Work on machine-translation evaluation measures continued as Kahn, Ostendorf and Roark developed Expected Dependency Pair Match (EDPM).

EDPM uses a popular PCFG syntactic parser [CJ05] to extract expected counts of dependency structures from hypothesis and reference translations and score them. EDPM is based on the intuitive and difficult-to-game F measure, where the tokens to be matched are drawn from 1-grams, 2-gram, and in- and out-bound dependency links. In this respect, it extends the dependency-scoring strategies of [OvGW07] but does so with a widely-used and publicly available PCFG parser and head-finding rules. Further, EDPM incorporates string-only features and uses the parser's own confidence to predict the (hidden) dependency structure. EDPM's free parameters thus include the number of parses n used in the calculation, the classes of tokens to be extracted for the F-measure, and the degree in which to trust the parser's confidence distribution over the n-best list.

In experiments over the Multiple Translation Chinese Corpus [LDC03, LDC06], we explored a variety of measures in the EDPM family and selected the variant that best-correlated with human judgments of fluency and adequacy per sentence. Over this corpus, the selected variant of EDPM correlated at $r = 0.240$, much better than popular measures TER ($r = -0.173$) and (add-one-smoothed) BLEU_4 ($r = 0.218$).

The utility of this chosen variant was further tested in predicting the difference in human-targeted translation edit rate (ΔHTER) between two translations of the same source from the unsequestered GALE 2.5 [DAR08] Arabic-to-English and Chinese-to-English task. EDPM's per-document correlation with ΔHTER over this corpus was better ($r = -0.47$) than BLEU_4 ($r = -0.32$) or TER ($r = 0.39$), with similar effects on per-sentence correlations.

EDPM was submitted to the 2008 NIST MetricsMATR MT-evaluation competition. In this competition, EDPM performed the best overall on document-level HTER correlation, and within the confidence of the best-performance on all HTER measures (and nearly all other human-derived measures as well).

Future work on EDPM will include the ability to change parsers (making it much easier to apply it to new domains or languages), partial-match scoring (using synonyms and other partial-match techniques) and exploration of other dependency graph-fragments.

References

[CJ05] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proc. ACL*, pages 173–180, 2005.

[DAR08] DARPA. Global Autonomous Language Exploitation (GALE). Mission, `http://www.darpa.mil/ipto/programs/gale/gale.asp`, 2008.

[LDC03] LDC. Multiple translation Chinese corpus, part 2, 2003. Catalog number LDC2003T17.

[LDC06] LDC. Multiple translation Chinese corpus, part 4, 2006. Catalog number LDC2006T04.

[OvGW07] Karolina Owczarzak, Josef van Genabith, and Andy Way. Labelled dependencies in machine translation evaluation. In *Proc. 2nd Workshop on Statistical MT*, pages 104–111, 2007.