Further Meta-Evaluation of Broad-Coverage Surface Realization

Dominic Espinosa and **Rajakrishnan Rajkumar** and **Michael White** and **Shoshana Berleant**
Department of Linguistics
The Ohio State University
Columbus, Ohio, USA
{espinosa,raja,mwhite,berleant}@ling.ohio-state.edu

Abstract

We present the first evaluation of the utility of automatic evaluation metrics on surface realizations of Penn Treebank data. Using outputs of the OpenCCG and XLE realizers, along with ranked WordNet synonym substitutions, we collected a corpus of generated surface realizations. These outputs were then rated and post-edited by human annotators. We evaluated the realizations using seven automatic metrics, and analyzed correlations obtained between the human judgments and the automatic scores. In contrast to previous NLG meta-evaluations, we find that several of the metrics correlate moderately well with human judgments of both adequacy and fluency, with the TER family performing best overall. We also find that all of the metrics correctly predict more than half of the significant system-level differences, though none are correct in all cases. We conclude with a discussion of the implications for the utility of such metrics in evaluating generation in the presence of variation. A further result of our research is a corpus of post-edited realizations, which will be made available to the research community.

1 Introduction and Background

In building surface-realization systems for natural language generation, there is a need for reliable automated metrics to evaluate the output. Unlike in parsing, where there is usually a single gold-standard parse for a sentence, in surface realization there are usually many grammatically-acceptable ways to express the same concept. This parallels the task of evaluating machine-translation (MT) systems: for a given segment in the source language, there are usually several acceptable translations into the target language. As human evaluation of translation quality is time-consuming and expensive, a number of automated metrics have been developed to evaluate the quality of MT outputs. In this study, we investigate whether the metrics developed for MT evaluation tasks can be used to reliably evaluate the outputs of surface realizers, and which of these metrics are best suited to this task.

A number of surface realizers have been developed using the Penn Treebank (PTB), and BLEU scores are often reported in the evaluations of these systems. But how useful is BLEU in this context? The original BLEU study (Papineni et al., 2001) scored MT outputs, which are of generally lower quality than grammar-based surface realizations. Furthermore, even for MT systems, the usefulness of BLEU has been called into question (Callison-Burch et al., 2006). BLEU is designed to work with multiple reference sentences, but in treebank realization, there is only a single reference sentence available for comparison.

A few other studies have investigated the use of such metrics in evaluating the output of NLG systems, notably (Reiter and Belz, 2009) and (Stent et al., 2005). The former examined the performance of BLEU and ROUGE with computer-generated weather reports, finding a moderate correlation with human fluency judgments. The latter study applied several MT metrics to paraphrase data from Barzilay and Lee's corpus-based system (Barzilay and Lee, 2003), and found moderate correlations with human adequacy judgments, but little correlation with fluency judgments. Cahill (2009) examined the performance of six MT metrics (including BLEU) in evaluating the output of a LFG-based surface realizer for

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 564–574,
MIT, Massachusetts, USA, 9-11 October 2010.

German, also finding only weak correlations with the human judgments.

To study the usefulness of evaluation metrics such as BLEU on the output of grammar-based surface realizers used with the PTB, we assembled a corpus of surface realizations from three different realizers operating on Section 00 of the PTB. Two human judges evaluated the adequacy and fluency of each of the realizations with respect to the reference sentence. The realizations were then scored with a number of automated evaluation metrics developed for machine translation. In order to investigate the correlation of targeted metrics with human evaluations, and gather other acceptable realizations for future evaluations, the judges manually repaired each unacceptable realization during the rating task. In contrast to previous NLG meta-evaluations, we found that several of the metrics correlate moderately well with human judgments of both adequacy and fluency, with the TER family performing best. However, when looking at statistically significant system-level differences in human judgments, we found that some of the metrics get some of the rankings correct, but none get them all correct, with different metrics making different ranking errors. This suggests that multiple metrics should be routinely consulted when comparing realizer systems.

Overall, our methodology is similar to that of previous MT meta-evaluations, in that we collected human judgments of system outputs, and compared these scores with those assigned by automatic metrics. A recent alternative approach to paraphrase evaluation is ParaMetric (Callison-Burch et al., 2008); however, it requires a corpus of annotated (aligned) paraphrases (which does not yet exist for PTB data), and is arguably focused more on paraphrase analysis than paraphrase generation.

