

Applying Machine Translation Metrics to Student-Written Translations

Lisa N. Michaud

Computer Science Department
Merrimack College
North Andover, MA, USA
michaudl@merrimack.edu

Patricia Ann McCoy

Language Department
Universidad de las Americas Puebla
Puebla, Mexico
patricia.mccoy@udlap.mx

Abstract

This paper discusses preliminary work investigating the application of Machine Translation (MT) metrics toward the evaluation of translations written by human novice (student) translators. We describe a study in which we apply the metric TERp (Translation Edit Rate Plus) to a corpus of student-written translations from Spanish to English and compare the judgments of TERp against assessments provided by a translation instructor.

1 Introduction

Extensive work in the field of Computational Linguistics has focused on the development of gold-standard metrics to automatically judge the accuracy of machine-generated translations. We are exploring whether these metrics, or a modified version thereof, may be applied to the translations generated by human novices.

While Machine Translation (MT) metrics have been shown to perform poorly when evaluating human-written translations due to their lack of tolerance for the high level of variation in human-written work, it is our belief that novice student translators keep much closer to the source text, and therefore will be easier to assess using automatic metrics.

Initial motivation for this work comes from developing the King Alfred translation environment (Michaud, 2008) supporting students of Anglo-Saxon English translating sentences into Modern English. Criticisms of the application of computational tools toward language learning have often highlighted the reality that the mainstays of modern

language teaching—dialogue and a focus on communicative goals over syntactic perfectionism—parallel the shortcomings of a computational environment. While efforts continue to extend the state of the art toward making the computer a conversational partner, they nevertheless often fall short of providing the language learner with learning assistance in the task of communicative competence that can make a real difference within or without the classroom. The modern learner of ancient or “dead” languages, however, has fundamentally different needs; the focus is on translation from source texts into the learner’s L1. An initial goal, therefore, was to provide the King Alfred system with ability to automatically judge and respond to student translations given a single instructor-provided reference.

The potential applications of this work extend beyond the learning of dead languages, however; translation skills in modern languages (until the field of MT reaches its full potential) are still needed for providing readers with access to cross-lingual information. The ability to assist translation instruction via a tutoring system outside of the classroom, or to assess translator skill automatically, is therefore greatly desirable.

The study described in this paper therefore focuses on a corpus of learner-written translations from a Spanish-English translation course; in Section 6 we discuss how these results may compare to those using a corpus of translations from Anglo-Saxon, which is one of our future tasks.

Reference	however ,	under	certain contexts a translator	may	intentionally	strive to	produce a literal translation .
		S		S		P	
Hyp After Shifts	however ,	in	certain contexts a translator	can	intentionally	try to	produce a literal translation .

Figure 1: Output from the TERp system.

2 Evaluating Student-Written Translations Using TERp

A primary challenge facing the assessment of translation fitness is the abstract nature of the definition of fitness with respect to the translating task. Most people approach this definition with two major foci: *fluency* (is it well-formed?) and *fidelity* (does it convey original meaning?) (Hovy et al., 2002). There are also stylistic concerns; translation can be defined as “rendering the meaning of a text into another language in the way the author intended the text” (Newmark, 1988)—and intention is difficult to precisely define. None of these viewpoints dictates that there exists only one way to write a translation.

We were drawn to the TERp (Translation Edit Rate Plus) translation metric (Snover et al., 2009) for our initial study because of its particular approach toward capturing this multiplicity of correct translations. Other metrics have addressed this issue; BLEU (Papineni et al., 2002), for example, uses multiple reference translations, in the hopes of capturing diversity through using diverse sources. The creators of TERp, however, create an alignment between reference and hypothesis strings in which direct matches are not required; they acknowledge synonymy by leveraging WordNet synsets (Fellbaum, 1998; Princeton University, 2010), in addition to using a stemmer, and a phrase table to handle probabilistic phrasal substitution. TERp also allows for words or phrases to be shifted into a different position, which nicely accounts for flexibility in terms of prepositional phrase or adverb placement or to handle modifiers that can take multiple forms.

There has been some dismissal of the appropriateness of MT metrics for Computer-Aided Language Learning (CALL) applications (cf. (McCarthy, 2006)) due to the fact that they often provide a holistic score comparing the hypothesis translation to one or more reference translations without identifying the source and nature of the differences. However, the output of TERp also includes more than a

holistic score; there is complete documentation of the alignments, with tags identifying the “edits” required to line up the hypothesis with the reference, as seen in Figure 1. This is an excellent resource from the perspective of translation pedagogy. While the METEOR system (Agarwal and Lavie, 2008) also uses WordNet synonymy and a stemmer to similar purpose, we believe that TERp comes the closest to embracing the multiplicity of translation paths while at the same time flagging issues of fundamental concern in a pedagogical application of MT metrics.

