BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation

Dennis N. Mehay
Department of Linguistics
The Ohio State University
Columbus, OH, USA
mehay@ling.osu.edu

Chris Brew
Department of Linguistics
The Ohio State University
Columbus, OH, USA
cbrew@ling.osu.edu

Abstract

This paper describes a novel approach to syntactically-informed evaluation of machine translation (MT). Using a statistical, treebank-trained parser, we extract word-word dependencies from reference translations and then compile these dependencies into a representation that allows candidate translations to be evaluated by string comparisons, as is done in n-gram approaches to MT evaluation. This approach gains the benefit of syntactic analysis of the reference translations, but avoids the need to parse potentially noisy candidate translations. Preliminary experiments using 15,242 judgments of reference-candidate pairs from translations of Chinese newswire text show that the correlation of our approach with human judgments is only slightly lower than other reported results. With the addition of multiple reference translations, however, performance improves markedly. These results are encouraging, especially given that our system is a prototype and makes no essential use of synonymy, paraphrasing or inflectional morphological information, all of which would be easy to add.

1 Introduction

Effective automatic translation evaluation (ATE) systems are crucial to the development of machine translation (MT) systems, as the relative performance gain of each minor system modification must be tested quickly and cheaply. A professional human evaluation of MT system output after each such modification is too expensive and time-consuming for rapid, cost-effective deployment of translation software.

For the past few years, n-gram precision metrics for MT evaluation such as BLEU (Papineni et al., 2002) and the related NIST metric (Doddington, 2002) have been the standard approach to ATE. In essence, BLEU and NIST measure the quality of a candidate translation as a function of the number of n-grams (typically, $1 \leq n \leq 4$) it shares with a set of (one

or more) reference translations. These metrics require a one-time investment of creating a reference corpus of translations to test the system against, but are fully automatic once this corpus has been created and are very portable, requiring only word tokenisers for the reference set (if it is not already tokenised).

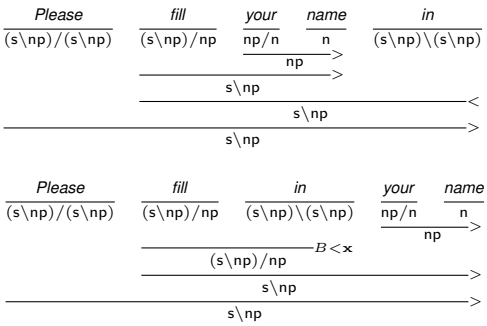
The portability of n-gram-based models, however, is one side of a trade-off with robustness: candidate translations are rewarded or penalised according to how well they match the *exact, contiguous word sequences in the reference set*. Candidates that contain legitimate word order variation will be penalised for not having these exact matches. Increasing the size of the reference set so as to capture more translational variation (as suggested by Thompson (1991)) is one possibility, but this is an expensive and time-consuming alternative. Moreover, given that adjuncts (e.g., adverbial modifiers), stacked attributive adjectives and a host of other grammatical elements can often "move around" without significantly affecting the meaning of a sentence, the strategy of padding the reference set with more examples for a word n-gram approach can only accommodate a fraction of the legitimate, syntactically-licensed variation in word order that a candidate translation should be allowed to display.

It seems reasonable, then, to explore approaches to ATE that exploit syntactic information so as not penalise legitimate syntactic variation. This paper describes such an approach. We describe here a prototype system called BLEUÂTRE[1] ("bluish"), a novel approach to syntactically-informed automatic machine translation evaluation that uses syntactic word-word dependencies from parses of reference translations. In this approach, we use a statistical Combinatory Categorial Grammar parser (Clark and Curran, 2004) to parse the reference set and extract word-word dependencies based on hierarchical head-dependent relationships (or "tectogrammatical" relationships). These dependencies are then compiled out into bags of dependent words that must appear to the left and right of each head word — essentially enforcing a partial linear ordering of dependents with respect to their heads. The quality of a candidate translation is then evaluated according to the number of these head word-dependent word partial orderings that it recalls. This approach is novel in that it only requires parses of reference translations, avoiding the need to parse (potentially noisy) candidate translations.

