E-rating Machine Translation

Kristen Parton[1] **Joel Tetreault**[2] **Nitin Madnani**[2] **Martin Chodorow**[3]
[1]Columbia University, NY, USA
kristen@cs.columbia.edu
[2]Educational Testing Service, Princeton, NJ, USA
{jtetreault, nmadnani}@ets.org
[3]Hunter College of CUNY, New York, NY, USA
martin.chodorow@hunter.cuny.edu

Abstract

We describe our submissions to the WMT11 shared MT evaluation task: MTeRater and MTeRater-Plus. Both are machine-learned metrics that use features from e-rater®, an automated essay scoring engine designed to assess writing proficiency. Despite using only features from e-rater and without comparing to translations, MTeRater achieves a sentence-level correlation with human rankings equivalent to BLEU. Since MTeRater only assesses fluency, we build a meta-metric, MTeRater-Plus, that incorporates adequacy by combining MTeRater with other MT evaluation metrics and heuristics. This meta-metric has a higher correlation with human rankings than either MTeRater or individual MT metrics alone. However, we also find that e-rater features may not have significant impact on correlation in every case.

1 Introduction

The evaluation of machine translation (MT) systems has received significant interest over the last decade primarily because of the concurrent rising interest in statistical machine translation. The majority of research on evaluating translation quality has focused on metrics that compare translation hypotheses to a set of human-authored reference translations. However, there has also been some work on methods that are not dependent on human-authored translations.

One subset of such methods is task-based in that the methods determine the quality of a translation in terms of how well it serves the need of an extrinsic task. These tasks can either be downstream NLP tasks such as information extraction (Parton et al., 2009) and information retrieval (Fujii et al., 2009) or human tasks such as answering questions on a reading comprehension test (Jones et al., 2007).

Besides extrinsic evaluation, there is another set of methods that attempt to "learn" what makes a good translation and then predict the quality of new translations without comparing to reference translations. Corston-Oliver et al. (2001) proposed the idea of building a decision tree classifier to simply distinguish between machine and human translations using language model (LM) and syntactic features. Kulesza and Shieber (2004) attempt the same task using an support vector machine (SVM) classifier and features derived from reference-based MT metrics such as WER, PER, BLEU and NIST. They also claim that the confidence score for the classifier being used, if available, may be taken as an estimate of translation quality. Quirk (2004) took a different approach and examined whether it is possible to explicitly compute a confidence measure for each translated sentence by using features derived from both the source and target language sides. Albrecht and Hwa (2007a) expanded on this idea and conducted a larger scale study to show the viability of regression as a sentence-level metric of MT quality. They used features derived from several other reference-driven MT metrics. In other work (Albrecht and Hwa, 2007b), they showed that one could substitute translations from other MT systems for human-authored reference translations and derive the regression features from them.

Gamon et al. (2005) build a classifier to distinguish machine-generated translations from human

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 108–115,
Edinburgh, Scotland, UK, July 30–31, 2011.

ones using *fluency-based* features and show that by combining the scores of this classifier with LM perplexities, they obtain an MT metric that has good correlation with human judgments but not better than the baseline BLEU metric.

The fundamental questions that inspired our proposed metrics are as follows:

- Can an operational English-proficiency measurement system, built with absolutely no forethought of using it for evaluation of translation quality, actually be used for this purpose?
- Obviously, such a system can only assess the fluency of a translation hypothesis and not the adequacy. Can the features derived from this system then be combined with metrics such as BLEU, METEOR or TERp—measures of adequacy—to yield a metric that performs better?

The first metric we propose (MTeRater) is an SVM ranking model that uses features derived from the ETS e-rater® system to assess fluency of translation hypotheses. Our second metric (MTeRater-Plus) is a meta-metric that combines MTeRater features with metrics such as BLEU, METEOR and TERp as well as features inspired by other MT metrics.

Although our work is intimately related to some of the work cited above in that it is a trained regression model predicting translation quality at the sentence level, there are two important differences:

1. We do not use *any* human translations – reference or otherwise – for MTeRater, not even when training the metric. The classifier is trained using human judgments of translation quality provided as part of the shared evaluation task.
2. Most of the previous approaches use feature sets that are designed to capture *both* translation adequacy and fluency. However, MTeRater uses only fluency-based features.

The next section provides some background on the e-rater system. Section 3 presents a discussion of the differences between MT errors and learner errors. Section 4 describes how we use e-rater to build our metrics. Section 5 outlines our experiments and Section 5 discusses the results of these experiments. Finally, we conclude in Section 6.

