Mach Translat (2009) 23:169–179
DOI 10.1007/s10590-009-9057-6

Expected dependency pair match: predicting translation quality with expected syntactic structure

Jeremy G. Kahn · Matthew Snover · Mari Ostendorf

Received: 15 May 2009 / Accepted: 10 October 2009 / Published online: 31 October 2009

Abstract Recent efforts to develop new machine translation evaluation methods have tried to account for allowable wording differences either in terms of syntactic structure or synonyms/paraphrases. This paper primarily considers syntactic structure, combining scores from partial syntactic dependency matches with standard local *n*-gram matches using a statistical parser, and taking advantage of N-best parse probabilities. The new scoring metric, expected dependency pair match (EDPM), is shown to outperform BLEU and TER in terms of correlation to human judgments and as a predictor of HTER. Further, we combine the syntactic features of EDPM with the alternative wording features of TERp, showing a benefit to accounting for syntactic structure on top of semantic equivalency features.

Keywords Machine translation evaluation · Syntax · Dependency trees

1 Introduction

A challenge in automatic machine translation (MT) evaluation is accounting for allowable variability: two equally good translations may be quite different in surface form. Currently, the most popular approaches are BLEU (Papineni et al. 2002), based on *n*-gram precision, and Translation Edit Rate (TER), an edit distance (Snover et al.

J. G. Kahn (✉) · M. Ostendorf
University of Washington, Seattle, WA, USA
e-mail: jgk@u.washington.edu

M. Ostendorf
e-mail: ostendor@u.washington.edu

M. Snover
University of Maryland, College Park, MD, USA
e-mail: snover@cs.umd.edu

2006). These measures can only account for variability when given multiple translations, and studies show that they may not accurately track translation quality (Charniak et al. 2003; Callison-Burch 2006).

Alternative measures that incorporate synonym knowledge sources include: METEOR (Banerjee and Lavie 2005), which uses synonym tables and morphological stemming to do progressively more forgiving matching; TER Plus (TERp) (Snover et al. 2009), which is an extension of the previously-mentioned TER that also incorporates synonym sets and stemming, along with automatically-derived paraphrase tables. Other metrics modeling syntactically-local (rather than string-local) word-sequences include: tree-local n-gram precision in various configurations of constituency and dependency trees (Liu and Gildea 2005); and the **d** and **d_var** measures proposed by Owczarzak et al. (2007a,b) that compare relational tuples derived from a lexical functional grammar (LFG) over reference and hypothesis translations.[1] These syntactically-oriented measures require a system for proposing dependency structure over the reference and hypothesis translations. Liu and Gildea (2005) use a PCFG parser with deterministic head-finding, while Owczarzak et al. (2007a) extract the semantic dependency relations from an LFG parser (Cahill et al. 2004). This work extends the dependency-scoring strategies of Owczarzak et al. (2007a), which reported substantial improvement in correlation with human judgment relative to BLEU and TER, by using a publicly-available probabilistic context-free grammar (PCFG) parser and deterministic head-finding rules. In addition, we consider more types of constituents and different score combinations, as well as combination with synonym-type scores.

MT measures are evaluated in a variety of ways. Some (Banerjee and Lavie 2005; Liu and Gildea 2005; Owczarzak et al. 2007a) compare the measure to human judgments of fluency and adequacy. In other work, e.g. Snover et al. (2006), measures are compared to human-targeted TER (HTER), a distance to a human-revised reference that uses wording closer to the MT system choices (keeping the original meaning) that is intended to measure the post-editing work required after translation. In this paper, we explore both kinds of evaluation.

We describe our approach to including syntax in MT evaluation by outlining a family of metrics in Sect. 2 and implementation details in Sect. 3. Section 4 examines the correlation of members of this family with human judgments of fluency and adequacy, using the Owczarzak et al. (2007a) paradigm to provide comparisons and select a best case configuration, expected dependency pair match (EDPM). The EDPM measure is then compared to BLEU and TER in terms of correlation with HTER, exploring language/genre effects in Sect. 5 and combination with TERp's synonym/paraphrase features in Sect. 6. Finally, findings and future work are summarized in Sect. 7.

