TerrorCat: a Translation Error Categorization-based MT Quality Metric

Mark Fishel@cl.uzh.ch, Rico Sennrich@cl.uzh.ch, Maja.Popović@dfki.de, Ondřej Bojar@ufal.mff.cuni.cz

Addicter:
- Ref-hyp word-align (lemma, context)
- Unaligned words: missing/extra
- Aligned words: inflection/lex.choice error (based on the lemmas)
- “Untangle” alignment to find min. amount of reordering errors
- punctuation errors as separate class

Hjerson:
- Computes WER and PER on lemmas
- missing/extra tokens are the ones that contribute to both WER and PER
- inflection errors do not contribute to either, but their lemmas don’t match
- order errors contribute only to WER
- The remaining are lex.choice errors

Pairwise Comparison:
- Binary SVM classifier, linear kernel
- WMT 2007–2011 manual rankings for training and development data sets (converted to comparison pairs)
- Input vectors are filled with differences of rel. error frequencies per category
- Addicter- and Hjerson-based analysis vectors are “concatenated”

Optimization:
- Variable reward to the winner proportional to the confidence estimate of the classifier
- PoS-specific features, e.g.: missing noun/misplaced verb/mis-inflected adj./...
- Generalizing mapping for PoS-tags
- Data from 2007 is an outlier; tried excluding it from the training material
- Addicter’s own alignment vs trained bilingual alignment
- Language-pair-specific vs tgt-language-specific vs universal models