Automatic Syntactic MT Evaluation with Expected Dependency Pair Match

Jeremy G. Kahn and **Mari Ostendorf**
Signal, Speech & Lang. Interpretation Lab
University of Washington, Seattle, WA
{jgk,mo}@ssli.ee.washington.edu

Brian Roark
Ctr. for Spoken Lang. Understanding
OHSU, Portland, OR
roark@cslu.ogi.edu

In improving MT metrics, we try to model acceptable variation, whether by modeling word-choice (e.g. METEOR (Banerjee and Lavie, 2005)), by weighting adjacent matches more than non-local matches (e.g. GTM (Turian et al., 2003)) or by modeling syntactic information (Liu and Gildea, 2005). Owczarzak et al. (2007) explore the correlation of their dependency-syntax based measure **d** and **d_var** with human judgment, and report substantial improvements relative to the popular measures BLEU and TER.

In keeping to the syntactic approach, we present Expected Dependency Pair Match (EDPM), which follows and extends the labelled-dependency match version of SPARSEVAL (Roark and others, 2006) and the **d/d_var** (Owczarzak et al., 2007) measures. These approaches evaluate hypothesis-reference similarity with an F measure over fragments of a labelled dependency structure, which may be generated by a PCFG with deterministic head-finding (Liu and Gildea, 2005; Roark and others, 2006) or by extracting the semantic dependencies from an LFG parser (Cahill and others (2004) in Owczarzak et al. (2007)).

EDPM extends these partial-dependency-scoring strategies with a widely used and publically available PCFG parser and deterministic head-finding rules instead of an LFG system. In addition, it incorporates word-level matching and weighted multiple parse alternatives for improved performance.

1 Definition of DPM family of metrics

We define a family of Dependency Pair Match (DPM) measures that is composed of extensions of the methods described inr Owczarzak et al. (2007). DPM is defined as the F measure over bags-of-subtrees of the hypothesis translation dependency tree as compared to bags-of-subtrees of the reference translation dependency tree. Figure 1 demonstrates a toy example of the bags-of-dependencies extracted from a hypothesis and reference tree. In the example presented here, we extract only dlh dependencies, which are tuples of the form $\langle$Dependent, arc-Label, Head$\rangle$.

1.1 DPM variations in subtree extraction

Different members of the DPM family of metrics may extract different subtrees. We denote the set of extracted tree-components with a trailing subscript: DPM_{dlh} extracts all $\langle$Dependent, arc-Label, Head$\rangle$ subtree tuples, roughly equivalent to labeled SPARSEVAL (Roark and others, 2006) and the Owczarzak et al. (2007) **d** measure. We also consider the $\text{DPM}_{1g,2g}$ extractions, which represent unigrams and bigrams, or the $\text{DPM}_{dl,lh}$, which extracts all the subtrees $\langle$Dependent, arc-Label$\rangle$ and $\langle$arc-Label, Head$\rangle$ (roughly equivalent to the Owczarzak et al. (2007) **d_var** method).

1.2 DPM variations using n-best lists and expected counts

Since the dependency structures of the hypothesis and reference text are hidden, we also explore alternative dependency structures predicted by the parser, to cope with genuine ambiguity (in both hypothesis and reference) and to mitigate the effects of parser error. DPM is well-defined over the n-best list of dependency-structures: when $n > 1$, DPM uses the expectation of bags-of-subtrees rather than the bags-of-subtrees derived from the 1-best parse.

An expectation requires a probability distribution over the n-best list, and we "flatten" the parser probabilities such that $\tilde{p}(x) = \frac{p(x)^\gamma}{\sum_i p(i)^\gamma}$ (where γ is a free parameter) to account for over-confidence in the parser. In all cases, the probabilities are normalized to sum to one over the the n-best list.

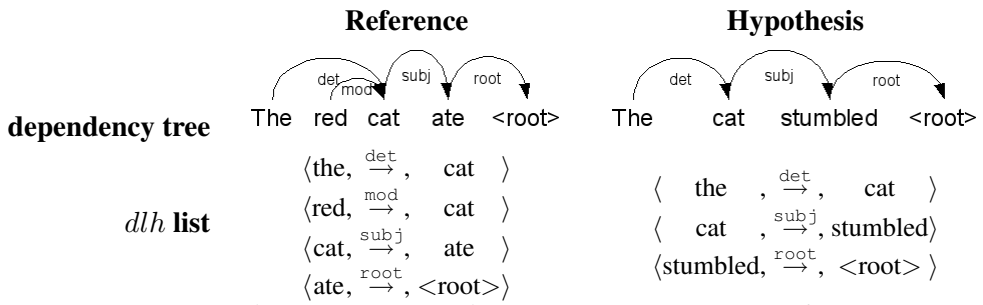

Figure 1: Example hypothesis and reference dependency trees and the dlh decomposition of each.

1.3 Implementation of the DPM family

In principle, the DPM family of measures may be implemented with any parser that generates a dependency graph (a single labelled arc for each word, pointing to its head-word).