Preliminary experiments using 15,242 judgments of reference-candidate pairs from translations of Chinese newswire text show that BLEUÂTRE's correlation with human judgments is competitive with, but lower than, other reported results. With the addition of multiple reference translations for each system judgment, however, performance improves markedly. These results are encouraging, especially given that BLEUÂTRE is a prototype and makes no essential use of synonymy, paraphrasing or inflectional morphological information. The essential contribution of this paper is a description of how syntactic dependencies can be "flattened" to a form suitable for evaluating unparsed candidate translation sentences. We anticipate that this approach can be profitably combined with other syntactic and non-syntactic approaches to ATE.

The remainder of this paper is organised as follows: Section 2 describes how we use the parser to extract dependencies and how BLEUÂTRE uses these dependencies for eval-

[1] Standing for **BLEU**'s **A**ssociate with **T**ectogrammatical **RE**lations.



(det name$_3$ your$_2$)	(det name$_4$ your$_3$)
(dobj fill$_1$ name$_3$)	(dobj fill$_1$ name$_4$)
(ncmod _ fill$_1$ in$_4$)	(ncmod _ fill$_1$ in$_2$)
(xcomp _ please$_0$ fill$_1$)	(xcomp _ please$_0$ fill$_1$)

Figure 1: A CCG derivations and corresponding dependency graphs for the word order variants *Please fill your name in* and *Please fill in your name*.
(Key: det=‘determiner’, dobj=‘direct object’, ncmod=‘non-clausal modifier’ and xcomp=‘externally controlled clausal complement’.)

uation. Section 3 describes related work. Section 4 describes our preliminary experiments, and Section 5 is a conclusion that also briefly outlines future work.

2 Extracting and Using Dependencies for ATE

In our experiments, we use a statistical Combinatory Categorial Grammar (CCG) parser (Clark and Curran, 2004). CCG (Steedman, 2000) is a “mildly context-sensitive” formalism that provides elegant analyses of coordination (including “non-constituent” coordination), extraction, right node raising and other constructions that have proved challenging in other frameworks.

Figure 1 illustrates the CCG derivation and corresponding Briscoe and Carroll-style grammatical role dependencies that the Clark and Curran (C&C) parser outputs for the sentence *Please fill your name in*.[2] A parse of the semantically identical *Please fill in your name* would give the identical dependency graph (*modulo*, of course, the different string indices on the words).

Note, however that, if the first sentence is a reference translation and the second sentence is a candidate translation, then an n-gram-based approach to ATE would heavily penalise this minor variation in word order, even though it is identical both in syntactic dependency structure and semantic content. This is because, although the two sentences share all the same unigrams, the second sentence only contains two of the four bigrams from the reference sentence (and none of the 3-grams or 4-grams), giving it a relatively low BLEU score. A method that compared the overlap of the syntactic dependencies of the two sentences, however, would not penalise this minor word-order variation at all.

Note, however, that only a *correct* parse of the second sentence would give the identical dependency graph as the first. In fact the C&C parser, despite its state of the art performance,[3] does not parse this well-formed sentence correctly. Instead, due to part-of-speech tagging errors, it improperly treats ‘*in*’ as a preposition and not a particle, giving a parse that treats ‘*in your name*’ as a PP modifying a non-phrasal verb ‘*fill*’. This induces the following (incor-

[2] For the uninitiated, the horizontal (underlining) lines are analogous to branchings in a traditional tree representation of a syntactic derivation, where the $\ldots < \ldots$ and $\ldots > \ldots$ annotate the direction and type of the combinatory mechanism that produced each such “branching”.

[3] With $\approx 85\%$ balanced F-score in recovering both local and long-distance labelled dependencies.

rect) dependency graph:

(det name$_4$ your$_3$)
(dobj fill$_1$ in$_2$)
(dobj _ in$_2$ name$_4$)
(xcomp _ please$_0$ fill$_1$)

Ignoring the errors in the labels of the dependency arcs, we can see that the *unlabelled* dependency structure is also wrong: the direct dependency between '*fill*' and '*name*' is lost.