2 E-rater

E-rater is a proprietary automated essay scoring system developed by Educational Testing Service (ETS) to assess writing quality.[1] The system has been used operationally for over 10 years in high-stakes exams such as the GRE and TOEFL given its speed, reliability and high agreement with human raters.

E-rater combines 8 main features using linear regression to produce a numerical score for an essay. These features are grammar, usage, mechanics, style, organization, development, lexical complexity and vocabulary usage. The grammar feature covers errors such as sentence fragments, verb form errors and pronoun errors (Chodorow and Leacock, 2000). The usage feature detects errors related to articles (Han et al., 2006), prepositions (Tetreault and Chodorow, 2008) and collocations (Futagi et al., 2008). The mechanics feature checks for spelling, punctuation and capitalization errors. The style feature checks for passive constructions and word repetition, among others. Organization and development tabulate the presence or absence of discourse elements and the length of each element. Finally, the lexical complexity feature details how complex the writer's words are based on frequency indices and writing scales, and the vocabulary feature evaluates how appropriate the words are for the given topic). Since many of the features are essay-specific, there is certainly some mismatch between what e-rater was intended for and the genres we are using it for in this experiment (translated news articles).

In our work, we separate e-rater features into two classes: sentence level and document level. The sentence level features consist of all errors marked by the various features for each sentence alone. In contrast, the document level features are an aggregation of the sentence level features for the entire document.

[1] A detailed description of e-rater is outside the scope of this paper and the reader is referred to (Attali and Burstein, 2006).

3 Learner Errors vs. MT Errors

Since e-rater is trained on human-written text and designed to look for errors in usage that are common to humans, one research question is whether it is even useful for assessing the fluency of machine translated text. E-rater is unaware of the translation context, so it does not look for common MT errors, such as untranslated words, mistranslations and deleted content words. However, these may get flagged as other types of learner errors: spelling mistakes, confused words, and sentence fragments.

Machine translations do contain learner-like mistakes in verb conjugations and word order. In an error analysis of SMT output, Vilar et al. (2006) report that 9.9% - 11.7% of errors made by a Spanish-English SMT system were incorrect word forms, including incorrect tense, person or number. These error types are also account for roughly 14% of errors made by ESL (English as a Second Language) writers in the Cambridge Learner Corpus (Leacock et al., 2010).

On the other hand, some learner mistakes are unlikely to be made by MT systems. The Spanish-English SMT system made almost no mistakes in idioms (Vilar et al., 2006). Idiomatic expressions are strongly preferred by language models, but may be difficult for learners to memorize ("kicked *a* bucket"). Preposition usage is a common problem in non-native English text, accounting for 29% of errors made by intermediate to advanced ESL students (Bitchener et al., 2005) but language models are less likely to prefer local preposition errors e.g., "he went *to* outside". On the other hand, a language model will likely not prevent errors in prepositions (or in other error types) that rely on long-distance dependencies.

4 E-rating Machine Translation

The MTeRater metric uses only features from e-rater to score translations. The features are produced directly from the MT output, with no comparison to reference translations, unlike most MT evaluation metrics (such as BLEU, TERp and METEOR).

An obvious deficit of MTeRater is a measure of adequacy, or how much meaning in the source sentence is expressed in the translation. E-rater was not developed for assessing translations, and the MTeRater metric never compares the translation to the source sentence. To remedy this, we propose the MTeRater-Plus meta-metric that uses e-rater features plus all of the hybrid features described below. Both metrics were trained on the same data using the same machine learning model, and differ only in their feature sets.

4.1 E-rater Features

Each sentence is associated with an e-rater sentence-level vector and a document-level vector as previously described and each column in these vectors was used a feature.

4.2 Features for Hybrid Models

We used existing automatic MT metrics as baselines in our evaluation, and also as features in our hybrid metric. The metrics we used were:

1. **BLEU** (Papineni et al., 2002): Case-insensitive and case-sensitive BLEU scores were produced using mteval-v13a.pl, which calculates smoothed sentence-level scores.

2. **TERp** (Snover et al., 2009): Translation Edit Rate plus (TERp) scores were produced using terp v1. The scores were case-insensitive and edit costs from Snover et al. (2009) were used to produce scores tuned for fluency and adequacy.