Our reference implementation[1] uses a state-of-the-art PCFG parser (the first stage of Charniak and Johnson (2005)) to generate a 50-best list of trees for each hypothesis and reference translation, using the parser's default WSJ training parameters. We use Magerman (1995) head-finding to construct dependency trees, using the Charniak parser's head-finding rules, with three modifications: prepositional and complementizer phrases choose nominal and verbal heads respectively and auxiliary verbs are modifiers of main verbs (rather than the converse).

Arc-labels $d \overset{A/B}{\rightarrow} h$ are determined from constituent labels, where the arc label A/B between dependent d and its head h is composed of A (the lowest constituent headed by h and dominating d) and B (the highest constituent headed by d). This strategy is the one adopted in labelled-dependency SPARSEVAL, and it acts as an approximation of the rich semantics generated by (Cahill and others, 2004) or another heavily knowledge-engineered parser, but with much less knowledge-engineering required. The A/B labels are not as descriptive as the LFG semantics, but they have a similar resolution, e.g. the $\overset{S/NP}{\rightarrow}$ arc label usually represents a subject dependent of a sentential verb.

[1]Download the EDPM source code — a collection of Perl scripts and libraries — at `http://ssli.ee.washington.edu/people/jgk/dist/edpm/`, or contact the first author.

Measure s	all-Arabic	all-Chinese	all
TER	0.51	0.19	0.39
BLEU$_4$	-0.40	-0.19	-0.32
EDPM	**-0.61**	**-0.25**	**-0.47**

Table 1: Per-document Pearson's r of Δs with ΔHTER over various measures s, examined for each genre in the corpus, for each language in the corpus, and as a whole.

1.4 Optimal EDPM

Experiments exploring correlation with fluency and adequacy judgments against the LDC Multiple Translation Chinese corpus parts 2 (LDC, 2003) and 4 (LDC, 2006) indicate that the best member of the DPM family uses the full 50-best parses produced by the system, with a "flattening" $\lambda = 0.25$. Using both the partial subtrees dl, lh and the string-only statistics $1g, 2g$ provides an optimal setting of

$$\text{EDPM} = \text{DPM}_{1g,2g,dl,lh}, n = 50, \gamma = 0.25$$

This configuration for EDPM has a correlation $r = 0.240$ against the average fluency and adequacy judgment per-sentence over these corpora.

2 Correlations with HTER

We explore this EDPM variant's utility on another task: predicting the human-targeted translation edit rate (HTER) on the (unsequestered) GALE 2.5 evaluation results, and find that per-document differences across systems in EDPM (ΔEDPM) are better correlated with changes in HTER (ΔHTER) than ΔBLEU$_4$ or ΔTER (table 1).

It is worth mentioning that TER has an advantage in that HTER uses a TER measure to calculate the

post-editing work between the hypothesized translation and the human-targeted reference which could, in principle, bias HTER towards a TER measure. EDPM shares no such advantage. EDPM nevertheless has the best correlation of the three measures in both Arabic and Chinese, as well as over the entire corpus.

3 Conclusion and future work

EDPM demonstrates a promising direction in exploiting syntactic structure for automatic evaluation of machine translation hypotheses. It differs in nature from other proposed evaluation methods in that it does *not* use word-substitution tables or tuned weights (beyond the λ free parameter described above), and yet substantially outperforms BLEU_4 and TER as a predictor of changes to HTER.

In future work, we would like to explore a number of related questions. Exploring a larger list of parse possibilities (by increasing n or scoring from packed parse forests (Huang, 2008)) might allow better diversity of link types and better estimates of their expected counts. Alternatively, we may use a different PCFG parser as another way of exploring the trade-offs between parse-quality, MT quality prediction, and speed. On the other hand, a poor-quality translation that is very difficult to parse may interfere with the quality of the measure; assessing this measure's sensitivity to sentence quality would also be worthwhile. In a different direction, we would like to ask whether other segmentations of the dependency tree are more appropriate than those explored here, following up on an approach suggested in Liu and Gildea (2005), which uses linked words in chains much larger than 2.

From the results correlating EDPM with both HTER and human judgments of fluency and accuracy, EDPM seems to be a superior tool for identifying improvements at the document level and at the sentence level, which is often where parameter tuning takes place. One possible use for this measure, since it is more computationally-costly than BLEU or TER, might be as a late pass evaluation metric in training to select among translation outputs already deemed to be very good.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0741585 and the Defense Advanced Research Projects Agency under Contract No. HR0011-06-C-0023.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic & Extrinsic Eval. Measures for MT*, pp. 65–72.

A. Cahill et al. 2004. Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proc. ACL*, pp. 319–326.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proc. ACL*, pp. 173–180.

L. Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proc. ACL: HLT*, pp. 586–594.

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

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

D. Liu and D. Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proc. ACL Workshop on Intrinsic & Extrinsic Eval. Measures for MT*, pp. 25–32.

D. M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proc. ACL*, pp. 276–283.

K. Owczarzak, J. van Genabith, and A. Way. 2007. Labelled dependencies in machine translation evaluation. In *Proc. 2nd Workshop on Statistical MT*, pp. 104–111.

B. Roark et al. 2006. SParseval: Evaluation metrics for parsing speech. In *Proc. LREC*.

J. P. Turian, L. Shen, and D. I. Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proc. MT Summit IX*.