3 Related Work

Other environments seeking to support student translations have addressed the issue of automatically determining translation accuracy. A English-Chinese translation environment described by (Wang and Seneff, 2007; Xu and Seneff, 2008) presents students with L1 sentences to translate into L2 speech. Because many of its L1 sentences are automatically generated, there is no possibility of prestored reference translations, so the system uses speech recognition to obtain the L2 sentence, and then parses both the English and Chinese sentences into a common interlingual representation in order to compare for accuracy. The authors report a high level of agreement between the system’s judgments on translation acceptability compared to that of a human expert, but unfortunately, the system cannot give a finer-grained judgment on student performance than *accept* or *reject*.

Another English-Chinese system is described by (Shei and Pain, 2002), creators of TMT, the Translation Method Tutor. In this case, students are translating from their L2 (English) into their L1 (Chinese) using source sentences from Jane Austen’s *Pride and Prejudice*, each selected to practice a particular linguistic structure. Students’ translations are matched against four possible reference translations: word-to-word (MT generated), literal (MT-

generated and then post-processed to obey word order rules), semantic (professional translations), and communicative (done by the authors), and the feedback provided to the student includes which translation she matched most closely and a lesson on how to deal with the structure at hand. Comparisons between the student translation and the references look at strict similarity and are heavily influenced by word selection rather than structure.

The Translator Choice Program (McCarthy, 2006) focuses on French-English translation for native English speakers. It presents passages in the L2 (French) and asks students to look at five candidate English translations written by students in previous years. Students either pick the best translation or rank them, and are scored in how similar their judgment is to that of their instructor. This system does not attempt, therefore, to handle novel translations performed by the student.

4 A Corpus of Student-Written Translations

In Spring 2012, we solicited participation from students of a Spanish-English translation course. In this course, students are asked to translate a sequence of articles in both Spanish and English, typically alternating the source language. The articles address varied topics from financial advice to current news. Thirteen students (both native English speakers and native Spanish speakers) opted to have their semester’s work collected as part of our study. Reference translations were provided for the entire corpus by the instructor of the course.

For our initial study, we have focused on only the Spanish-to-English translations, as many aspects of the metric we used focus on comparing an English hypothesis sentence against an English reference sentence. This yielded a total of 2,982 sentences. They are described in Table 1.

Table 1: Our Student-Written Translation Corpus.

Number of Subjects	13
Native English Speakers	3
Native Spanish Speakers	10
Number of Articles Translated	11
Average Number of Sentences per Article	28
Total Translated Sentences	2982

5 Comparing Human Judgments to TERp

Before analyzing the translations with the MT metric, we post-processed the corpus to create an alignment between student translations, source sentences, and the instructor reference. One of the challenges we faced in this step is that these students, unlike an MT system, are actively encouraged to recognize the stylistic differences between English and Spanish native writing in terms of sentence brevity. The students therefore sometimes create translations that do not always perfectly match sentence boundaries of the source text; in some cases a single Spanish sentence has been split into multiple English sentences (following a general principle that English native speakers typically use more concise utterances), but sometimes also the opposite occurs, where two source sentences are combined into one translated sentence. While most translations (more than 99%) did obey source sentence boundaries, for alignment purposes whenever a sentence was split both target sentences were concatenated into a single string (including the end-of-sentence punctuation, which is ignored by TERp)¹ for comparison against the reference. Where the student had merged two sentences, the clauses were separated at an appropriate boundary and treated as separate utterances. The instructor-provided references obeyed a 1:1 correspondence between source and target sentences.

Our entire corpus has been graded using the TERp-A variant, with unchanged parameters². The TERp system scores sentences on an interval of [0,100], where a lower score indicates closer agreement to the reference translation, and 100 indicates no agreement; for the ease of our human grader, we normalized the TERp scores to invert the scale and better match a human-intuitive scale of 100 for excellence and 0 for no agreement.

Figure 2 illustrates for those subjects submitting more than three assignments to the study the longitudinal progress of the average TERp score (inverted) across the sentences in each assignment given over

¹The insertion of a connector, such as ‘and,’ to form a unified sentence could be penalized by TERp, so it was avoided; the alternative to avoid penalty would be to include whatever connector the original author used, but this would not be available during automated analysis later.