The plan of the paper is as follows: Section 2 discusses the preparation of the corpus of surface realizations. Section 3 describes the human evaluation task and the automated metrics applied. Sections 4 and 5 present and discuss the results of these evaluations. We conclude with some general observations about automatic evaluation of surface realizers, and some directions for further research.

2 Data Preparation

We collected realizations of the sentences in Section 00 of the WSJ corpus from the following three sources:

1. OpenCCG, a CCG-based chart realizer (White, 2006)

2. The XLE Generator, a LFG-based system developed by Xerox PARC (Crouch et al., 2008)

3. WordNet synonym substitutions, to investigate how differences in lexical choice compare to grammar-based variation.[1]

Although all three systems used Section 00 of the PTB, they were applied with various parameters (e.g., language models, multiple-output versus single-output) and on different input structures. Accordingly, our study does not compare OpenCCG to XLE, or either of these to the WordNet system.

2.1 OpenCCG realizations

OpenCCG is an open source parsing/realization library with multimodal extensions to CCG (Baldridge, 2002). The OpenCCG chart realizer takes logical forms as input and produces strings by combining signs for lexical items. Alternative realizations are scored using integrated n-gram and perceptron models. For robustness, fragments are greedily assembled when necessary. Realizations were generated from 1,895 gold standard logical forms, created by constrained parsing of development-section derivations. The following OpenCCG models (which differ essentially in the way the output is ranked) were used:

1. Baseline 1: Output ranked by a trigram word model

2. Baseline 2: Output ranked using three language models (3-gram words + 3-gram words with named entity class replacement + factored language model of words, POS tags and CCG supertags)

[1] Not strictly surface realizations, since they do not involve an abstract input specification, but for simplicity we refer to them as realizations throughout.

3. Baseline 3: Perceptron with syntax features and the three LMs mentioned above

4. Perceptron full-model: n-best realizations ranked using perceptron with syntax features and the three n-gram models, as well as discriminative n-grams

The perceptron model was trained on sections 02-21 of the CCGbank, while a grammar extracted from section 00-21 was used for realization. In addition, oracle supertags were inserted into the chart during realization. The purpose of such a non-blind testing strategy was to evaluate the quality of the output produced by the statistical ranking models in isolation, rather than focusing on grammar coverage, and avoid the problems associated with lexical smoothing, i.e. lexical categories in the development section not being present in the training section.

To enrich the variation in the generated realizations, dative-alternation was enforced during realization by ensuring alternate lexical categories of the verb in question, as in the following example:

(1) the executives gave [the chefs] [a standing ovation]

(2) the executives gave [a standing ovation] [to the chefs]

2.2 XLE realizations

The corpus of realizations generated by the XLE system contained 42,527 surface realizations of approximately 1,421 section 00 sentences (an average of 30 per sentence), initially unranked. The LFG f-structures used as input to the XLE generator were derived from automatic parses, as described in (Riezler et al., 2002). The realizations were first tokenized using Penn Treebank conventions, then ranked using perplexities calculated from the same trigram word model used with OpenCCG. For each sentence, the top 4 realizations were selected. The XLE generator provides an interesting point of comparison to OpenCCG as it uses a manually-developed grammar with inputs that are less abstract but potentially noisier, as they are derived from automatic parses rather than gold-standard ones.

2.3 WordNet synonymizer

To produce an additional source of variation, the nouns and verbs of the sentences in section 00 of the PTB were replaced with all of their WordNet synonyms. Verb forms were generated using verb stems, part-of-speech tags, and the *morphg* tool.[2] These substituted outputs were then filtered using the n-gram data which Google Inc. has made available.[3] Those without any 5-gram matches centered on the substituted word (or 3-gram matches, in the case of short sentences) were eliminated.

3 Evaluation

From the data sources described in the previous section, a corpus of realizations to be evaluated by the human judges was constructed by randomly choosing 305 sentences from section 00, then selecting surface realizations of these sentences using the following algorithm:

1. Add OpenCCG's best-scored realization.

2. Add other OpenCCG realizations until all four models are represented, to a maximum of 4.

3. Add up to 4 realizations from either the XLE system or the WordNet pool, chosen randomly.

The intent was to give reasonable coverage of all realizer systems discussed in Section 2 without overloading the human judges. "System" here means any instantiation that emits surface realizations, including various configurations of OpenCCG (using different language models or ranking systems), and these can be multiple-output, such as an n-best list, or single-output (best-only, worst-only, etc.). Accordingly, more realizations were selected from the OpenCCG realizer because 5 different systems were being represented. Realizations were chosen randomly, rather than according to sentence types or other criteria, in order to produce a representative sample of the corpus. In total, 2,114 realizations were selected for evaluation.