3. **METEOR** (Lavie and Denkowski, 2009): Meteor scores were produced using Meteor-next v1.2. All types of matches were allowed (exact, stem, synonym and paraphrase) and scores tuned specifically to rank, HTER and adequacy were produced using the "-t" flag in the tool.

We also implemented features closely related to or inspired by other MT metrics. The set of these auxiliary features is referred to as "Aux".

1. **Character-level statistics**: Based on the success of the i-letter-BLEU and i-letter-recall metrics from WMT10 (Callison-Burch et al., 2010), we added the harmonic mean of precision (or recall) for character n-grams (from 1 to 10) as features.

2. **Raw n-gram matches**: We calculated the precision and precision for word n-grams (up to n=6) and added each as a separate feature (for a total of 12). Although these statistics are also calculated as part of the MT metrics above, breaking them into separate features gives the model more information.

3. **Length ratios**: The ratio between the lengths of the MT output and the reference translation was calculated on a character level and a word level. These ratios were also calculated between the MT output and the source sentence.

4. **OOV heuristic**: The percentage of tokens in the MT that match the source sentence. This is a low-precision heuristic for counting out of vocabulary (OOV) words, since it also counts named entities and words that happen to be the same in different languages.

4.3 Ranking Model

Following (Duh, 2008), we represent sentence-level MT evaluation as a ranking problem. For a particular source sentence, there are N machine translations and one reference translation. A feature vector is extracted from each {source, reference, MT} tuple. The training data consists of sets of translations that have been annotated with relative ranks. During training, all ranked sets are converted to sets of feature vectors, where the label for each feature vector is the rank. The ranking model is a linear SVM that predicts a relative score for each feature vector, and is implemented by SVM-rank (Joachims, 2006). When the trained classifier is applied to a set of N translations for a new source sentence, the translations can then be ranked by sorting the SVM scores.

5 Experiments

All experiments were run using data from three years of previous WMT shared tasks (WMT08, WMT09 and WMT10). In these evaluations, annotators were asked to rank 3-5 translation hypotheses (with ties allowed), given a source sentence and a reference translation, although they were only required to be fluent in the target language.

Since e-rater was developed to rate English sentences only, we only evaluated tasks with English as the target language. All years included source languages French, Spanish, German and Czech. WMT08 and WMT09 also included Hungarian and multisource English. The number of MT systems was different for each language pair and year, from as few as 2 systems (WMT08 Hungarian-English) to as many as 25 systems (WMT10 German-English). All years had a newswire testset, which was divided into stories. WMT08 had testsets in two additional genres, which were not split into documents.

All translations were pre-processed and run through e-rater. Each document was treated as an essay, although news articles are generally longer than essays. Testsets that were not already divided into documents were split into pseudo-documents of 20 contiguous sentences or less. Missing end of sentence markers were added so that e-rater would not merge neighboring sentences.

6 Results

For assessing our metrics prior to WMT11, we trained on WMT08 and WMT09 and tested on WMT10. The metrics we submitted to WMT11 were trained on all three years. One criticism of machine-learned evaluation metrics is that they may be too closely tuned to a few MT systems, and thus not generalize well as MT systems evolve or when judging new sets of systems. In this experiment, WMT08 has 59 MT systems, WMT09 has 70 different MT systems, and WMT10 has 75 different systems. Different systems participate each year, and those that participate for multiple years often improve from year to year. By training and testing across years rather than within years, we hope to avoid overfitting.

To evaluate, we measure correlation between each metric and the human annotated rankings according to (Callison-Burch et al., 2010): Kendall's tau is calculated for each language pair and the results are averaged across language pairs. This is preferable to averaging across all judgments because the number of systems and the number of judgments vary based on the language pair (e.g., there were 7,911 ranked pairs for 14 Spanish-English systems, and 3,575 ranked pairs for 12 Czech-English systems).

It is difficult to calculate the statistical significance of Kendall's tau on these data. Unlike the

Source language	cz	de	es	fr	avg
Individual Metrics & Baselines					
MTeRater	.32	.31	.19	.23	.26
bleu-case	.26	.27	.28	.22	.26
meteor-rank	.33	.36	.33	.27	.32
TERp-fluency	.30	.36	.28	.28	.30
Meta-Metric & Baseline					
BMT+Aux+MTeRater	.38	.42	.37	.38	.39
BMT	.35	.40	.35	.34	.36
Additional Meta-Metrics					
BMT+LM	.36	.41	.36	.36	.37
BMT+MTeRater	.38	.42	.36	.38	.38
BMT+Aux	.38	.41	.38	.37	.39
BMT+Aux+LM	.39	.42	.38	.36	.39

Table 1: Kendall's tau correlation with human rankings. BMT includes bleu, meteor and TERp; Aux includes auxiliary features. BMT+Aux+MTeRater is MTeRater-Plus.

