

Confidence Estimation for Statistical Machine Translation

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur,
Cyril Goutte, Alex Kulesza, Alberto Sanchis, Nicola Ueffing

Confidence Estimation for SMT

- estimate probabilities of correctness for MT output■
- Jelinek: estimate probabilities of correctness for *incorrect* MT output■
- try to distinguish between slightly incorrect and very incorrect translations■
- applications exist!

Automatically Assessing MT Correctness

- find some sentence-level error function:

$$E(\text{translation, reference-translations})$$

that correlates well with human judgements of translation quality

- candidates: WER, WER-d, PER, WER+PER, Melamed's F-measure, NIST, delta BLEU, smoothed BLEU:

$$\text{smoothed-precision} = \frac{\text{matches} + \epsilon}{\text{count} + 2\epsilon}$$

Automatically Assessing (cont)

- evaluate by comparing to human ratings of approx 1000 sentences by FE's:
 - adequacy-biased 1-5 scale
 - reduce annotator effects by transforming to quantiles
- preliminary results: inter-annotator agreement fair; correlation with automatic measures poor
- \Rightarrow choose top one or two best measures for further experiments

Learning Setup

- data from nbest lists generated by base SMT system
- examples of the form: (source, nbest hyp, E)
- extract features from source and hyp
- do:
 - direct regression on E
 - classification: okay if E below a threshold τ , for various values of τ intended to reflect different applications

Learning (cont)

- evaluate using:
 - direct error measure: MSE or NLL
 - discriminability: ROC curves, IROC, classification error
 - applications

Features

- how hard is the source sentence to translate?
 - length of source sentence
 - distribution of its ngrams across quartiles
 - trigram LM probability
 - search features: avg active hyps post pruning, avg hyps pruned per cardinality, avg best and worst hyp scores
 - nbest length and density
 - number of unique content word stems in nbest list

Features (cont)

- how strange is the target hypothesis?
 - trigram LM probability
 - LM probabilities from ngrams over nbest list
 - nbest rank and score relative to best score
 - semantic distance to top hypotheses
 - Levenshtein distance to center hypothesis
 - mismatched quotes and parens

Features (cont)

- how related are the source sentence and target hypothesis?
 - base model scores
 - length ratio
 - average source fertility
 - alignment monotonicity
 - word posterior probabilities over nbest list
 - M1 probability

Status

- feature generation almost complete
- preliminary results for WER and NIST measures:
 - feature ranking: no clear pattern (base-model score, length ratio, source quartiles)
 - model ranking: MLP > NB > base score
- to do:
 - ML experiments: models, features, classification/regression, regime, learning curve
 - applications: model combination (CMU data), reranking

Sub-sentence CE

- don't need error measures; can use exact match on words or ngrams
- but several definitions of “correctness”: exact position match in ref, presence/absence in ref, presence/absence without replacement, Levenshtein aligned with self, Melamed-aligned
- rough correspondence with error measures, resp.: SER, PER, BLEU/NIST, WER, F-measure
- applications: post-editing, recombination