[2] http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/morph.html

[3] http://www.ldc.upenn.edu/Catalog/docs/LDC2006T13/readme.txt

3.1 Human judgments

Two human judges evaluated each surface realization on two criteria: *adequacy*, which represents the extent to which the output conveys all and only the meaning of the reference sentence; and *fluency*, the extent to which it is grammatically acceptable. The realizations were presented to the judges in sets containing a reference sentence and the 1-8 outputs selected for that sentence. To aid in the evaluation of adequacy, one sentence each of leading and trailing context were displayed. Judges used the guidelines given in Figure 1, based on the scales developed by the NIST Machine Translation Evaluation Workshop.

In addition to rating each realization on the two five-point scales, each judge also repaired each output which he or she did not judge to be fully adequate and fluent. An example is shown in Figure 2. These repairs resulted in new reference sentences for a substantial number of sentences. These repaired realizations were later used to calculate *targeted* versions of the evaluation metrics, i.e., using the repaired sentence as the reference sentence. Although targeted metrics are not fully automatic, they are of interest because they allow the evaluation algorithm to focus on what is actually wrong with the input, rather than all textual differences. Notably, targeted TER (HTER) has been shown to be more consistent with human judgments than human annotators are with one another (Snover et al., 2006).

3.2 Automatic evaluation

The realizations were also evaluated using seven automatic metrics:

- IBM's BLEU, which scores a hypothesis by counting n-gram matches with the reference sentence (Papineni et al., 2001), with smoothing as described in (Lin and Och, 2004)
- The NIST n-gram evaluation metric, similar to BLEU, but rewarding rarer n-gram matches, and using a different length penalty
- METEOR, which measures the harmonic mean of unigram precision and recall, with a higher weight for recall (Banerjee and Lavie, 2005)
- TER (Translation Edit Rate), a measure of the number of edits required to transform a hypothesis sentence into the reference sentence (Snover et al., 2006)
- TERP, an augmented version of TER which performs phrasal substitutions, stemming, and checks for synonyms, among other improvements (Snover et al., 2009)
- TERPA, an instantiation of TERP with edit weights optimized for correlation with adequacy in MT evaluations
- GTM (General Text Matcher), a generalization of the F-measure that rewards contiguous matching spans (Turian et al., 2003)

Additionally, targeted versions of BLEU, METEOR, TER, and GTM were computed by using the human-repaired outputs as the reference set. The human repair was different from the reference sentence in 193 cases (about 9% of the total), and we expected this to result in better scores and correlations with the human judgments overall.

4 Results

4.1 Human judgments

Table 1 summarizes the dataset, as well as the mean adequacy and fluency scores garnered from the human evaluation. Overall adequacy and fluency judgments were high (4.16, 3.63) for the realizer systems on average, and the best-rated realizer systems achieved mean fluency scores above 4.

4.2 Inter-annotator agreement

Inter-annotator agreement was measured using the κ-coefficient, which is commonly used to measure the extent to which annotators agree in category judgment tasks. κ is defined as $\frac{P(A)-P(E)}{1-P(E)}$, where $P(A)$ is the observed agreement between annotators and $P(E)$ is the probability of agreement due to chance (Carletta, 1996). Chance agreement for this data is calculated by the method discussed in Carletta's squib. However, in previous work in MT meta-evaluation, Callison-Burch et al. (2007), assume the less strict criterion of uniform chance agreement, i.e. $\frac{1}{5}$ for a five-point scale. They also

Score	*Adequacy*	*Fluency*
5	All the meaning of the reference	Perfectly grammatical
4	Most of the meaning	Awkward or non-native; punctuation errors
3	Much of the meaning	Agreement errors or minor syntactic problems
2	Meaning substantially different	Major syntactic problems, such as missing words
1	Meaning completely different	Completely ungrammatical

Figure 1: Rating scale and guidelines

Ref.	It wasn't clear how NL and Mr. Simmons would respond if Georgia Gulf spurns them again
Realiz.	It weren't clear how NL and Mr. Simmons would respond if Georgia Gulf again spurns them
Repair	It wasn't clear how NL and Mr. Simmons would respond if Georgia Gulf again spurns them

Figure 2: Example of repair

introduce the notion of "relative" κ, which measures how often two or more judges agreed that $A > B$, $A = B$, or $A < B$ for two outputs A and B, irrespective of the specific values given on the five-point scale; here, uniform chance agreement is taken to be $\frac{1}{3}$. We report both absolute and relative κ in Table 2, using actual chance agreement rather than uniform chance agreement.

