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

Abstract

Previous work on dependency-based machine translation evaluation has shown that an LFG-based F-measure of labeled partial dependencies over an n-best list has improved correlation with human judgments of fluency and adequacy as compared to BLEU and TER. Inspired by SParseval, we demonstrate that a statistical syntactic parser based on PCFGs may be used in place of the LFG dependencies, and we explore variations on other aspects of the dependency-matching method. We find that including $n > 1$ parses helps, especially if one incorporates probabilistic weights from the parser. Additionally, we find that matching simple words (along with partial-dependencies) further improves correlations with human judgments.

From these results, we design a new scoring metric "Expected Dependency Pair Match" (EDPM), and demonstrate that ΔEDPM is superior to ΔBLEU and ΔTER as a per-document and per-sentence predictor of ΔHTER.

Introduction

This paper presents the Expected Dependency Pair Match metric and algorithm. This algorithm uses a popular PCFG parser to extract expected counts of dependency structures from hypothesis and reference translations and score them.

The reference implementation of this algorithm is available from the web[1]. The metric itself is implemented as a collection of Perl scripts and libraries and is therefore portable to almost any modern platform, while the parser required (the first stage of (Charniak and Johnson, 2005)) is available for architectures that support the gcc compiler. Please see the distribution for more details on use.

The remainder of this work describes the motivation for this algorithm, and presents a series of experiments exploring its correlation with human judgments and its correlation with the human-targeted HTER metric.

1 Background & Motivation

Machine translation (MT) evaluation is a challenge for research because the space of good translations is large, and two equally good translations may appear to be quite different at first glance. The challenges of choosing among translations are compounded when this evaluation is done automatically. Human evaluation, however, is both time-consuming and difficult, so research has turned increasingly towards automatic measures of translation quality, usually by comparing the system translation to one or more reference (human) translations. Automatic measures of this kind (e.g. BLEU (Papineni et al., 2002) and TER (Snover and others, 2006)) not only provide a well-defined evaluation standard but are also required for training on error criteria, e.g. with minimum error rate training (Och, 2003).

The most popular evaluation measures (BLEU and related measure NIST (Doddington, 2002)) are based on n-gram precision. More recent research has found that these measures do not always accurately track translation quality both empirically (Charniak et al., 2003) and theoretically (Callison-Burch, 2006).

One direction to look for improving metrics is to try to model acceptable variation, whether word-choice (e.g. METEOR (Banerjee and Lavie, 2005), which does progressively more forgiving word matching), by weighting adjacent matches more than non-local matches (e.g. GTM (Turian et

[1] Download the EDPM source code at http://ssli.ee.washington.edu/people/jgk/dist/edpm/ or contact the first author.

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.

These measures have been evaluated in a number of ways. Some (Banerjee and Lavie, 2005; Liu and Gildea, 2005; Owczarzak et al., 2007) have evaluated their success by comparing the measure to human judgments of fluency and adequacy. In other work, e.g. Snover and others (2006), measures are evaluated by comparison to 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 pursue both kinds of evaluation.

Keeping to the syntactic approach, the work here 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)).

This paper presents a new metric EDPM that extends the dependency-scoring strategies of Owczarzak et al. (2007) but do so with a widely used and publically available PCFG parser and deterministic head-finding rules instead of an LFG system. In addition, EDPM incorporates word-level matching and weighted multiple parse alternatives for improved performance.

The remainder of this work is as follows. In section 2, we define a family of measures DPM that include SPARSEVAL and **d/d_var** measures and describe our strategy for extracting labelled dependencies. In section 3, we explore a selection of variants of that family with human judgments over the Multiple Translation Chinese corpus (LDC, 2003; LDC, 2006). We explore questions of dependency-graph decomposition, using multiple parses, and parse-confidence, and we select a member of that family as a new best measure EDPM that we believe to be a superior representative of the DPM family. In section 4, we explore EDPM's ability to predict changes in a human-generated score (HTER) over a corpus of hypothesis translations in the GALE 2.5 (DARPA, 2008) evaluation. In section 5, we discuss future work and conclude.