²As will be discussed in Section 6, a future goal is to tune the parameters for performance on this data.

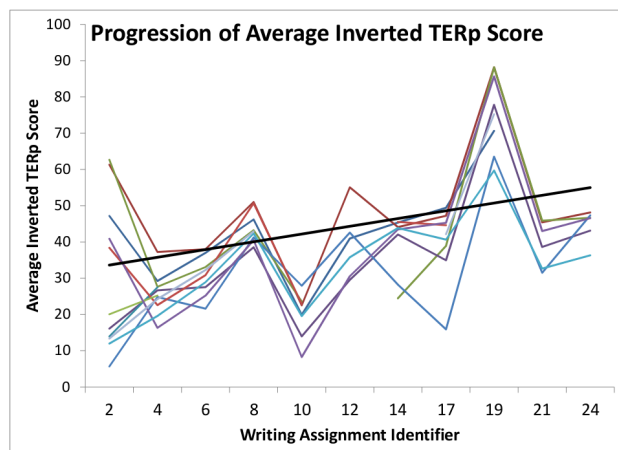


Figure 2: TERp scores across development.

the term. Although there were clearly a couple of assignments that were very challenging to all of the students, the trend line shown indicates that the scores were rising over the course of the semester.

We have also collected instructor-assigned scores on a portion of our corpus in order to compare them against these TERp scores. An example of the rubric used by the instructor as part of her regular grading practices in the course is shown in Table 2. Each of these categories receive a score from 0-10 with 10 being *excellent*, 9 *good*, 8 *satisfactory*, and 0-7 *deficient*.

Table 2: Instructor rubric for assigning sentence grades.

Conveys original meaning	55%
Written in natural language	20%
Uses appropriate vocabulary	10%
Written in accurate language	15%

Our preliminary study has yielded some interesting results. The Pearson correlation between the two sets of scores is $r=0.232236$, which on a $[-1,1]$ interval indicates weak positive correlation. But if TERp does not have significant agreement with the students' instructor, what is the source of the disagreement? One illustration of this disagreement is the distribution of the grades; Figure 3 shows that the instructor's grades are heavily slanted toward the high end of the scale, with 42% of the sentences

scored receiving a grade of 90 or higher; TERp, by contrast, gave very few sentences higher normalized accuracy scores. This is most likely due to the instructor's heavy emphasis placed on communicative rather than syntactic accuracy, as shown in the rubric. We are in the process of rescoring the corpus with a revised rubric that places stronger emphasis on syntactic accuracy.

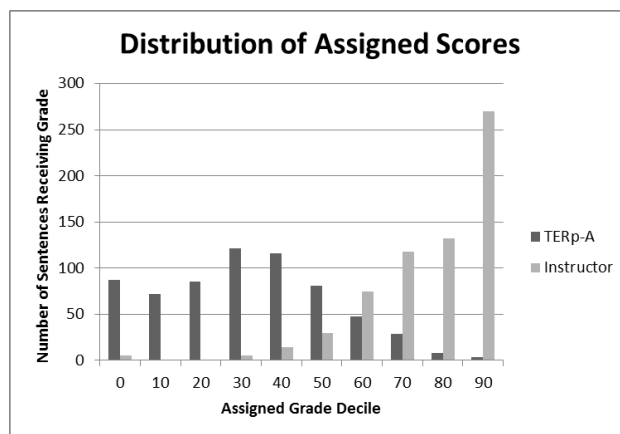


Figure 3: TERp score distribution compared against the human expert.

While TERp has already been evaluated in terms of its correlation to human judgment, this has not been done before with learner-written sentences³. We also performed an analysis of a randomized sample of individual sentences with a particular focus on the four edits designed to accommodate divergence but equivalence (or near equivalence): phrase equivalency, stemming, synonymy, and shifts. Our pilot study results indicate that TERp's identified edits have very high precision: 100% for the stemmer, which is to be expected, but also 92% for appropriate shifts, 89% for synonymy, and 83% for phrase equivalency. In recall, the edits performed less well; for example, synonymy achieved a recall of only 65%. This is possibly a limitation of the synset resource.