The κ scores of 0.60 for adequacy and 0.63 for fluency across the entire dataset represent "substantial" agreement, according to the guidelines discussed in (Landis and Koch, 1977), better than is typically reported for machine translation evaluation tasks; for example, Callison-Burch et al. (2007) reported "fair" agreement, with κ = 0.281 for fluency and κ = 0.307 for adequacy (relative). Assuming the uniform chance agreement that the previously cited work adopts, our inter-annotator agreements (both absolute and relative) are still higher. This is likely due to the generally high quality of the realizations evaluated, leading to easier judgments.

4.3 Correlation with automatic evaluation

To determine how well the automatic evaluation methods described in Section 3 correlate with the human judgments, we averaged the human judgments for adequacy and fluency, respectively, for each of the rated realizations, and then computed both Pearson's correlation coefficient and Spearman's rank correlation coefficient between these scores and each of the metrics. Spearman's correlation makes fewer assumptions about the distribution of the data, but may not reflect a linear relationship that is actually present. Both are frequently reported in the literature. Due to space constraints, we show only Spearman's correlation, although the TER family scored slightly better on Pearson's coefficient, relatively.

The results for Spearman's correlation are given in Table 3. Additionally, the average scores for adequacy and fluency were themselves averaged into a single score, following (Snover et al., 2009), and the Spearman's correlation of each of the automatic metrics with these scores are given in Table 4. All reported correlations are significant at $p < 0.001$.

4.4 Bootstrap sampling of correlations

For each of the sub-corpora shown in Table 1, we computed confidence intervals for the correlations between adequacy and fluency human scores with selected automatic metrics (BLEU, HBLEU, TER, TERP, and HTER) as described in (Koenh, 2004). We sampled each sub-corpus 1000 times with replacement, and calculated correlations between the rankings induced by the human scores and those induced by the metrics for each reference sentence. We then used these coefficients to estimate the confidence interval, after excluding the top 25 and bottom 25 coefficients, following (Lin and Och, 2004). The results of this for the BLEU metric are shown in Table 5. We determined which correlations lay within the 95% confidence interval of the best performing metric in each row of Table Table 3; these figures are italicized.

5 Discussion

5.1 Human judgments of systems

The results for the four OpenCCG perceptron models mostly confirm those reported in (White and Rajkumar, 2009), with one exception: the B-3 model was below B-2, though the P-B (perceptron-best) model still scored highest. This may have been due to differences in the testing scenario. None of the differences in adequacy scores among the individual systems are significant, with the exception of the WordNet system. In this case, the lack of word-sense disambiguation for the substituted words results in a poor overall adequacy score (e.g., *wage floor* → *wage story*). Conversely, it scores highest for fluency, as substituting a noun or verb with a synonym does not usually introduce ungrammaticality.

5.2 Correlations of human judgments with MT metrics

Of the non-human-targeted metrics evaluated, BLEU and TER/TERP demonstrate the highest correlations with the human judgments of fluency ($r = 0.62, 0.64$). The TER family of evaluation metrics have been observed to perform very well in MT-evaluation tasks, and although the data evaluated here differs from typical MT data in some important ways, the correlation of TERP with the human judgments is substantial. In contrast with previous MT evaluations where TERP performs considerably better than TER, these scored close to equal on our data, possibly because TERP's stem, synonym, and paraphrase matching are less useful when most of the variation is syntactic.

The correlations with BLEU and METEOR are lower than those reported in (Callison-Burch et al., 2007); in that study, BLEU achieved adequacy and fluency correlations of 0.690 and 0.722, respectively, and METEOR achieved 0.701 and 0.719. The correlations for these metrics might be expected to be lower for our data, since overall quality is higher, making the metrics' task more difficult as the outputs involve subtler differences between acceptable and unacceptable variation.

The human-targeted metrics (represented by the prefixed *H* in the data tables) correlated even more strongly with the human judgments, compared to the non-targeted versions. HTER demonstrated the best correlation with realizer fluency ($r = 0.75$).