2 Definition of DPM family of metrics

In this work, we define a family of measures Dependency Pair Match (DPM) that is composed of extensions of the methods described in Owczarzak et al. (2007). DPM is defined as the F measure (harmonic mean of precision and recall) 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$\rangle$. We describe different variants in this family below, and compare their effectiveness in section 3.

2.1 DPM variations in subtree extraction

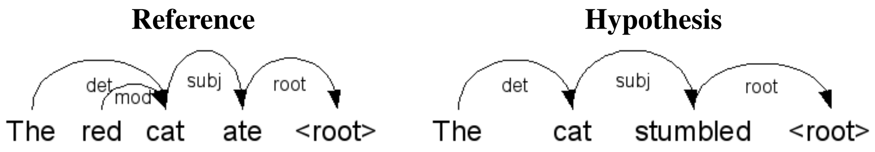
Different members of the DPM family of metrics may extract different subtrees. We denote the set of extracted tree-components with a trailing subscript: DPM_{dlh} extracts all $\langle$Dependent, arc-Label, Head$\rangle$ subtree tuples, which is equivalent (modulo dependency-extraction methods) to labeled SPARSEVAL (Roark and others, 2006) and the Owczarzak et al. (2007) **d** measure.

$\text{DPM}_{dl,lh}$, by contrast, extracts all the subtrees $\langle$Dependent, arc-Label$\rangle$ and $\langle$arc-Label, Head$\rangle$, which is equivalent to the Owczarzak et al. (2007) **d_var** method (again, modulo the dependency-extraction method). The hypothesis tree in figure 1 produces the six items below for scoring with $\text{DPM}_{dl,lh}$:

dl	lh
$\langle$ the , $\xrightarrow{\texttt{det}}$ $\rangle$	$\langle$ $\xrightarrow{\texttt{det}}$, cat $\rangle$
$\langle$ cat , $\xrightarrow{\texttt{subj}}$ $\rangle$	$\langle$ $\xrightarrow{\texttt{subj}}$, stumbled $\rangle$
$\langle$stumbled, $\xrightarrow{\texttt{root}}$ $\rangle$	$\langle$ $\xrightarrow{\texttt{root}}$, <root> $\rangle$

For the hypothesis and reference trees in figure 1,

dependency tree



dlh **list**

Reference	Hypothesis
$\langle \text{the}, \overset{\texttt{det}}{\rightarrow}, \text{cat} \rangle$	$\langle \text{the}, \overset{\texttt{det}}{\rightarrow}, \text{cat} \rangle$
$\langle \text{red}, \overset{\texttt{mod}}{\rightarrow}, \text{cat} \rangle$	$\langle \text{cat}, \overset{\texttt{subj}}{\rightarrow}, \text{stumbled} \rangle$
$\langle \text{cat}, \overset{\texttt{subj}}{\rightarrow}, \text{ate} \rangle$	$\langle \text{stumbled}, \overset{\texttt{root}}{\rightarrow}, \text{<root>} \rangle$
$\langle \text{ate}, \overset{\texttt{root}}{\rightarrow}, \text{<root>} \rangle$	

Precision$_{dlh}$ is $\frac{1}{3}$ and Recall$_{dlh}$ is $\frac{1}{4}$, yielding $\mathrm{DPM}_{dlh} = F = \frac{2}{7} = 0.286$

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

Precision$_{dl,lh}$ is $\frac{3}{6}$ and Recall$_{dl,lh}$ is $\frac{3}{8}$, giving a $\mathrm{DPM}_{dl,lh}$ of $\frac{3}{7} = 0.429$.

2.2 DPM variations using n-best lists and expected counts

Since the dependency structures of the hypothesis and reference text are hidden, we may also be interested in exploring alternative dependency structures predicted by the parser, to cope with genuine ambiguity (in both the translation 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 consider three options: uniform, the parser probabilities, and a flattened version of 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 the fact that the parser tends to be over-confident. In all cases, the probabilities are normalized to sum to one over the the n-best list, where the maximum n in this work is 50. The uniform distribution ($\gamma = 0$) is intended to be equivalent to the Owczarzak et al. (2007) **d_50** and **d_50_var** measures.[2] We note in table 1 those measures in the DPM family that correspond to Owczarzak et al. (2007) measures.