Metrics MATR annotations (Przybocki et al., 2009), (Peterson and Przybocki, 2010), the WMT judgments do not give a full ranking over all systems for all judged sentences. Furthermore, the 95% confidence intervals of Kendall's tau are known to be very large (Carterette, 2009) – in Metrics MATR 2010, the top 7 metrics in the paired-preference single-reference into-English track were within the same confidence interval.

To compare metrics, we use McNemar's test of paired proportions (Siegel and Castellan, 1988) which is more powerful than tests of independent proportions, such as the chi-square test for independent samples.[2] As in Kendall's tau, each metric's relative ranking of a translation pair is compared to that of a human. Two metrics, A and B, are compared by counting the number of times both A and B agree with the human ranking, the number of times A disagrees but B agrees, the number of times A agrees but B disagrees, and the number of times both A and B disagree. These counts can be arranged in a 2 x 2 contingency table as shown below.

	A agrees	A disagrees
B agrees	a	b
B disagrees	c	d

McNemar's test determines if the cases of mismatch in agreement between the metrics (cells b and c) are symmetric or if there is a significant difference in favor of one of the metrics showing more agreement with the human than the other. The two-tailed probability for McNemar's test can be calculated using the binomial distribution over cells b and c.

6.1 Reference-Free Evaluation with MTeRater

The first group of rows in Table 1 shows the Kendall's tau correlation with human rankings of MTeRater and the best-performing version of the three standard MT metrics. Even though MTeRater is blind to the MT context and does not use the source or references at all, MTeRater's correlation with human judgments is the same as case-sensitive bleu (bleu-case). This indicates that a metric trained to assess English proficiency in non-native speakers is applicable to machine translated text.

6.2 Meta-Metrics

The second group in Table 1 shows the correlations of our second metric, MTeRater-Plus (BMT+Aux+MTeRater), and a baseline meta-metric (BMT) that combined BLEU, METEOR and TERp. MTeRater-Plus performs significantly better than BMT, according to McNemar's test.

We also wanted to determine whether the e-rater features have any significant impact when used as part of meta-metrics. To this end, we first created two variants of MTeRater-Plus: one that removed the MTeRater features (BMT+Aux) and another that replaced the MTeRater features with the LM likelihood and perplexity of the sentence (BMT+Aux+LM).[3] Both models perform as well as MTeRater-Plus, i.e., adding additional fluency features (either LM scores or MTeRater) to the BMT+Aux meta-metric has no significant impact.

To determine whether this was generally the case, we also created two variants of the BMT baseline meta-metric that added fluency features to it: one in the form of LM scores (BMT+LM) and another in the form of the MTeRater score (BMT+MTeRater). Based on McNemar's test, both models are significantly better than BMT, indicating that these reference-free fluency features indeed capture an aspect of translation quality that is absent from the standard MT metrics. However, there is no signficant difference between the two variants of BMT.

[2] See http://faculty.vassar.edu/lowry/propcorr.html for an excellent description.

[3] The LM was trained on English Gigaword 3.0, and was provided by WMT10 organizers.

1) Ref: Gordon Brown has discovered yet another hole to fall into; his way out of it remains the same MT+: Gordon Brown discovered a new hole in which to sink; even if it resigned, the position would not change. *Errors: None marked* MT-: Gordon Brown has discovered a new hole in which could, Even if it demissionnait, the situation does not change not. *Errors: Double negative, spelling, preposition*
2) Ref: Jancura announced this in the Twenty Minutes programme on Radiozurnal. MT+: Jancura said in twenty minutes Radiozurnal. *Errors: Spelling* MT-: He said that in twenty minutes. *Errors: none marked*

Table 2: Translation pairs ranked correctly by MTeRater but not bleu-case (1) and vice versa (2).

6.3 Discussion

Table 2 shows two pairs of ranked translations (MT+ is better than MT-), along with some of the errors detected by e-rater. In pair 1, the lower-ranked translation has major problems in fluency as detected by e-rater, but due to n-gram overlap with the reference, bleu-case ranks it higher. In pair 2, MT- is more fluent but missing two named entities and bleu-case correctly ranks it lower.