The fact that parsers can and often do err on well-formed sentences suggests that their performance will degrade considerably on less well-formed MT system output. This motivates the principle innovation of BLEUÂTRE: namely, we compile out the dependency triples from the parse of a candidate translation into bags of dependent words that must appear either to the left or right of each head word. This is essentially a partial linear ordering of dependents with respect to their heads. The essential point of this approach is that it avoids parsing MT system output. The following illustrates this process on our hypothetical reference sentence *Please fill your name in*:

$\emptyset$	$\overleftarrow{left}$	'**Please**'	$\overrightarrow{right}$	{'fill'}
$\emptyset$	$\overleftarrow{left}$	'**fill**'	$\overrightarrow{right}$	{'in','name'}
{'your'}	$\overleftarrow{left}$	'**name**'	$\overrightarrow{right}$	$\emptyset$

These partial orderings of dependents — which we shall sometimes call "*left and right contexts*" — allow candidates to be evaluated by a simple string search, verifying whether each of the dependents is either to the right or to the left of the head word as the case may be. The score of a candidate with respect to a reference is the number of such left-right orderings that it recalls multiplied by an exponentially decaying "length penalty", which is inspired by BLEU's brevity penalty. The intuition is that, the longer a candidate translation is, the more of the reference dependency orderings it is likely to recover, and, thus, candidate sentences longer than the reference must be penalised. Candidates shorter than the reference, in effect, penalise themselves, as they do not contain as many words that could match those in the left-right contexts, and, as such, no brevity penalty is assessed. In symbols, a candidate c's dependent ordering score for a single head word h that is in the reference r is the following:

$$\text{DEP}_{c,h,r} = \sum_{d_i \in \text{lf}(h)} \Lambda_c(d_i, h) + \sum_{d_j \in \text{rt}(h)} \rho_c(d_j, h)$$

where c is the candidate translation, $\text{lf}(h)$ is the left context of h in r, $\text{rt}(h)$ is the right context of h in r, and the functions $\Lambda_c(d_i, h)$ and $\rho_c(d_j, h)$ have value 1 if both $h \in c$ and d_i (or d_j, respectively) is to the left (or right) of h in c, and 0 otherwise.[4]

The BLEUÂTRE recall score of a candidate c with respect to a reference r is then:

$$\text{BLEUÂTRE}_{c,r} = LP_{c,r} \cdot \Big(\frac{\sum_{h \in r} \text{DEP}_{c,h,r}}{\sum_{h \in r} |\{d : d \in lf(h) \vee d \in rt(h)\}|} \Big)$$

Where $LP_{c,r}$, the length penalty of a candidate with respect to a reference, is simply BLEU's brevity penalty with the roles of the candidate and reference lengths reversed:

$$LP_{c,r} = \begin{cases} 1, \text{ if } len(c) < len(r) \\ e^{\left(1 - \frac{len(c)}{len(r)}\right)}, \text{ otherwise} \end{cases}$$

As a concrete example, take our hypothetical candidate translation *Please fill in your name*. This candidate scores a perfect 1.0, because

[4] Essentially, these functions signal whether the dependent is properly ordered with respect to the head in the candidate translation.

'*fill*' is to the right of '*Please*', '*in*' and '*name*' are to the right of '*fill*' and '*your*' is to the left of '*name*', and the sentences have the same length. Thus the syntactically licit word order variation is not penalised. Imagine further a less well-formed candidate translation from Dutch '*Vul even uw naam in*' $\Rightarrow$ '*Fill please your name in*'. Even though this candidate has only 1 bigram (and no 3- and 4-grams) in common with the reference (thus, giving it a low BLEU score), it still receives a fairly high BLEUÂTRE score of 0.75, since only '*please*' and '*fill*' are out of the order specified by the parse of the reference. This accords with our intuitions that '*Fill please your name in*' is only mildly "Dutch-sounding" and conveys the gist of the reference.