DPM Subgraph	n	γ	Owczarzak et al. (2007) equivalence
dlh	1	—	**d**
dlh	50	0	**d_50**
dl, lh	1	—	**d_var**
dl, lh	50	0	**d_50_var**

Table 1: **Correspondences between the** DPM **family of measures and Owczarzak et al. (2007) d_* measures.** Differences between dependency extraction methods are ignored for these equivalencies.

2.3 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).

In this work, we use 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. We use the parser's default Wall Street Journal training parameters. We use Magerman (1995) style context-free 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 (rather than functional heads) and auxiliary verbs are modifiers of main verbs (rather than the converse).

Given the dependency arcs, we determine the arc-

[2] Since Owczarzak et al. (2007) report no use of parse weights, **d_50** and **d_50_var** may be using a sum of counts over the 50-best list rather than expected-counts over a uniform distribution. These two approaches are equivalent — so long as the n-best list is always the same length for hypothesis and reference. In our implementation (section 2.3), the n-best list does not always reach 50 candidate parses on short sentences, so the expectation matches our intent better than a sum of counts over the n-best.

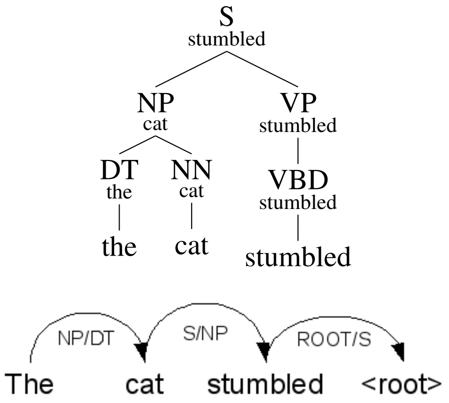


Figure 2: An example constituent tree (heads of each constituent are listed small below the label) and the labelled dependency tree derived from it using the strategy described in section 2.3.

labels $d \overset{A/B}{\rightarrow} h$ from the constituent labels, where the arc label A/B between dependent d and its head h is composed of A (the lowest constituent-label headed by h and dominating d) and B (the highest constituent label headed by d). For example, in figure 2, the arc between *cat* and *stumbled* is labelled $\overset{S/NP}{\rightarrow}$ because 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*.

This strategy is very similar to one adopted in the reference implementation of labelled-dependency SParseval, and may be considered 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.

Measure	r
$\mathrm{DPM}_{dl,lh}$ ($\sim$**d_var**)	0.226
1+BLEU$_4$	0.218
DPM_{dlh} ($\sim$**d**)	0.185
TER	-0.173

Table 2: Correlation of various measures with the average of fluency and adequacy over the sentences in the MTC corpus. These DPM results use only the one-best parse ($n = 1$).

3 Correlation with human judgments of fluency & adequacy

To select a good member of the DPM family, we explore the correlation of these measures against a corpus of human judgments of fluency and adequacy.

3.1 Corpus

For these experiments, we use LDC Multiple Translation Chinese corpus parts 2 (LDC, 2003) and 4 (LDC, 2006). These corpora include multiple human judgments of fluency and adequacy for each sentence, with each judgment using a different human judge and a different reference translation. For a rough[3] comparison with Owczarzak et al. (2007), we treat each judgment as a separate segment and use Pearson's r as a measure of correlation. This treatment of this corpus yields 16,815 tuples of ⟨hypothesis, reference, fluency, adequacy⟩. In these experiments, we extend this tuple with automatic scores derived from ⟨hypothesis, reference⟩ and examine the correlations[4] between those automatic scores and the arithmetic mean of the fluency and adequacy measures. Because BLEU is known to drop to zero too easily on short sentences, we perform add-one smoothing for BLEU's component precisions (we denote this 1+BLEU$_k$).