2 Approach

The specific family of dependency pair match (DPM) measures explored here combines precision and recall scores of various decompositions of a syntactic dependency

[1] Owczarzak et al. (2007a) extend their previous line of research (Owczarzak et al. 2007b) by variably-weighting dependencies and by including synonym matching, two directions not pursued here. Hence, the earlier paper is cited in comparisons.

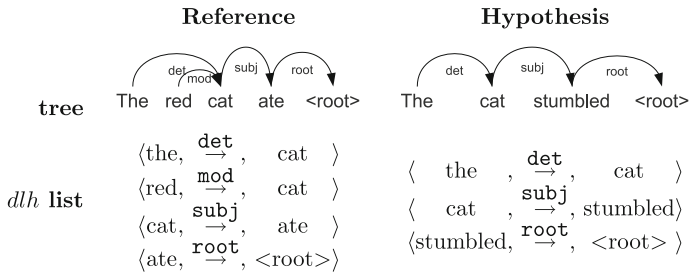

Fig. 1 Example dependency trees and their *dlh* decompositions

Fig. 2 The *dl* and *lh* decompositions of the hypothesis tree in Fig. 1

dl	*lh*
⟨ the , $\overset{\text{det}}{\rightarrow}$⟩	⟨ $\overset{\text{det}}{\rightarrow}$, cat ⟩
⟨ cat , $\overset{\text{subj}}{\rightarrow}$⟩	⟨ $\overset{\text{subj}}{\rightarrow}$, stumbled⟩
⟨stumbled, $\overset{\text{root}}{\rightarrow}$⟩	⟨ $\overset{\text{root}}{\rightarrow}$, <root> ⟩

tree. These measures are extensions of the dependency-pair F measures found in Owczarzak et al. (2007b). Rather than comparing string sequences, as BLEU does with its *n*-gram precision, this approach defers to a parser for an indication of the relevant word tuples associated with meaning—in these implementations, the head on which that word depends. Each sentence (both reference and hypothesis) is converted to a labeled syntactic dependency tree and then relations from each tree are extracted and compared.

We motivate the use of dependencies with actual translations:

Ref: Authorities have also closed southern Basra's airport and seaport.
S1: The authorities also closed the airport and seaport in the southern port of Basra.
S2: Authorities closed the airport and the port of.

A human judged the system 1 result (S1) as equivalent to the reference, but the system 2 (S2) result as having errors. BLEU gives both a similar score (0.199 vs. 0.203). TER scores S2 as better (errors of 0.9 vs. 0.7, respectively), since a simple deletion requires fewer edits that rephrasing. By representing matches of dependencies, we obtain a score for S1 from the new EDPM measure that is higher than that for S2 (0.414 vs. 0.356). The two phrases "southern Basra's airport and seaport" and "the airport and seaport in the southern port of Basra" have more in similarities in terms of dependencies than word order.

The particular relations that are extracted from the dependency tree are referred to here as *decompositions*. Figure 1 illustrates the *dependency-link-head* decomposition of a toy dependency tree into a list of $\langle d, l, h \rangle$ tuples. Some members of the DPM family may apply more than one decomposition; other good examples are the *dl* decomposition, which generates a bag of dependent words with outbound links, and the *lh* decomposition, which generates a bag of inbound link labels, with the head word for each included. Figure 2 shows the *dl* and *lh* decompositions for the same hypothesis tree.

It is worth noting here that the *dlh* and *lh* decompositions (but not the *dl* decomposition) "overweight" the headwords, in that there are *n* elements in the resulting bag, but if a word has no dependents it is found in the resulting bag exactly one time (in the *dlh* case) or not at all (in the *lh* case). Conversely, syntactically "key" words, that are directly modified by many other words in the tree, are included multiple times in the decomposition (once for each inbound link). This "overweighting" leverages syntactic indications of which words are more important to translate correctly (e.g., "Basra" in the example).