For several kinds of acceptable variation involving the rearrangement of constituents (such as dative shift), TERP gives a more reasonable score than BLEU, due to its ability to directly evaluate phrasal shifts. The following realization was rated 4.5 for fluency, and was more correctly ranked by TERP than BLEU:

(3) *Ref:* The deal also gave Mitsui access to a high-tech medical product.

(4) *Realiz.:* The deal also gave access to a high-tech medical product to Mitsui.

For each reference sentence, we compared the ranking of its realizations induced from the human scores to the ranking induced from the TERP score, and counted the rank errors by the latter, informally categorizing them by error type (see Table 7). In the 50 sentences with the highest numbers of rank errors, 17 were affected by punctuation differences, typically involving variation in comma placement. Human fluency judgments of outputs with only punctuation problems were generally high, and many realizations with commas inserted or removed were rated fully fluent by the annotators. However, TERP penalizes such insertions or deletions. Agreement errors are another frequent source of ranking errors for TERP. The human judges tended to harshly penalize sentences with number-agreement or tense errors, whereas TERP applies only a single substitution penalty for each such error. We expect that with suitable optimization of edit weights to avoid over-penalizing punctuation shifts and under-penalizing agreement errors, TERP would exhibit an even stronger correlation with human fluency judgments.

None of the evaluation metrics can distinguish an acceptable movement of a word or constituent from an unacceptable movement, with only one reference sentence. A substantial source of error for both TERP and BLEU is variation in adverbial placement, as shown in (7).

Similar errors are seen with prepositional phrases and some commonly-occurring temporal adverbs, which typically admit a number of variations in placement. Another important example of acceptable variation which these metrics do not generally rank correctly is dative alternation:

Ref. We need to clarify what exactly is wrong with it.

(7)

Realiz.	*Flu.*	TERP	BLEU
We need to clarify exactly what is wrong with it.	5	0.1	0.5555
We need to clarify exactly what 's wrong with it.	5	0.2	0.4046
We need to clarify what , exactly , is wrong with it.	5	0.2	0.5452
We need to clarify what is wrong with it exactly.	4.5	0.1	0.6756
We need to clarify what exactly , is wrong with it.	4	0.1	0.7017
We need to clarify what , exactly is wrong with it.	4	0.1	0.7017
We needs to clarify exactly what is wrong with it.	3	0.103	0.346

(5) *Ref.* When test booklets were passed out 48 hours ahead of time, she says she copied questions in the social studies section and gave the answers to students.

(6) *Realiz.* When test booklets were passed out 48 hours ahead of time , she says she copied questions in the social studies section and gave students the answers.

The correlations of each of the metrics with the human judgments of fluency for the realizer systems indicate at least a moderate relationship, in contrast with the results reported in (Stent et al., 2005) for paraphrase data, which found an inverse correlation for fluency, and (Cahill, 2009) for the output of a surface realizer for German, which found only a weak correlation. However, the former study employed a corpus-based paraphrase generation system rather than grammar-driven surface realizers, and the resulting paraphrases exhibited much broader variation. In Cahill's study, the outputs of the realizer were almost always grammatically correct, and the automated evaluation metrics were ranking markedness instead of grammatical acceptability.

5.3 System-level comparisons

In order to investigate the efficacy of the metrics in ranking different realizer systems, or competing realizations from the same system generated using different ranking models, we considered seven different "systems" from the whole dataset of realizations. These consisted of five OpenCCG-based realizations (the best realization from three baseline models, and the best and the worst realization from the full perceptron model), and two XLE-based systems (the best and the worst realization, after ranking the outputs of the XLE realizer with an n-gram model). The mean of the combined adequacy and fluency scores of each of these seven systems was compared with that of every other system, resulting in 21 pairwise comparisons. Then Tukey's HSD test was performed to determine the systems which differed significantly in terms of the average adequacy and fluency rating they received.[4] The test revealed five pairwise comparisons where the scores were significantly different.

Subsequently, for each of these systems, an overall system-level score for each of the MT metrics was calculated. For the five pairwise comparisons where the adequacy-fluency group means differed significantly, we checked whether the metric ranked the systems correctly. Table 8 shows the results of a pairwise comparison between the ranking induced by each evaluation metric, and the ranking induced by the human judgments. Five of the seven non-targeted metrics correctly rank more than half of the systems. NIST, METEOR, and GTM get the most comparisons right, but neither NIST nor GTM correctly rank the OpenCCG-baseline model 1 with respect to the XLE-best model. TER and TERP get two of the five comparisons correct, and they incorrectly rank two of the five OpenCCG model comparisons, as well as the comparison between the XLE-worst and OpenCCG-best systems.