3.2 Utility of alternative dependency extraction

Our research suggests that the dependency extraction strategy from section 2.3 works in place of the richer semantics from an LFG parser. The results in table 2, which use only the top parse for each sentence, are very similar to those reported in Owczarzak et al. (2007), in that the syntactic mea-

[3] Our segment count differs from Owczarzak et al. (2007), who use 16,800 segments from the same corpus, and our baseline metric correlations differ from theirs (we find that 1+BLEU$_4$ does better and TER worse than reported there). The results presented here are thus not directly comparable with that paper, though we also demonstrate similar gains over those baselines in essentially the same corpus (section 3.2).

[4] The independence of each of these segments is questionable, since the same hypothesis translations are used in multiple items, but for the sake of methodological comparison with prior work, this strategy is preserved. As Turian et al. (2003) points out, ranking correlations might be a better choice than linear correlation, but treating each segment as independent forbids this.

Measure	r
$\mathrm{DPM}_{1g,2g,dl,lh}$	0.237
$\mathrm{DPM}_{1g,dl,lh}$	0.234
$\mathrm{DPM}_{1g,2g}$($\equiv$ bag-of-ngrams(2))	0.227
$\mathrm{DPM}_{dl,lh}$	0.226
$\mathrm{DPM}_{1g,dl,dlh}$	0.227
1+BLEU$_4$	0.218
DPM_{dlh}	0.185
TER	-0.173

Table 3: As in table 2, but with alternative dependency-graph constituents to compute the F measure. Again, $n = 1$ for all DPM correlations.

sure $\mathrm{DPM}_{dl,lh}$ (∼**d_var**) is much-better correlated with human judgment than the non-syntactic measures 1+BLEU$_4$ and TER.

3.3 Alternative dependency sub-graphs

The DPM family allows us to easily explore alternative sub-graphs of the dependency graph, and we find that we achieve small improvements in correlation with human judgments by including the unigram ($1g$) and bigram ($2g$) in the dependency-tree decomposition, ignoring the dependency arc. Owczarzak et al. (2007) found that their original proposal, **d**, scoring full dependent-arc-head triples, was not as well-correlated with human judgment as **d_var**, which examined only dependent-arc and arc-head tuples. Table 2 confirms this as well for a statistical constituent parser with simple dependency-extraction. In table 3, we extend this search to consider whether it is useful to include other subgraphs of the dependency tree into the bag of tree-fragments to be scored.

Table 3 shows that we can combine the benefits of string-local n-grams ($\mathrm{DPM}_{1g,2g}$) with the benefits of dependency information ($\mathrm{DPM}_{dl,lh}$) for a further improved correlation with human judgment, with the best correlation in $\mathrm{DPM}_{1g,2g,dl,lh}$. Including progressively larger chunks of the dependency graph (as in $\mathrm{DPM}_{1g,dl,dlh}$, which is inspired by the BLEU$_k$ idea of progressively larger n-grams) does not seem to be an improvement over $\mathrm{DPM}_{dl,lh}$.

Measure	**parameters**	r
$\mathrm{DPM}_{1g,2g,dl,lh}$	$\gamma = 0, n = 50$	0.239
$\mathrm{DPM}_{1g,2g,dl,lh}$	$n = 1$	0.237
$\mathrm{DPM}_{1g,dl,lh}$	$\gamma = 0, n = 50$	0.237
$\mathrm{DPM}_{1g,dl,lh}$	$n = 1$	0.234
$\mathrm{DPM}_{dl,lh}$ (∼ **d_50_var**)	$\gamma = 0, n = 50$	0.234
$\mathrm{DPM}_{dl,lh}$ (∼ **d_var**)	$n = 1$	0.226

Table 4: As in table 3, but considering variants of the best DPM measures uniform probability distribution over multiple parses ($\gamma = 0, n = 50$).

$\mathrm{DPM}_{1g,2g,dl,lh}$			$\mathrm{DPM}_{dl,lh}$		
n	γ	r	n	γ	r
50	0.25	0.240	50	0.25	0.234
50	0.5	0.240	50	0.5	0.234
50	0.75	0.240	50	0	0.234
50	1	0.239	50	0.75	0.233
50	0	0.239	50	1	0.232
1	—	0.237	1	—	0.226

Table 5: As in table 4, but considering various values of γ and n for two different DPM sub-graph lists (dl, lh and $1g, 2g, dl, lh$).