3 Related Work

There is a growing concern in the MT research community as to the correlation of BLEU with human judgments of translation quality, even at the document level (Callison-Burch et al., 2006). This is of particular concern, as statistical MT systems are now trained to minimise error with respect to ATE metrics (Och, 2003).

There have been many attempts to improve upon the performance of BLEU. The NIST metric mentioned above (Doddington, 2002) uses n-gram precision scores as BLEU does, but it weights the information contributed by certain n-grams. In this approach, rare n-grams count more than frequent n-grams in a candidate's precision score. Turian et al.'s (2003) approach (called General Text Matcher or GTM) is to compute both precision and recall of a candidate's match to the reference set, scoring contiguous sequences higher than discontiguous matches. Kulesza and Shieber (2004) describe a machine learning-based approach to combining various metrics such as BLEU-style n-gram precision ($1 \leq n \leq 5$), word error rate, position-independent word error rate, etc. These values are passed as features to a support vector machine (Vapnik, 1995) which learns to discriminate human from machine-generated translations. The farther a candidate translation's feature encoding is on the human side of the hyperplane separating human from machine translations, the better it is judged to be.

(Banerjee and Lavie, 2005) describes METEOR, a word-based generalised unigram matching approach that rewards sentence alignments between references and candidates that minimise the number of crossing word alignments. Stemming and WordNet synonyms are used to improve the match between translations that may differ only in their lexical choice or grammatical use of a particular base word form. All of these approaches, however, are still based on matching a candidate to a reference at the word level, and, as such, they are ultimately still susceptible to reduced performance due to syntactically acceptable variation.

Thus, some authors have attempted to use syntactic information in ATE. Liu and Gildea (2005) parse both reference and candidate translations. The count of subtrees up to a fixed, uniform depth that the candidate recalls is one metric used. Also, by decomposing each parse tree into a vector of counts of all subtrees, the authors compute the cosine between the reference and candidate vectors. Both metrics are also computed for dependency parses, as extracted from the phrase-structure parses of the candidate and reference translations. Finally, the authors compute the fraction of dependency chains (up to some fixed length) in the reference that are also in the candidate. The authors report improved correlation with human judgments as compared with BLEU.

Recently, Owczarzak et al. (2007) have reported using Lexical Functional Grammar (LFG) grammatical functional dependency triples to evaluate translation quality. Their approach is also to parse both the reference and candidate translations. They directly compute the dependency precision and recall of the candidate translation with respect to the reference. These authors perform an extensive comparison of their system to various ATE metrics over the Linguistic Data Consortium's Multiple Translation Chinese corpus (parts 2 and 4). When supplementing the dependency matches with WordNet synonyms, they achieve the highest correlation to human judgments in fluency and second place in an average of fluency and accuracy, as compared to BLEU, NIST, GTM, Translation Error Rate (TER, (Snover et al., 2006)) and METEOR. We have used this same corpus and, as such, can compare our results to theirs, as well as the other approaches they tested over this corpus. Our approach is distinguished from these last two approaches in that we do not attempt to parse candidate translations.

4 Preliminary Experiments

To test our system, we used sections 2 and 4 of the TIDES 2003 Chinese-to-English Multiple Translation corpus (MTC) of newswire text (released by the LDC). This corpus contains various commercial off-the-shelf (COTS) and research MT systems' translations of a set of Chinese source sentences. There are 4 human-produced reference translations for each source sentence. There are also human translation quality (fluency and accuracy) judgments for a subset of the machine-produced translations. We use these quality judgments to track the performance of BLEUÂTRE.

