CUTTING THE LONG TAIL: HYBRID LANGUAGE MODELS FOR TRANSLATION STYLE ADAPTATION

ARIANNA BISAZZA & MARCELLO FEDERICO

Scenario: Translation of TED talks ⇒ large topic diversity
Problem: Not enough in-domain data to ensure good coverage.
Large out-of-domain data (news) improves coverage but alters style

Approach: Map infrequent (= topic-specific) words to Part-of-Speech
Train LM on mapped in-domain data ⇒ word/POS-level hybrid LM
Integrate into decoding to boost the style of the talks
Result: Better n-gram coverage, more discriminative than POS-level LM

Translation of TED talks

Target LM training corpora

Language styles

In-domain target data

Hybridly mapped word/POS data

Translation quality

Talk-level analysis

Talk-level evaluation on Arabic-English tst2010:
LM perplexities and impact on BLEU