3.4 Using parse n-best lists

We explore the use of multiple parses in table 4, which presents DPM variants with $\gamma = 0$ of the most successful DPM sub-graph lists shown in previous tables. We use $\gamma = 0$ (uniform probability over the n-best list) to compare as closely as possible to Owczarzak et al. (2007), which uses a parser with ranks but no weights.

We find that using multiple parses with a uniform distribution improves correlations further, although the improvement from varying the dependency-tree sub-graphs is not as large for $n = 50$ variants of DPM as for $n = 1$ variants.

3.5 Including parse confidence

Since the parser in our implementation provides a confidence in each parse, we explore the use of that confidence with the γ free parameter. Table 5 explores various "flattenings" (values of γ) of the parse confidence in the DPM measure. $\gamma = 1$ is not always the best, suggesting that the parse probabilities $p(\text{tree}|\text{words})$ are overconfident. We find that $\gamma = 0.25$ is generally the best flattening of the parse confidence for the variants of DPM that we have tested. The differences are small, but the trends are consistent across the variants.

3.6 Summary

In this section, we have presented experiments exploring a number of parameters to the DPM measure. The experiments suggest a best-case variant EDPM, where we set:

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

in which we choose a $1g, 2g, dl, lh$ sub-graph decomposition based on the improvements from better sub-graphs (table 3), multiple parses ($n = 50$) based on table 4, and $\gamma = 0.25$, hinted at by table 5. We use these EDPM parameter-settings in the experiments exploring correlations with HTER (below).

4 Correlations with HTER

Having chosen EDPM as a good candidate member of the DPM family, we explore its utility on another task: predicting the human-targeted translation edit rate (HTER) on the (unsequestered) GALE 2.5 evaluation results.

4.1 Corpus

The GALE 2.5 translation corpus is made up of system translations into English from three sites. The three sites all use system combination to integrate results from multiple systems, some of which are phrase-based and some which may use syntax on either the source or target side. No system provided system-generated parses. The corpus being translated comes from Arabic and Chinese in four genres: bc (broadcast conversation), bn (broadcast news), nw (newswire), and wb (web text), with corpus sizes shown in table 6. The corpus includes one English reference translation r_i (LDC, 2008) for each sentence i and a system translation $t_{i,z}$ for each of the three systems z. Additionally, each translation of each segment i has a corresponding human-targeted reference aligned at the sentence level, so we have available the HTER score of each segment $s_{\mathrm{HTER}}(t_{i,z})$ at both the sentence and document level.

4.2 Prediction of ΔHTER per-document

	Arabic		Chinese		Total	
	doc	sent	doc	sent	doc	sent
bc	59	750	56	1061	115	1811
bn	63	666	63	620	126	1286
nw	68	494	70	440	138	934
wb	69	683	68	588	137	1271
Total	259	2593	257	2709	516	5302

Table 6: Corpus statistics for the GALE 2.5 translation corpus.

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 7: Per-doc 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.

We evaluate our new measure EDPM by testing its prediction of improvements in HTER over different translations of a document. We choose one of the systems arbitrarily as baseline b and define

$$\Delta s(i, b, z) \quad = \quad s(t_{i,z}) - s(t_{i,b}) \qquad (1)$$

as a measure of change that system z provides over segment i with respect to system b (by providing translation $t_{i,z}$ instead of $t_{i,b}$).

We would like an automatic measure s that provides a good correlation[5] r between Δs and ΔHTER. For each of the 516 documents[6] in the corpus described in table 6, and for each of the three pairs of systems, we generate the Δs scores, ordering the document translations such that $\Delta\mathrm{HTER} \geq 0$ for each document pair.

Table 7 shows the per-document correlations with ΔHTER over this set of data, broken out into per-genre and per-language correlations.

We compare EDPM's correlation with HTER against both that of TER and that of BLEU$_4$, two very popular automatic measures. It is worth mentioning that TER has an advantage in that HTER uses a TER measure to calculate the post-editing

[5] Here we prefer Pearson's linear correlation r to a rank correlation, because we are trying to predict one (roughly) linear score with another.

