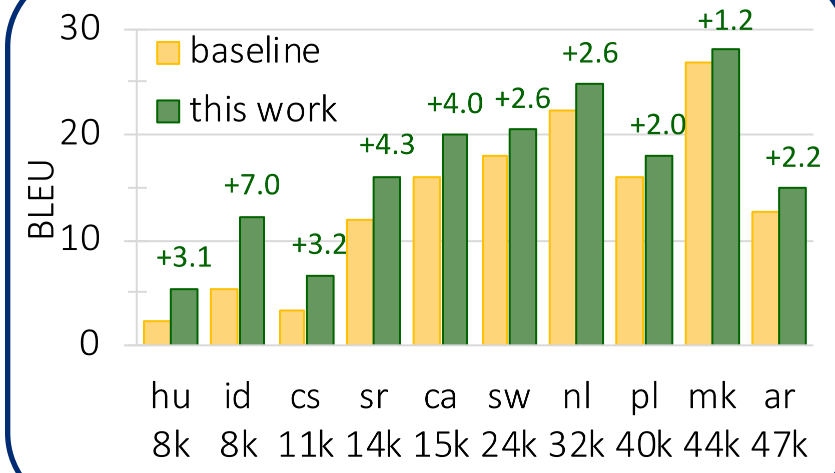
Simulated Multiple Reference Training: Leveraging Paraphrases for Machine Translation

Huda Khayrallah| huda@jhu.edu | cs.jhu.edu/~huda | Johns Hopkins University

NMT is sensitive to the ***quality*** and ***quantity*** of the training data

use target language ***paraphrasing*** to overcome data sparsity in ***low-resource*** settings



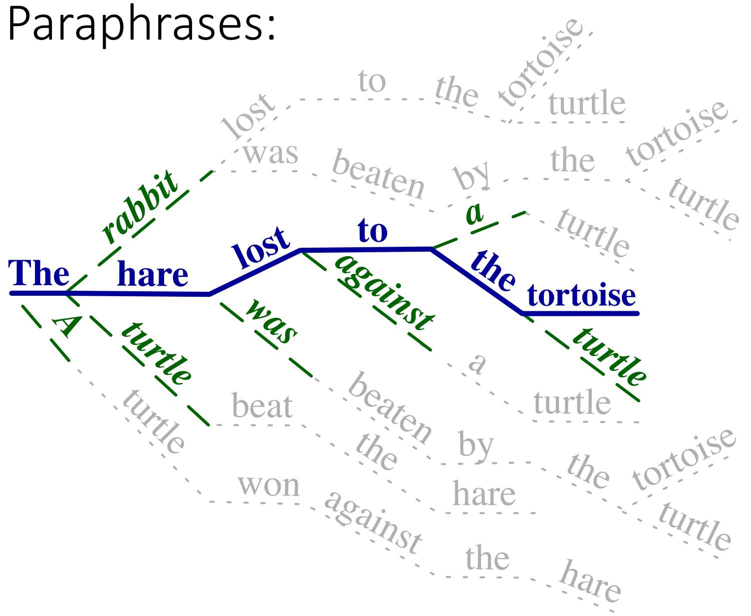
Source:

La tortuga ganó contra la liebre

Translation:

The turtle beat the hare

Paraphrases:



$$\text{NLL:}\quad -\sum_{v\in\mathcal{V}}\Big[\underbrace{\boxed{\mathbb{1}\{y_i = v\}}}_{\text{Gold Target}} \times \log \underbrace{\boxed{p_{\text{MT}}(y_i = v \mid x; y_{j<i})}}_{\text{MT Model output distribution}}\Big]$$

$$\text{SMRT:}\quad -\sum_{v\in\mathcal{V}}\Big[\underbrace{\boxed{p_{\text{para}}(y'_i = v \mid y; y'_{j<i})}}_{\text{Paraphraser output distribution}} \times \log \underbrace{\boxed{p_{\text{MT}}(y'_i = v \mid x; y'_{j<i})}}_{\text{MT Model output distribution}}\Big]$$

- Simulated Multiple Reference Training Improves Low-Resource Machine Translation by: **Khayrallah**, Thompson, Post, Koehn (@ EMNLP 2020)
- SMRT Chatbots: Improving Non-Task-Oriented Dialog with Simulated Multi-Reference Training by **Khayrallah** & Sedoc (@ EMNLP Findings 2020)
- Code, data & more: data.statmt.org/SMRT