A statistical parser provides confidences associated with parses in an *n*-best list, which we use to compute expected counts for each decomposition in both reference and hypothesized translations. The expected counts lead to partial matches (or weighted counts) used in computing precision and recall. This approach addresses both error in the best parse and ambiguity in the translations (reference and hypothesis).

When multiple decomposition types are used together, we may combine these subscores in a variety of ways. Here, we experiment with using two variations of a harmonic mean: computing precision and recall over all decompositions as a group (giving a single precision and recall number) vs. computing precision and recall separately for each decomposition. We distinguish between these using the notation:

$$F[dl, lh] = \mu_h \left(\text{Prec}\left(dl \cup lh\right), \text{Recall}\left(dl \cup lh\right)\right) \quad (1)$$

$$\mu_{PR}[dl, lh] = \mu_h \left(\text{Prec}\left(dl\right), \text{Recall}\left(dl\right), \text{Prec}\left(lh\right), \text{Recall}\left(lh\right)\right) \quad (2)$$

where μ_h represents a harmonic mean. Dependency-based SParseval Roark et al. (2006) and the **d** approach from Owczarzak et al. (2007a) may each be understood as $F[dlh]$, while the latter's **d_var** method may be understood as something close to $F[dl, lh]$. Both the combination methods F and μ_{PR} are "naive" in that they treat each component score as equivalent to the next. When we introduce syntactic/paraphrasing features, we will move to a weighted combination.

3 Parsing and dependency extraction

The family of DPM measures may be implemented with any parser that generates a dependency graph (a single labelled arc for each word, pointing to its head-word). Prior work (Owczarzak et al. 2007a) on related measures has used an LFG parser (Cahill et al. 2004) or an unlabelled dependency tree (Liu and Gildea 2005).

In this work, we use a state-of-the-art PCFG (the first stage of Charniak and Johnson 2005) and context-free head-finding rules (Magerman 1995) to generate an N-best list of dependency trees for each hypothesis and reference translation. We use the parser's default (English) Wall Street Journal training parameters. Head-finding uses the Charniak parser's rules, with three modifications to make the semantic (rather than syntactic) relations more dominant in the dependency tree: prepositional and complementizer phrases choose nominal and verbal heads respectively (rather than functional heads) and auxiliary verbs are modifiers of main verbs (rather than the converse). These changes capture the fact that main verbs are more important for

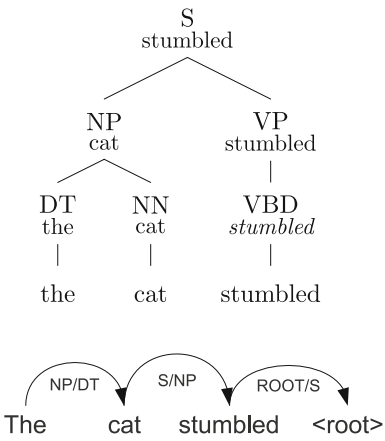

Fig. 3 An example constituent tree (heads of each constituent are listed below the label) and the labelled dependency tree derived from it

adequacy in translation, as illustrated by the functional equivalence of "have also closed" vs. "also closed" in the example in Sect. 2.

Having constructed the dependency tree, we label the arc between dependent d and its head h as A/B when A is the lowest constituent-label headed by h and dominating d and B is the highest constituent label headed by d. For example, in Fig. 3, the S node is the lowest node headed by *stumbled* that dominates *cat*, and the NP node is the highest constituent label headed by *cat*, so the arc linking *cat* to *stumbled* is labelled S/NP. This strategy is very similar to one adopted in the reference implementation of labelled-dependency SPARSEVAL (Roark et al. 2006), and may be considered as a shallow approximation of the rich semantics generated by LFG parsers (Cahill et al. 2004). The A/B labels are not as descriptive as the LFG semantics, but they have a similar resolution, e.g. the S/NP arc label usually represents a subject dependent of a sentential verb.