6 Conclusion and Future Work

We have seen that TERp's identification of the source and nature of divergences between a student

³The word *learner* here refers to the fact that the writer is a student of translation, not to whether he or she is writing in an L2.

translation and a teacher's reference translation is reliable; it correctly identifies the nature of the divergence from the reference in a high percentage of cases. This can provide a tutoring environment with sufficient information to address the translation's problems in feedback to the student, and indicates that holistic scores will be much more correlated with human scores that place equal emphasis on syntactic quality. A future version of the King Alfred system will use these error identifications to drive its feedback.

Once the rescoring of the corpus with an emphasis on syntactic accuracy is complete, further work will include tuning the TERp parameters for higher performance on the student corpus, with the aim of greatly improving the correlation of the scores.

We are also looking at post-processing TERp's scores so that certain divergences are not penalized. There is a *cost* associated with the edits that represent mismatches between the reference and hypothesis texts. While the idea of flexible phrase order, and the equality of synonym choice or phrase choice is captured by the metric, the application of such edits worsens the grade of the translation. We believe that stemming and substitution, deletion, or insertion should be penalized, but that synonymy, phrase matches, and shifts should be *free of charge*; those costs will therefore be added back into the final score.

As part of our larger investigation, we will continue to evaluate the applicability of machine translation metrics in general to the learner translation problem. The Mult-Eval suite of metrics (Clark et al., 2011) is a short term target, and iBLEU (Madnani, 2011) may provide useful data for a pedagogical context.

With a recent addition of 14 more subjects, we would also like to do an investigation of whether the performance of an MT metric is affected by whether the novice translator is translating L1→L2, or L2→L1. English native speakers are a minority in our subject pool, but with doubling the size of our corpus, we may be able to explore this more reliably.

One of our other interests going forward is to accommodate the distinct errors made by a very novice human translator. One such error is a tendency to fall prey to false cognates or *faux amis*—false friends, words that look similar (like Spanish *em-*

barazada and English *embarrassed*) that have significantly different meanings (*embarazada*, for example, meaning “pregnant”). We have a working hypothesis that student translators are often misled by these similar-looking words. We are currently working to automatically extract potential *faux amis* from parallel Spanish/English dictionaries with the hope of augmenting TERp's ability to align parallel elements between the student and reference translation. We are leveraging the spellcheck algorithm *Hunspell* to identify the similarly-spelled words.

Finally, it is our intention to do a comparative study between evaluating learner translations from modern languages and learner translations from ancient languages such as Anglo-Saxon. One challenge that may arise is that many ancient languages such as Anglo-Saxon are morphologically rich and therefore not strict word order languages; the source text will be fluid with its own order and this may introduce more diversity than in a modern language translation even among novice translators.

Acknowledgments

We wish to sincerely thank the students who have volunteered to share their semester's work with us for the purpose of this study. We would also like to thank the reviewers for their helpful comments and additional references.

References

- Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio, June. ACL.
- Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 176–181. ACL.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- Eduard Hovy, Margaret Kine, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17:43–75.

- Hunspell: open source spell checking, stemming, morphological analysis and generation under GPL, LGPL or MPL licenses. Website. <http://hunspell.sourceforge.net/> Accessed February 2013.
- Nitin Madnani. 2011. iBLEU: Interactively debugging and scoring statistical machine translation systems. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing, ICSC '11*, pages 213–214, Washington, DC, USA. IEEE Computer Society.
- Brian McCarthy. 2006. Tutoring translation skills: Reflections on a computer-managed teaching-learning-research triangle. *CALL-EJ Online*, 7(2), January.
- Lisa N. Michaud. 2008. King Alfred: A translation environment for learners of anglo-saxon english. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications, an ACL-HLT '08 Workshop*, pages 19–26, Columbus, Ohio, June. ACL.
- Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall International, New York.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, July 6-12. ACL.
- Princeton University. 2010. WordNet. Website. <http://wordnet.princeton.edu> Accessed July 2011.
- Chi-Chiang Shei and Helen Pain. 2002. Computer-assisted teaching of translation methods. *Literary & Linguistic Computing*, 17(3):323–343.
- Matthew Snover, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? exploring different judgments with a tunable MT metric. In *Proceedings of the EACL 2009 Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, March 30-31. ACL.
- Chao Wang and Stephanie Seneff. 2007. Automatic assessment of student translations for foreign language tutoring. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 468–475, Rochester, NY, April 22-27. ACL.
- Yushi Xu and Stephanie Seneff. 2008. Mandarin learning using speech and language technologies: A translation game in the travel domain. In *Proceedings of the 6th International Symposium on Chinese Spoken Language Processing (ISCSLP08)*, Kunming, China, December.