For the targeted metrics, HNIST is correct for all five comparisons, while neither HBLEU nor HMETEOR correctly rank all the OpenCCG models. On the other hand, HTER and HGTM incorrectly rank the XLE-best system versus OpenCCG-based models.

In summary, some of the metrics get some of the rankings correct, but none of the non-targeted metrics get all of them correct. Moreover, different metrics make different ranking errors. This argues for

[4] This particular test was chosen since it corrects for multiple post-hoc analyses conducted on the same data-set.

the use of multiple metrics in comparing realizer systems.

6 Conclusion

Our study suggests that although the task of evaluating the output from realizer systems differs from the task of evaluating machine translations, the automatic metrics used to evaluate MT outputs deliver moderate correlations with combined human fluency and adequacy scores when used on surface realizations. We also found that the MT-evaluation metrics are useful in evaluating different versions of the same realizer system (e.g., the various OpenCCG realization ranking models), and finding cases where a system is performing poorly. As in MT-evaluation tasks, human-targeted metrics have the highest correlations with human judgments overall. These results suggest that the MT-evaluation metrics are useful for developing surface realizers. However, the correlations are lower than those reported for MT data, suggesting that they should be used with caution, especially for cross-system evaluation, where consulting multiple metrics may yield more reliable comparisons. In our study, the targeted version of TERP correlated most strongly with human judgments of fluency.

In future work, the performance of the TER family of metrics on this data might be improved by optimizing the edit weights used in computing its scores, so as to avoid over-penalizing punctuation movements or under-penalizing agreement errors, both of which were significant sources of ranking errors. Multiple reference sentences may also help mitigate these problems, and the corpus of human-repaired realizations that has resulted from our study is a step in this direction, as it provides multiple references for some cases. We expect the corpus to also prove useful for feature engineering and error analysis in developing better realization models.[5]

Acknowledgements

We thank Aoife Cahill and Tracy King for providing us with the output of the XLE generator. We also thank Chris Callison-Burch and the anonymous reviewers for their helpful comments and suggestions.

This material is based upon work supported by the National Science Foundation under Grant No. 0812297.

[5]The corpus can be downloaded from `http://www.ling.ohio-state.edu/~mwhite/data/emnlp10/`.

References

Jason Baldridge. 2002. *Lexically Specified Derivational Control in Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

S. Banerjee and A. Lavie. 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 Machine Translation and/or Summarization*, pages 65–72.

R. Barzilay and L. Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *proceedings of HLT-NAACL*, volume 2003, pages 16–23.

Aoife Cahill. 2009. Correlating human and automatic evaluation of a german surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 97–100, Suntec, Singapore, August. Association for Computational Linguistics.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*, volume 2006, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Morristown, NJ, USA. Association for Computational Linguistics.

C. Callison-Burch, T. Cohn, and M. Lapata. 2008. Parametric: An automatic evaluation metric for paraphrasing. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 97–104. Association for Computational Linguistics.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2008. Xle documentation. Technical report, Palo Alto Research Center.

Philip Koenh. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 501, Morristown, NJ, USA. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report, IBM Research.

E. Reiter and A. Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. III Maxwell, and Mark Johnson. 2002. Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

M. Snover, N. Madnani, B.J. Dorr, and R. Schwartz. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Proceedings of CICLing*.

J.P. Turian, L. Shen, and I.D. Melamed. 2003. Evaluation of machine translation and its evaluation. *recall (C— R)*, 100:2.

Michael White and Rajakrishnan Rajkumar. 2009. Perceptron reranking for CCG realization. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Singapore, August. Association for Computational Linguistics.

Michael White. 2006. Efficient Realization of Coordinate Structures in Combinatory Categorial Grammar. *Research on Language and Computation*, 4(1):39–75.

