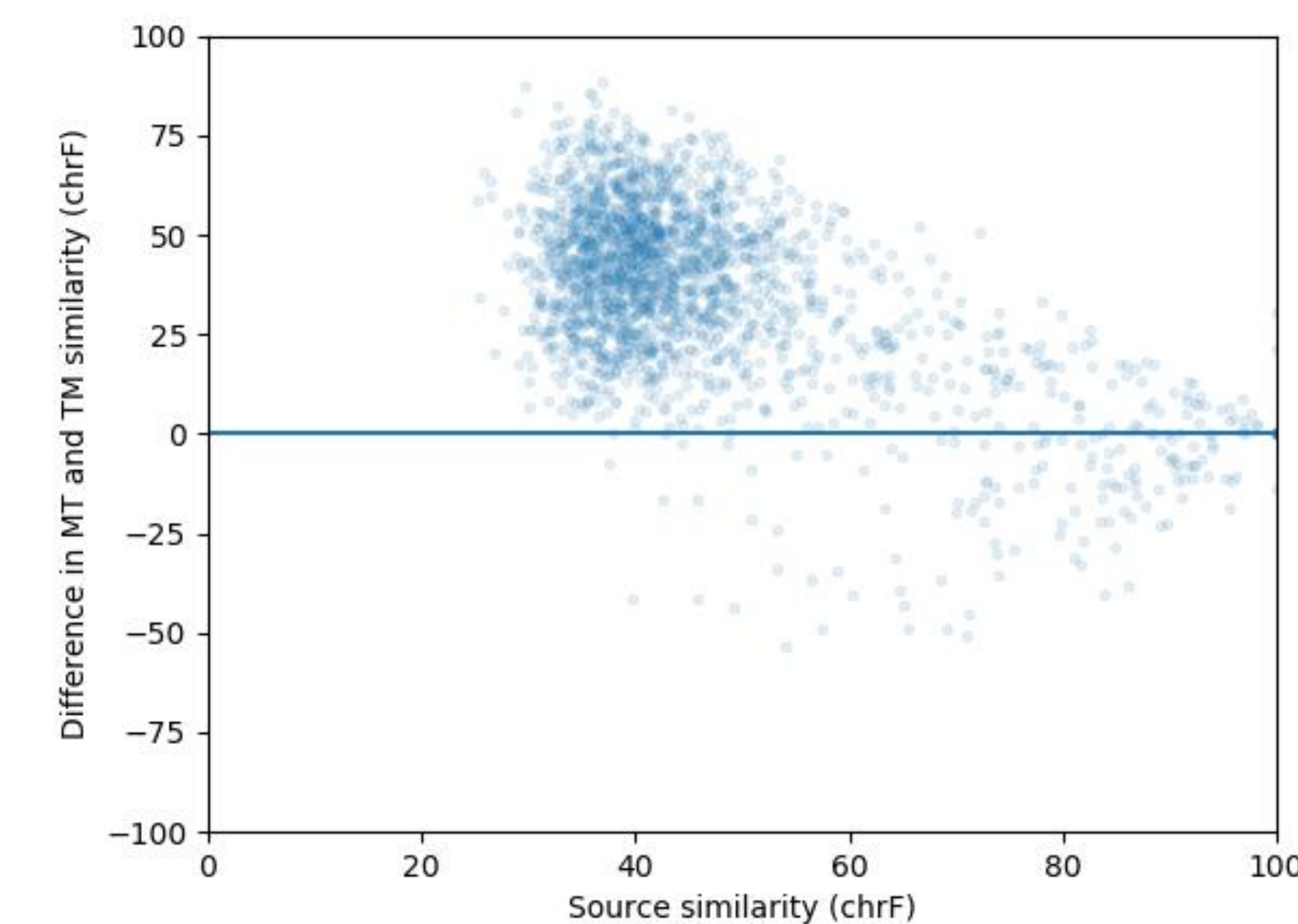
Compute your preferred automatic metric score over the returned hypotheses as compared to test references. As with prior work, we find it is best to use the same automatic metric for the TM and for scoring.

Observations

- Different low-resource tasks vary in their data size and how well-matched (in terms of domain, orthography, dialect, etc.) their training, development, and test data are
- In some recent shared tasks, development data is much more closely related to test data than training data is

	Close Match	Mismatch
Low-Resource	Upper Sorbian	AmericasNLP
Mid.-Resource	Inuktitut Hansard	Inuktitut News

Fine-grained analysis



- Using sentence-level automatic metrics, we can estimate how much MT systems improve over TM baselines (above the horizontal axis)
- This can provide preliminary insight into whether MT systems might be more useful to translators than TM data

Limitations and Future Work

- Current approach is not optimized; quite slow for larger datasets (larger datasets might be able to use tools like FuzzyMatch-cli to balance the speed-coverage tradeoff)
- Not a replacement for human evaluation or analysis; complementary to existing approaches
- TM baselines could be used to help tease apart the contributions of data size increase vs. model/architecture/training improvements