4.1 Experiment 1

The human judges were only shown a single "best" reference translation (as determined by an independent expert), and, so, following Owczarzak et al. (2007), we compute Pearson's correlation coefficient of the BLEUÂTRE score to each reference-candidate-judgment triple for our first experiment. This gives 15,242 total points of comparison (triples). This number is less than the 16,800 triples used by Owczarzak et al. (2007), as the C&C parser was only able to find a spanning analysis for 98.2% of the reference sentences, and many of these reference sentences are used several times as a gold standard for the human evaluators.[5]

The results of BLEUÂTRE's correlation to human fluency, accuracy and an average of the two are displayed in Table 1. To the extent that our approach is comparable with the results in (Owczarzak et al., 2007), we have listed their relevant results for comparison. Note that TER is negatively correlated with human judgments. This is because 0 is a perfect TER score. Owczarzak et al.(2007) note, however, that this still allows comparison of the absolute values of the correlation coefficients. Our system uses word-word dependencies, with no recourse to external morphological or thesaurus-based resources, such as WordNet. We therefore compare only with systems that use the same type of input. Future work may use a wider range of lexical resources and allow a wider range of meaningful comparisons.

We note that BLEUÂTRE does as well as

[5]The parser employs a back-off strategy that expands the parse search space incrementally to five back-off levels. After five unsuccessful back-off retries, however, the parser returns a failure notice and moves on to the next sentence. These settings are the off-the-shelf settings of the C&C parser with an additional, less-restrictive back-off level, as well as with a larger maximum size on the parse chart.

FL	HAC	AVE
BLEU 0.155*	MET 0.278*	MET 0.242*
OEtAl 0.154*	NIST 0.273*	NIST 0.238*
MET 0.149*	GTM 0.260*	OEtAl 0.236*
NIST 0.146*	OEtAl 0.224*	GTM 0.230*
GTM 0.146*	BA 0.202	BLEU 0.197*
TER -0.133*	BLEU 0.199*	BA 0.186
BA 0.128	TER -0.192*	TER -0.182*

Table 1: Pearson's correlation between various evaluation metrics and human judgments. BLEUÂTRE's results are our own. * indicates that the results are as reported in (Owczarzak et al., 2007) for the same set of reference-candidate-judgment triples (modulo C&C parsing failures). (Key: **BA**=BLEUÂTRE; **OEtAl**=Owczarzak et al.'s "predicate-argument dependency" system; **MET**=METEOR without WordNet or stemming; **FL**= Human fluency judgments; **HAC**=human accuracy judgments; **AVE**=Average of FL and HAC. Other abbreviations are given above.)

TER in fluency and both TER and BLEU in accuracy and fluency-accuracy average.[6]

Perhaps surprisingly, BLEUÂTRE correlates better with human accuracy judgments than with fluency judgments. We would expect approaches that pay appropriate attention to syntax to do well on fluency, because it is closely associated with grammatical well-formedness. We suspect that that BLEUÂTRE is still too conservative about word order variation. It seems to over-enforce partial orderings of dependents with respect to their heads [7]. It appears that human raters are better able to overlook this kind of variation, and that this emerges in their fluency judgments.

[6]Only a change of 0.015 or greater is significant at the 95% confidence level for both ours and Owczarzak et al.'s (2007) results.

[7]E.g. "*Fill your name in, please*" does not satisfy the partial (right-hand side) ordering of '*fill*' to '*Please*' as extracted from our hypothetical reference translation above.

FL	HAC	AVE
UFS 0.143	BA 0.208	BA 0.190
LFS 0.142	UFS 0.196	UFS 0.189
BA 0.130	LFS 0.194	LFS 0.188

Table 2: Pearson's correlation between BLEUÂTRE, and C&C parser-based f-score evaluation (labelled and unlabelled). Key: **BA**=BLEUÂTRE; **LFS**=Labelled F-score; **UFS**=Unlabelled F-score; (correlations to) **FL**=Human fluency judgments; **HAC**=human accuracy judgments; **AVE**=Average of FL and HAC. Only a difference of ± 0.016 is significant with 95% confidence (no significant differences).