Type	System	#Refs	#Paraphrases	Average Paraphrases/Ref	#Exact Matches	Adq	Flu
Single output	OpenCCG Baseline 1	296	296	1.0	72	4.17	3.65
	OpenCCG Baseline 2	296	296	1.0	82	4.34	3.94
	OpenCCG Baseline 3	296	296	1.0	76	4.31	3.86
	OpenCCG Perceptron Best	296	1.0	1.0	112	4.37	4.09
	OpenCCG Perceptron Worst	117	117	1.0	5	4.34	3.36
	XLE Best	154	154	1.0	24	4.41	4.07
	XLE Worst	157	157	1.0	13	4.08	3.73
Multiple output	OpenCCG-Perceptron All	296	767	2.6	158	4.45	3.91
	OpenCCG All	296	1131	3.8	162	4.20	3.61
	XLE All	174	557	3.2	54	4.17	3.81
	Wordnet Subsitutions	162	486	3.0	0	3.66	4.71
	Realizer All	296	1628	5.0	169	4.16	3.63
	All	296	2114	7.1	169	4.05	3.88

Table 1: Descriptive statistics

System	Adq			Flu		
	p(A)	p(E)	κ	p(A)	p(E)	κ
OpenCCG-Abs	0.73	0.47	0.48	0.70	0.24	0.61
OpenCCG-Rel	0.76	0.47	0.54	0.76	0.34	0.64
XLE-Abs	0.68	0.42	0.44	0.69	0.27	0.58
XLE-Rel	0.73	0.45	0.50	0.69	0.37	0.50
Wordnet-Abs	0.57	0.25	0.43	0.77	0.66	0.33
Wordnet-Rel	0.74	0.34	0.61	0.73	0.60	0.33
Realizer-Abs	0.70	0.44	0.47	0.69	0.24	0.59
Realizer-Rel	0.74	0.41	0.56	0.73	0.33	0.60
All-Abs	0.67	0.38	0.47	0.71	0.29	0.59
All-Rel	0.74	0.36	**0.60**	0.75	0.34	**0.63**

Table 2: Corpora-wise inter-annotator agreement (absolute and relative κ values shown)

Sys	N	B	M	G	TP	TA	T	HT	HN	HB	HM	HG
OpenCCG-Adq	0.27	*0.39*	0.35	0.18	*0.39*	0.34	**0.4**	**0.43**	0.3	**0.43**	**0.43**	0.23
OpenCCG-Flu	0.49	0.55	0.4	0.42	**0.6**	0.46	**0.6**	**0.72**	0.58	0.69	0.57	0.53
XLE-Adq	*0.52*	*0.51*	**0.55**	0.31	*0.5*	*0.5*	*0.5*	0.52	0.47	0.51	**0.61**	0.4
XLE-Flu	**0.56**	**0.56**	0.48	0.37	*0.55*	0.5	*0.55*	**0.61**	0.54	**0.61**	0.51	0.51
Wordnet-Adq	0.17	0.14	0.24	0.15	**0.37**	0.26	0.22	**0.64**	0.52	0.56	0.32	0.6
Wordnet-Flu	0.26	0.21	0.24	0.24	0.22	**0.27**	0.26	**0.34**	0.32	**0.34**	0.3	**0.34**
Realizer-Adq	0.47	**0.6**	*0.57*	0.42	*0.59*	*0.57*	**0.6**	*0.62*	0.49	*0.62*	**0.65**	0.48
Realizer-Flu	0.51	*0.62*	0.52	0.5	*0.63*	0.53	**0.64**	**0.75**	0.59	*0.73*	0.65	0.63
All-Adq	0.37	0.37	0.33	0.32	*0.42*	0.31	**0.43**	**0.53**	0.44	0.48	0.44	0.45
All-Flu	0.21	**0.62**	0.51	0.32	*0.61*	0.55	*0.6*	*0.7*	0.33	**0.71**	0.62	0.48

Table 3: Spearman's correlations among NIST (N), BLEU (B), METEOR (M), GTM (G), TERp (TP), TERpa (TA), TER (T), human variants (HN, HB, HM, HT, HG) and human judgments (-Adq: adequacy and -Flu: Fluency); Scores which fall within the 95 %CI of the best are italicized.