[6] Using per-document comparison avoids the problems of per-sentence comparisons, e.g. that BLEU falls to zero too easily on a sentence. Accordingly, we use BLEU rather than 1+BLEU, though differences are small.

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.

In table 8, we break out the results into individual language×genre subcorpora for further comparison. We find that EDPM is the best measure in nearly all subcorpora (Arabic bn standing as an exception).

4.3 Prediction of ΔHTER per-sentence

Like the per-document correlations, at the sentence level we are interested in identifying the predictive power of changes in score s with respect to changes in the gold-standard score (here, HTER). In table 9, we present per-sentence correlations of ΔHTER improvements weighted by sentence length[7], also broken out by individual language×genre subcorpora. In contrast to the per-document scores, we use 1+BLEU$_4$ rather than BLEU$_4$ as the representative of the BLEU metric.

Table 9's per-sentence results are largely similar to the per-document analysis (table 8), although both the absolute and relative differences are smaller.

5 Conclusion and future work

In this paper, we described DPM, a family of metrics for evaluating machine translation quality using a labelled dependency tree. Using the Multiple Translation Chinese corpus, we selected a member of that family EDPM and found a good value for free parameter γ, taking advantage of parse structure and parser probabilities. We then tested the EDPM measure against the GALE 2.5 translation corpus and evaluated its ability to predict ΔHTER (the change in HTER) on a per-document and per-sentence level. We have shown EDPM to be superior to both BLEU$_4$ and TER in most cases and for most of the subcorpora available to us.

EDPM has the same advantages as the Owczarzak et al. (2007) **d_*** measures, but is implemented in a more portable way. The method described in this paper may be implemented with any PCFG-based parser — a treebanked corpus for training that parser should be sufficient. Both methods require substantially more time to run than BLEU or TER, because parses of both the hypothesis and the reference are required.

The correlations with BLEU and TER are both worth closer examination. Many of the systems used in the comparison were optimized on BLEU, which raises the possibility that current MT systems are over-fitted to BLEU, reducing BLEU's utility as a predictor of quality translation. Conversely, TER's relationship to HTER makes one suspicious about its good correlation; it might be worth testing against another human-directed measure as well.

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

[7] When calculating per-sentence correlations, we want to scale differences in $\Delta s(i, b, z)$ by the length of reference r_i, since an improvement of $x\%$ on score s in a long sentence is ordinarily understood to be worth more than the same $x\%$ on a short sentence.

	Arabic				Chinese			
Measure s	`bc`	`bn`	`nw`	`wb`	`bc`	`bn`	`nw`	`wb`
TER	0.59	**0.24**	0.22	0.26	0.06	0.13	0.35	0.14
BLEU$_4$	-0.50	-0.10	**-0.30**	-0.31	-0.01	-0.22	**-0.36**	-0.07
EDPM	**-0.80**	-0.10	**-0.31**	**-0.33**	**-0.14**	**-0.30**	**-0.37**	**-0.16**

Table 8: Per-document r of Δs with ΔHTER for various measures s. Correlations by individual language×genre subcorpora.

	Arabic				Chinese			
Measure s	`bc`	`bn`	`nw`	`wb`	`bc`	`bn`	`nw`	`wb`
TER	0.54	**0.18**	0.11	0.19	0.15	0.14	0.26	**0.13**
1+BLEU$_4$	-0.36	-0.16	-0.07	-0.18	-0.09	-0.12	**-0.28**	-0.09
EDPM	**-0.59**	-0.12	**-0.15**	**-0.21**	**-0.18**	**-0.17**	**-0.27**	**-0.13**

Table 9: Per-sentence r of length-weighted Δs with length-weighted ΔHTER. Correlations by individual language×genre subcorpora.

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.

C. Callison-Burch. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. EACL*, pp. 249–256.

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

E. Charniak, K. Knight, and K. Yamada. 2003. Syntax-based language models for statistical machine translation. In *Proc. MT Summit IX*.

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

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. HLT*, pp. 138–145.

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.

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

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.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*.

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.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pp. 311–318.

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

M. Snover et al. 2006. A study of translation edit rate with targeted human annotation. In *Proc. AMTA*.

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