For the cases where we have N-best parse hypotheses, we use the associated parse probabilities (or confidences) to compute expected counts. The sentence will then be represented with more tuples, corresponding to alternative analyses. For example, if the N-best parses include two different roles for dependent "Basra", then two different *dl* tuples are included, each with the weighted count that is the sum of the confidences of all parses having the respective role.[2] The parse confidence $\tilde{p}$ is normalized so that the N-best confidences sum to one. Because the parser is overconfident, we explore a flattened estimate: $\tilde{p}(k) = \frac{p(k)^\gamma}{\sum_i p(i)^\gamma}$, where k, i index the parse and γ is a free parameter.

4 Correlation with human judgments of fluency and adequacy

We explore various configurations of the DPM by assessing the results against a corpus of human judgments of fluency and adequacy, specifically the LDC Multiple

[2] The use of expectations with N-best parses is different from **d_50** and **d_50_pm** in Owczarzak et al. (2007a) in that the latter uses the best-matching pair of trees rather than an aggregate over the tree sets and they do not use parse confidences.

Translation Chinese corpus parts 2 (LDC 2003) and 4 (LDC 2006), which are composed of translations of written Chinese news stories. These corpora include multiple human judgments of fluency and adequacy for each sentence (assigned on a five-point scale), with each judgment using a different human judge and a different reference translation. For a rough[3] comparison with Owczarzak et al. (2007a), we treat each judgment as a separate segment, which yields 16,815 tuples of ⟨hypothesis, reference, fluency, adequacy⟩. We compute per-segment correlations.[4] The baselines for comparison are case-sensitive BLEU (4-grams, with add-one smoothing) and TER.

The specific dimensions of DPM explored include:

Decompositions. We compute precision and recall of:
- *dlh* ⟨Dependent, arc Label, Head⟩—full triple
- *dl* ⟨Dependent, arc Label⟩—marks how the word fits into its syntactic context (what it modifies)
- *lh* ⟨arc Label, Head⟩—implicitly marks how key the word is to the sentence
- *dh* ⟨Dependent, Head⟩—drops syntactic-role information.
- *1g,2g* —simple measures of unigram (bigram) precision and recall.

Parser variations. When using more than one parse, we explore:

Size of n-best list. 50 (as in Owczarzak et al. 2007a)

Parse confidence. The distribution flattening parameter is varied from $\gamma = 0$ (uniform distribution) to $\gamma = 1$ (no flattening).

Score combination. Global F vs. component harmonic mean μ_{PR}.

Considering only the 1-best parse, we compare DPM with different decompositions to the baseline measures. Table 1 shows that all decompositions except $[dlh]$ have a better per-segment correlation with the fluency/adequacy scores than TER or BLEU_4. Dependencies $[dl, lh]$ and string-local n-grams $[1g, 2g]$ give similar results, but the combination gives further improvement. The results also confirm, with a PCFG, what Owczarzak et al. (2007a) found with an LFG parser: that partial-dependency matches are better correlated with human judgments than full-dependency links. Including progressively larger chunks of the dependency graph with $F[1g, dl, dlh]$, inspired by the BLEU_k idea of progressively larger n-grams, did not give an improvement over $[dl, lh]$. F gives substantially better results than μ_{PR}, which is never better than BLEU for these decompositions.

Table 2 shows the impact of using N-best parses for different decompositions. For the $n = 50$ cases, we set $\gamma = 0$ to assign uniform probabilities, which was slightly better than $\gamma = 1$. While not all of these differences are significant, there is a consistent trend of correlation r improving with 50 vs. 1 parse. Tuning experiments find that increasing γ to 0.25 can increase the r reported here for $F[1g, 2g, dl, lh]$ (but insignificantly).

[3] Our segment count differs slightly from Owczarzak et al. (2007a) for the same corpus: 16,807 vs. 16,815. As a result, the baseline per-segment correlations differ slightly (BLEU_4 is higher here, while TER here is lower), but the trends in gains over those baselines are very similar.

[4] The use of the same hypothesis translations in multiple comparisons in the Multiple Translation Corpus means that scored segments are not strictly independent, but for methodological comparison with prior work, this strategy is preserved.