Sys	N	B	M	G	TP	TA	T	HT	HN	HB	HM	HG
OpenCCG	0.49	0.57	0.42	0.4	0.61	0.46	**0.62**	**0.73**	0.58	0.7	0.59	0.51
XLE	0.63	**0.64**	0.59	0.39	0.62	0.58	0.63	**0.69**	0.6	0.68	0.63	0.54
Wordnet	0.21	0.14	0.21	0.19	**0.38**	0.25	0.23	**0.65**	0.56	0.57	0.31	0.63
Realizer	0.55	0.68	0.57	0.5	0.68	0.58	**0.69**	**0.78**	0.61	0.77	0.7	0.63
All	0.34	0.58	0.47	0.38	**0.61**	0.48	**0.61**	**0.75**	0.48	0.73	0.61	0.58

Table 4: Spearman's correlations among NIST (N), BLEU (B), METEOR (M), GTM (G), TERp (TP), TERpa (TA), TER (T), human variants (HN, HB, HM, HT, HG) and human judgments (combined adequacy and fluency scores)

System	Adq			Flu		
	Sp	95%L	95%U	Sp	95%L	95%U
Realizer	0.60	0.58	0.63	0.62	0.59	0.65
XLE	0.51	0.47	0.56	0.56	0.51	0.61
OpenCCG	0.39	0.35	0.42	0.55	0.52	0.59
All	0.37	0.34	0.4	0.62	0.6	0.64
Wordnet	0.14	0.06	0.21	0.21	0.13	0.28

Table 5: Spearman's correlation analysis (bootstrap sampling) of the BLEU scores of various systems with human adequacy and fluency scores

Sys	HJ	N	B	M	G	TP	TA	T	HT	HN	HB	HM	HG	HJ1-HJ2
OpenCCG	HJ-1	0.44	0.52	0.39	0.36	0.56	0.43	**0.58**	**0.75**	0.58	0.72	0.62	0.52	0.76
	HJ-2	0.5	0.58	0.43	0.4	0.62	0.46	**0.63**	**0.7**	0.55	0.68	0.56	0.49	
XLE	HJ-1	**0.6**	**0.6**	0.55	0.37	0.57	0.55	0.58	**0.69**	0.63	0.68	0.64	0.54	0.75
	HJ-2	0.6	0.6	0.56	0.39	0.6	0.55	**0.61**	**0.64**	0.54	0.61	0.57	0.51	
Wordnet	HJ-1	0.2	0.18	0.26	0.16	**0.37**	0.28	0.24	**0.7**	0.59	0.64	0.35	0.65	0.72
	HJ-2	0.25	0.16	0.23	0.19	**0.37**	0.25	0.25	**0.59**	0.52	0.51	0.32	0.56	
Realizer	HJ-1	0.51	0.65	0.56	0.49	0.64	0.56	**0.66**	**0.8**	0.62	0.78	0.72	0.64	0.82
	HJ-2	0.55	**0.68**	0.56	0.5	0.67	0.57	**0.68**	**0.74**	0.58	0.73	0.66	0.6	
All	HJ-1	0.32	0.53	0.45	0.37	**0.57**	0.44	**0.57**	**0.77**	0.5	0.74	0.62	0.59	0.79
	HJ-2	0.35	0.58	0.46	0.37	**0.61**	0.47	0.6	**0.71**	0.44	0.69	0.57	0.54	

Table 6: Spearman's correlations of NIST (N), BLEU (B), METEOR (M), GTM (G), TERp (TP), TERpa (TA), human variants (HT, HN, HB, HM, HG), and individual human judgments (combined adq. and flu. scores)

Factor	*Count*
Punctuation	17
Adverbial shift	16
Agreement	14
Other shifts	8
Conjunct rearrangement	8
Complementizer ins/del	5
PP shift	4

Table 7: Factors influencing TERP ranking errors for 50 worst-ranked realization groups

Metric	Score	Errors
nist	4	C1-XB
bleu	3	XB-PW C1-XB
meteor	4	XW-PB
ter	2	PW-PB XW-PB C1-PB
terp	2	PW-PB XW-PB C1-PB
terpa	3	XW-PB C1-PB
gtm	4	C1-XB
hnist	5	
hbleu	3	PW-PB XW-PB
hmeteor	2	PW-PB XW-PB C1-PB
hter	3	XB-PW C1-XB
hgtm	3	XB-PW C1-XB

Table 8: Metric-wise ranking performance in terms of agreement with a ranking induced by combined adequacy and fluency scores; each metric gets a score out of 5 (i.e. number of system-level comparisons that emerged significant as per the Tukey's HSD test)
Legend: Perceptron Best (PB); Perceptron Worst (PW); XLE Best (XB); XLE Worst (XW); OpenCCG baseline models 1 to 3 (C1 ... C3)