4.2 Experiment 2

An obvious question raised by the above results is whether our decision not to parse candidate translations is helpful — it may be that the differences between Owczarzak et al. (2007)'s results and ours are not due to this feature of the system but rather to other differences such as the nature of the parsers or grammatical formalisms used (LFG vs. CCG). To investigate this, we compare BLEUÂTRE's correlation to human judgments to that of a re-implementation of the Owczarzak et al. (2007) approach by computing the f-score between parses of the candidate translations and the corresponding reference translations using the C&C parser. We compute this score for both labelled and unlabelled dependencies and compare it with BLEUÂTRE's correlation to a subset of the reference-candidate-triples where

both BLEUÂTRE and the f-score methods were able to provide a score.[8] This results in a set of 14,138 scores by BLEUÂTRE and the f-score methods compared against reference-candidate-judgment triples.

Table 2 gives the correlation of BLEUÂTRE and the two f-score methods to the relevant 14,138 human judgments. Although BLEUÂTRE differs slightly from the other methods, none of the differences is statistically significant. This confirms our intuition that BLEUÂTRE is proving effective at extracting and applying syntactic criteria when assigning scores to candidate translations. In effect, it is an alternative means of doing the job for which (Owczarzak et al., 2007) use the parser.

4.3 Experiment 3

In a third experiment, we include multiple reference translations to provide more partial orderings, thus minimising BLEUÂTRE's sensitivity to partial orderings extracted from a single reference translation. For this, we simply compute BLEUÂTRE scores for each candidate-reference pairing and pick the highest score as the BLEUÂTRE multiple-reference score. Owczarzak et al. (2007) do not describe such an experiment, and so our results are not comparable to theirs. Liu and Gildea (2005), however, do perform such an experiment, as do Banerjee and Lavie (2005). Accordingly, we performed two sub-experiments for comparison with these authors' work:[9]

[8] As the C&C parser only achieves 98% coverage on the reference set and 91% on the test set, we compare BLEUÂTRE and the f-score approach on the intersection of the parsed reference and candidate examples.

[9] Keeping in mind that the data sets are not identical due to C&C parsing failures. These failures, however, only lead to a few instances where there is no parsable reference sentence for a candidate. 915 sentences in E14 and 910 sentences in E15 were given BLEUÂTRE scores. Liu and Gildea report having 925 sentences per section, which means we have a loss of coverage of 1% and 2%, respectively, on these sections.

E14-FL	**E15-FL**
BA 0.199	**BA** 0.188
LG_dt 0.159*	**LG_pt** 0.144*
LG_dc 0.157*	**LG_dt** 0.137*
LG_pt 0.147*	**LG_dc** 0.128*
BLEU 0.132*	**BLEU** 0.122*
LG_dtvc 0.090*	**LG_ptvc** 0.089*
LG_ptvc 0.065*	**LG_dtvc** 0.066*

Table 3: Correlation of BLEUÂTRE and Liu and Gildea's metrics to human fluency judgments for systems E14 and E15. (Key: * indicates that the score is from (Liu and Gildea, 2005); **BA**=BLEUÂTRE; **LG**=Liu and Gildea — different approaches: **_dt**=dependency subtrees, **vc**=vector-cosines, **_pt** structural subtrees; **_dc**=dependency chains.)

First, following Liu and Gildea (2005), we ran BLEUÂTRE to compute scores for systems E14 and E15 on part 4 of the Chinese Multiple Translation corpus using three reference translations (namely, those from E01, E03 and E04). We compare the segment-level BLEUÂTRE scores to human fluency scores for those same sentences.[10] We list these scores next to their best reported per-system scores (including their figures for BLEU over the same set) in Table 3.[11]

Second, we compute BLEUÂTRE scores individually for systems E09, E11, E12, E14, E15 and E22 (MTC, Part 4) using all four reference translations in E01-E04. We list the average

[10] Liu and Gildea also compute "overall" scores, which they describe as the sum of the fluency and accuracy score. We do not compare with these numbers.

[11] In our correlation tests, a difference of 0.06 is significant at the 95% confidence level. It is difficult to say how this compares with Liu and Gildea's results, but their data set is essentially the same as ours.

	BLEUÂTRE	METEOR
E09	0.338	0.351
E11	0.193	0.253
E12	0.216	0.264
E14	0.257	0.285
E15	0.238	0.237
E22	0.273	0.284
AVE	0.253	0.279

Table 4: BLEUÂTRE and METEOR's correlation to an average of human judgments of fluency and accuracy for various MT systems.