Table 1 Per-segment correlation with human fluency/adequacy judgments of baselines and different decompositions

Metric	$\|r\|$
$F[1g, 2g, dl, lh]$	0.237
$F[1g, 2g]$	0.227
$F[dl, lh]$	0.226
BLEU_4	0.218
$F[dlh]$	0.185
TER	0.173

Table 2 Per-segment correlation with human fluency/adequacy judgments comparing $n = 1$ vs. 50 parses for different decompositions

Metric	n	r
$F[1g, 2g, dl, lh]$	50	0.239
$F[1g, 2g, dl, lh]$	1	0.237
$F[1g, dl, lh]$	50	0.237
$F[1g, dl, lh]$	1	0.234
$F[dl, lh]$	50	0.234
$F[dl, lh]$	1	0.226

In summary, exploring a number of variants of the DPM metric against an average fluency/adequacy judgment leads to a best-case of:

$$\text{EDPM} = F[1g, 2g, dl, lh], n = 50, \gamma = 0.25$$

We use this configuration in experiments assessing correlations with HTER.

5 Correlating EDPM with HTER

In this section, we compare EDPM to baseline metrics in terms of document- and segment-level correlation with HTER scores using the GALE 2.5 translation corpus. The corpus includes system translations into English from three sites, all of which use system combination to integrate results from several systems, some phrase-based and some that use syntax on either the source or target side. No system provided system-generated parses. The source data comprises Arabic and Chinese in four genres: `bc` (broadcast conversation), `bn` (broadcast news), `nw` (newswire), and `wb` (web text), with corpus sizes shown in Table 3. The corpus includes one English reference translation (LDC 2008) for each sentence and a system translation for each of the three systems. Additionally, each of the system translations has a corresponding human-targeted reference aligned at the sentence level, so we have available the HTER score at both the sentence and document level.

HTER and automatic scores all degrade on average for more difficult sentences. Since there are multiple system translations in this corpus, it is possible to roughly factor out this source of variability by correlating mean normalized scores,[5]

[5] Previous work (Kahn et al. 2008) reported HTER correlations against pairwise differences among translations derived from the same source to factor out sentence difficulty, but this violates independence assumptions used in the Pearson's r tests.

Table 3 Corpus statistics for the GALE 2.5 translation corpus

	Arabic		Chinese		Total	
	doc	sent	doc	sent	doc	sent
bc	59	750	56	1,061	115	1,811
bn	63	666	63	620	126	1,286
nw	68	494	70	440	138	934
wb	69	683	68	588	137	1,271
Total	259	2,593	257	2,709	516	5,302

Table 4 Per-document correlations of $\overline{\text{EDPM}}$ and others to $\overline{\text{HTER}}$, by genre and by source language

r vs. $\overline{\text{HTER}}$	bc	bn	nw	wb	All Arabic	All Chinese	All
$\overline{\text{TER}}$	0.59	**0.35**	**0.47**	*0.17*	**0.54**	**0.32**	0.44
$\overline{\text{BLEU}}$	−0.42	**−0.32**	**−0.46**	**−0.27**	−0.42	**−0.33**	−0.37
$\overline{\text{EDPM}}$	**−0.69**	**−0.39**	**−0.47**	**−0.27**	**−0.60**	**−0.39**	**−0.50**

Bold numbers are within 95% significance of the best per column; italics indicate that the sign of the r value has less than 95% confidence

Table 5 Per-sentence, length-weighted correlations of $\overline{\text{EDPM}}$ and others to $\overline{\text{HTER}}$, by genre and by source language

r vs. $\overline{\text{HTER}}$	bc	bn	nw	wb	All Arabic	All Chinese	All
$\overline{\text{TER}}$	**0.44**	**0.29**	**0.33**	0.25	**0.44**	0.25	**0.36**
$\overline{\text{BLEU}}$	−0.31	−0.24	−0.29	−0.25	−0.31	−0.24	−0.28
$\overline{\text{EDPM}}$	**−0.46**	**−0.31**	**−0.34**	**−0.30**	**−0.44**	**−0.30**	**−0.37**

Bold numbers indicate significance as above

$\overline{m}(t_i) = m(t_i) - \frac{1}{I}\sum_{j=1}^{I} m(t_j)$ where m can be HTER, TER, BLEU or EDPM and t_i represents the ith translation of segment t. Mean-removal ensures that the reported correlations are among differences in the translations rather than among differences in the underlying segments.

In Table 4, we show per-document Pearson's r between $\overline{\text{EDPM}}$ and $\overline{\text{HTER}}$, as well as the $\overline{\text{TER}}$ and $\overline{\text{BLEU}_4}$ baselines. EDPM has the highest correlation in each of the subcorpora created; by division of genre or by source language, as well as on the corpus as a whole. In structured data (bn and nw), these differences are not always significant, but in the unstructured domains (wb and bc), EDPM is always significantly better than at least one of the comparison baselines.

Table 5 presents per-sentence correlations based on scores normalized by sentence length in order to get a per-word measure which reduces variance across sentences. (Even with length weighting, the r values are smaller in magnitude due to the higher variability at the sentence level.) EDPM again has the largest correlation in each category, but TER has r values within 95% confidence of EDPM scores on nearly every breakdown.

6 Combining syntax, edit and semantic knowledge sources

While our results show that EDPM is as good or better than other measures, the correlation is still low, which is consistent with the example in Sect. 2, where the EDPM score is much less than 1 for the good translation. For that reason, we investigated combining the alternative wording features of TERp with the EDPM syntactic features.

The TERp tools (Snover et al. 2009) provide an optimizer for weighting multiple simple subscores. The TERp optimizer performs a hill-climbing search, with randomized restarts, to maximize the correlation of a linear combination of the subscores with a set of human judgments. Within the TERp framework the subscores are the counts of the various edit types normalized for the length of the reference, where the counts are determined after aligning the MT output to the reference using default edit costs.

The experiments here leverage the TERp optimizer but extend the set of subscores by including the syntactic and n-gram overlap features (modified to reflect false and missed detection rates for the TERp format rather than precision and recall). The subscores explored include:

E: the 8 fully syntactic subscores from the EDPM family, including false/miss error rates for the dl, lh, dlh, and dh decompositions.
N: the 4 n-gram subscores from the DPM family; specifically, error rates for the $1g$ and $2g$ decompositions.
T: the 11 subscores from TERp, which include matches, insertions, deletions, substitutions, shifts, synonym and stem matches, and four paraphrase edit scores.

For these experiments, we again use the GALE 2.5 data, but with 2-fold cross-validation in order to have independent tuning and test data. Documents are partitioned randomly, such that each subset has the same document distribution across source-language and genre. The objective is length-normalized per-sentence correlation with HTER, using mean-removed scores as before. In Fig. 4, we plot the Pearson's r (with 95% confidence interval) for the results on the two test sets combined, after linearly normalizing the predicted scores to account for magnitude differences in the learned weight vectors. The baseline scores, which involve no tuning, are not normalized.

The figure shows that TER and EDPM are significantly more correlated with HTER than BLEU, consistent with the overall results of the previous section. The N + E combination outperforms E alone (i.e. it is helpful to use both n-gram and dependency overlap) but gives lower performance than EDPM because of the particular combination technique. Both findings are consistent with the fluency/adequacy experiments. The TERp features, which account for synonym/paraphrase differences, have much higher correlation with HTER than the syntactic E + N subscores. However, a significant additional improvement is obtained by adding syntactic features to TERp (T + E). Adding the n-gram features to TERp (T + N) gives almost as much improvement, probably because most dependencies are local. There is no further gain from using all three subscore types.

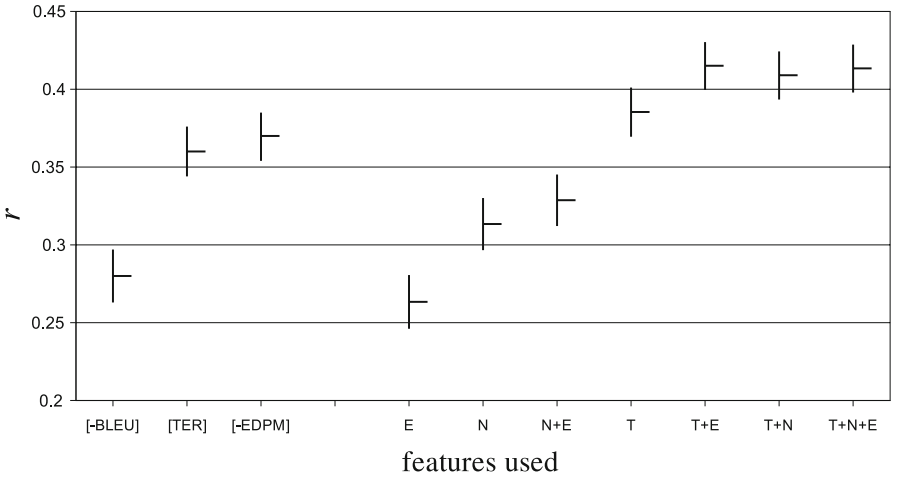

Fig. 4 Pearson's *r* for various feature tunings, with 95% confidence intervals

7 Conclusion

In summary, we explore a family of dependency pair match measures. Through a corpus of human fluency and adequacy judgments, we settle on EDPM, a member of that family with promising predictive power. We find that EDPM is superior to BLEU and TER in terms of correlation with human judgments and as a per-document and per-sentence predictor of mean-normalized HTER. We also experiment with including syntactic and synonym/paraphrase features in a TERp-style linear combination, and find that the combination improves correlation with HTER over either method alone.

One difference with respect to the work of Owczarzak et al. (2007a) is the use of a PCFG vs. an LFG parser. The PCFG has the advantage that it is publicly available and easily adaptable to new domains. However, the performance varies depending on the amount of labeled data for the domain, which raises the question of how sensitive EDPM and related measures are to parser quality.

A limitation of this method for MT system tuning is the computational cost of parsing compared to word-based measures such as BLEU or TER. Two alternative low-cost use scenarios include late-pass evaluation, for choosing between different system architectures, or system diagnostics, looking at relative quality of these component scores compared to those of an alternative configuration.

Acknowledgements This material is based on work supported by the National Science Foundation under Grant No. 0741585 and the Defense Advanced Research Projects Agency under Contract Nos. HR0011-06-C-0022 and HR0011-06-C-0023. Any opinions, findings, conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization, pp 65–72

Cahill A, Burke M, O'Donovan R, Van Genabith J, Way A (2004) Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In: Proceedings of the ACL, pp 319–326

Callison-Burch C (2006) Re-evaluating the role of BLEU in machine translation research. In: Proceedings of the EACL, pp 249–256

Charniak E, Johnson M (2005) Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the ACL, pp 173–180

Charniak E, Knight K, Yamada K (2003) Syntax-based language models for statistical machine translation. In: Proceedings of the MT Summit IX

Kahn JG, Roark B, Ostendorf M (2008) Automatic syntactic MT evaluation with expected dependency pair match. In: MetricsMATR: NIST metrics for machine translation challenge

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

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

LDC (2008) GALE phase 2 + retest evaluation references. Catalog number LDC2008E11

Liu D, Gildea D (2005) Syntactic features for evaluation of machine translation. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization, pp 25–32

Magerman DM (1995) Statistical decision-tree models for parsing. In: Proceedings of the ACL, pp 276–283

Owczarzak K, Genabith Jvan , Way A (2007a) Evaluating machine translation with LFG dependencies. Mach Transl 21(2):95–119

Owczarzak K, van Genabith J, Way A (2007b) Labelled dependencies in machine translation evaluation. In: Proceedings of the second workshop on statistical machine translation, pp 104–111

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL, pp 311–318

Roark B, Harper M, Charniak E, Dorr B, Johnson M, Kahn JG, Liu Y, Ostendorf M, Hale J, Krasnyanskaya A, Lease M, Shafran I, Snover M, Stewart R, Yung L (2006) SParseval: evaluation metrics for parsing speech. In: Proceedings of the LREC

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the AMTA

Snover M, Madnani N, Dorr B, Schwartz R (2009) Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In: Proceedings of the workshop on statistical machine translation at EACL