One disadvantage of machine-learned metrics is that it is not always clear which features caused one translation to be ranked higher than another. We did a feature ablation study for MTeRater which showed that document-level collocation features significantly improve the metric, as do features for sentence-level preposition errors. Discourse-level features were harmful to MT evaluation. This is unsurprising, since MT sentences are judged one at a time, so any discourse context is lost.

Overall, a metric with only document-level features does better than one with only sentence-level features due to data sparsity – many sentences have no errors, and we conjecture that the document-level features are a proxy for the quality of the MT system. Combining both document-level and sentence-level e-rater features does significantly better than either alone. Incorporating document-level features into sentence-level evaluation had one unforeseen effect: two identical translations can get different scores depending on how the rest of the document is translated. While using features that indicate the relative quality of MT systems can improve overall correlation, it fails when the sentence-level signal is not strong enough to overcome the prior belief.

7 Conclusion

We described our submissions to the WMT11 shared evaluation task: MTeRater and MTeRater-Plus. MTeRater is a fluency-based metric that uses features from ETS's operational English-proficiency measurement system (e-rater) to predict the quality of any translated sentence. MTeRater-Plus is a meta-metric that combines MTeRater's fluency-only features with standard MT evaluation metrics and heuristics. Both metrics are machine-learned models trained to rank new translations based on existing human judgments of translation.

Our experiments showed that MTeRater, by itself, achieves a sentence-level correlation as high as BLEU, despite not using reference translations. In addition, the meta-metric MTeRater-Plus achieves higher correlations than MTeRater, BLEU, METEOR, TERp as well as a baseline meta-metric combining BLEU, METEOR and TERp (BMT). However, further analysis showed that the MTeRater component of MTeRater-Plus does not contribute significantly to this improved correlation. However, when added to the BMT baseline meta-metric, MTeRater does make a significant contribution.

Our results, despite being a mixed bag, clearly show that a system trained to assess English-language proficiency can be useful in providing an indication of translation fluency even outside of the specific WMT11 evaluation task. We hope that this work will spur further cross-pollination between the fields of MT evaluation and grammatical error detection. For example, we would like to explore using MTeRater for confidence estimation in cases where reference translations are unavailable, such as task-oriented MT.

Acknowledgments

The authors wish to thank Slava Andreyev at ETS for his help in running e-rater. This research was supported by an NSF Graduate Research Fellowship for the first author.

References

Joshua Albrecht and Rebecca Hwa. 2007a. A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. In *Proceedings of ACL*.

Joshua Albrecht and Rebecca Hwa. 2007b. Regression for Sentence-Level MT Evaluation with Pseudo References. In *Proceedings of ACL*.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3).

John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on esl student writing. *Journal of Second Language Writing*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ben Carterette. 2009. On rank correlation and the distance between rankings. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 436–443, New York, NY, USA. ACM.

Martin Chodorow and Claudia Leacock. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 140–147.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 148–155.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 191–194, Stroudsburg, PA, USA. Association for Computational Linguistics.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2009. Evaluating Effects of Machine Translation Accuracy on Cross-lingual Patent Retrieval. In *Proceedings of SIGIR*, pages 674–675.

Yoko Futagi, Paul Deane, Martin Chodorow, and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21:353–367.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT Evaluation Without Reference Translations: Beyond Language Modeling. In *Proceedings of the European Association for Machine Translation (EAMT)*.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2):115–129.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, pages 217–226.

Douglas Jones, Martha Herzog, Hussny Ibrahim, Arvind Jairam, Wade Shen, Edward Gibson, and Michael Emonts. 2007. ILR-Based MT Comprehension Test with Multi-Level Questions. In *HLT-NAACL (Short Papers)*, pages 77–80.

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 (TMI)*.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23:105–115, September.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, Wei Xu, and Sibel Yaman. 2009. Who, What, When, Where, Why? Comparing Multiple Approaches to the Cross-Lingual 5W Task. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 423–431.

Kay Peterson and Mark Przybocki. 2010. Nist 2010 metrics for machine translation evaluation (metricsmatr10) official release of results. `http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2010/results`.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challenge–overview, methodology, metrics, and results. *Machine Translation*, 23:71–103, September.

Christopher Quirk. 2004. Training a Sentence-level Machine Translation Confidence Measure. In *Proceedings of LREC*.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, 2 edition.

Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard 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*, StatMT '09, pages 259–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joel Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 865–872.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May.