FL	HAC	AVE
0.235	0.328	0.315

Table 5: BLEUÂTRE correlation to across-judge human judgments using multiple references (MTC 2 and 4). Key: **FL**= Human fluency judgments; **HAC**=human accuracy judgments; **AVE**=Average of FL and HAC.

of these scores next to the relevant METEOR score (without WordNet or Porter stemming) in Table 4. This set of systems is different from those reported in (Banerjee and Lavie, 2005) — which also includes system E17 — as we do not have E17 in our LDC corpus. The METEOR scores were obtained by running METEOR (v 0.5) on the above-mentioned data.

These scores demonstrate that, with multiple reference translations, BLEUÂTRE's performance improves markedly and becomes competitive with other systems that report results using multiple references. It is notable that only a difference of ± 0.016 is significant with 95% confidence ($p \leq$ 3.609e-11) for both systems (BLEUÂTRE and METEOR). Thus, the difference in performance between our system and METEOR is not shown to be significant here.

Finally, for all judgments in MTC Parts 2 and 4, Table 5 gives BLEUÂTRE's correlation with an average of each of the human fluency and accuracy judgments, as well as to the average of the averages of each fluency-accuracy pair while using all four references. We are not aware of any study that has reported these figures. We simply offer them for comparison.

5 Conclusion and Future Work

We have shown that it is possible to extract syntactic dependency information from a reference translation and compile it to a form that allows candidate translations to be evaluated by simple string searches. While our approach currently does not achieve state-of-the-art performance with only one reference translation, we are encouraged by the fact that it is at least competitive with other methods such as TER and BLEU, and its performance is not significantly different from a direct parse-to-parse f-measure comparison on the same data set, using the same parser. Further, when BLEUÂTRE is allowed to maximise its score over multiple reference translations, its performance improves markedly. Here it is competitive with state-of-the-art approaches such as METEOR (v 0.5), and perhaps superior to more complicated syntax-based methods such as that in (Liu and Gildea, 2005), all while avoiding the overhead of parsing at evaluation-time.

A strength of our approach is that it is compatible with any parsing approach that outputs dependency triples and relative string positions. To improve the performance of our system, we would like to experiment with different parsers, as well as with stemming, electronic thesauri such as WordNet, and sources of synonymy and paraphrasing such as that described in (Owczarzak et al., 2006).

Finally, some dependencies (e.g. determiner-

noun dependencies) are unsurprising and perhaps "easier" to get right, so they should arguably not contribute much to assessments of progress in the field. We would like to explore schemes for using NIST-like weights to reward candidate translations for recalling more "valuable" dependencies such as, e.g., verb-object dependencies that are systematically missed by well-known benchmark systems.

6 Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions. Thanks also to Detmar Meurers for helpful feedback when BLEUÂTRE was a half-formed idea. The CCG parses in this paper were produced using Ben Wing's "wccg" extension to OpenCCG.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings the ACL*, Ann Arbor, MI, USA.

Chris M Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the EACL-2006*, Trento, Italy.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the ACL*, Barcelona, Spain.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference*, San Diego, CA, USA.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore, MD, USA.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, MI, USA.

Franz Joseph Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the the ACL*, Sapporo, Japan.

Karolina Owczarzak, Declan Groves, Josef van Genabith, and Andy Way. 2006. Contextual bitext-derived paraphrases in automatic MT evaluation. In *Proceedings of the HLT-NAACL 2006 Workshop on Statistical Machine Translation*, New York, NY, USA.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In *Proceedings of the HLT-NAACL Workshop on Syntax and Structure in Statistical Translation*, Rochester, NY, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, Philadelphia, PA, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, and John Makhoul. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, Cambridge, MA, USA.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press, Cambridge, Massachusetts.

Henry Thompson. 1991. Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. In *(ISSCO) Proceedings of the Evaluators Forum*, Geneva, Switzerland.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*, New Orleans, LA, USA.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin.