Identifying Common Challenges for Human and Machine Translation: A Case Study from the GALE[1] Program

Lauren Friedman
Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104 USA
lf@ldc.upenn.edu

Stephanie Strassel
Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104 USA
strassel@ldc.upenn.edu

Abstract

The dramatic improvements shown by statistical machine translation systems in recent years clearly demonstrate the benefits of having large quantities of manually translated parallel text for system training and development. And while many competing evaluation metrics exist to evaluate MT technology, most of those methods also crucially rely on the existence of one or more high quality human translations to benchmark system performance. Given the importance of human translations in this framework, understanding the particular challenges of human translation-for-MT is key, as is comprehending the relative strengths and weaknesses of human versus machine translators in the context of an MT evaluation. Vanni (2000) argued that the metric used for evaluation of competence in human language learners may be applicable to MT evaluation; we apply similar thinking to improve the prediction of MT performance, which is currently unreliable. In the current paper we explore an alternate model based upon a set of genre-defining features that prove to be consistently challenging for both humans and MT systems.

1 Introduction

The dramatic improvements shown by statistical machine translation systems in recent years clearly demonstrate the benefits of having large quantities of manually translated parallel text for system training and development. And while many competing evaluation metrics exist to evaluate MT technology, most of those methods also crucially rely on the existence of one or more high quality human translations to benchmark system performance. Given the importance of human translations in this framework, understanding the particular challenges of human translation-for-MT is key, as is comprehending the relative strengths and weaknesses of human versus machine translators in the context of an MT evaluation. Understanding such correlations will highlight areas for improvement on both ends and will inform the production of linguistic resources to support MT. Moreover, if a common set of challenges for HT and MT can be defined, source data in the human translation pipeline can be analyzed to better predict MT performance, thus enabling more informed data selection for system training and evaluation. Vanni (2000) argued that the metric used for evaluation of competence in human language learners may be applicable to MT evaluation; we apply similar thinking to improve the prediction of MT performance, which is currently unreliable. Similarly, Clifford (2004) reported that while Interagency Language Roundtable Scale (ILR) difficulty ratings are moderately useful, they are not consistently reliable in predicting MT performance. In the current paper we explore an alternate model based upon a set of genre-defining features that prove to be consistently challenging for both humans and MT systems.

[1] This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

2 The GALE MT Evaluation

Machine translation is one of three technology components for the DARPA GALE Program, which includes an annual MT evaluation administered by NIST (NIST 2007a). LDC creates training and test data for the GALE program, including gold standard translations for system evaluation. The GALE MT evaluation metric is edit distance, measured by HTER (human translation edit rate) (Snover 2006). HTER calculates the minimum number of changes required for highly-trained human editors to correct MT output so that it has the same meaning as the reference translation, through a process called "post-editing" (NIST 2007b).

GALE evaluation translation references undergo eight distinct phases of translation and quality control. First, raw translations are created by translators under contract to LDC. Translators follow LDC's GALE translation guidelines (LDC 2006a, LDC2006b) that include rules for handling idioms, non-standard grammar, misspellings, ambiguity, proper names and other common complexities, as well as genre-specific issues like URLs and emoticons in web text. While traditional translation tasks and some machine translation evaluation protocols would accept the output of the raw translation stage as adequate, GALE requires reference translations to undergo seven stages of additional annotation and quality control to correct errors, improve translation adequacy, add translation variants[2] , standardize proper nouns, verify technical terms and so on, with the ultimate goal of having the gold standard translations that are absolutely faithful to the source in terms of meaning, fluency, structure and style.

3 Common Translation Challenges

To begin exploring the relative difficulty of GALE evaluation data for human translators versus MT systems, we selected ~250 Chinese evaluation documents from the four genres currently evaluated in GALE: newswire (NW), broadcast news transcripts (BN), broadcast conversation/talk show transcripts (BC), and web text drawn primarily from weblogs and discussion groups (WB). We then analyzed the differences between raw translations and the final gold standard translations, and calculated a correction rate (CR) for each document .[3]

Since the CR score is not a standard metric, it is most useful as a relative score on a reasonably sized pool of data. We analyzed the CR on a per-genre rather than a per-file basis as CR is not considered a reliable or consistent at a high-level of granularity. Nevertheless, it provides a straightforward measurement particular to the context of the QC process, wherein changes of any kind can be construed as translation error correction. Analyzing the CR at the genre level indicates trends in the level of difficulty of a particular set of files.

Examination of CR results clearly reveals that some genres are more challenging for human translators (i.e., have a higher CR) than others. GALE MT system performance also varied widely by genre, as indicated by average HTER in the table below.

[2] Variants are introduced for all idioms (both literal and idiomatic translation are provided) and whenever the meaning of the source document is ambiguous or underspecified

[3] CR divides the number changes made during the seven-step QC process (excluding variant insertions) by the total number of words; CR = (diffs – variants) / words (in translation)

Genre	Features	Average CR (%)	Standard Deviation	Average combined HTER (% error) for GALE MT systems	Standard Deviation
NW	Structured Text	0.93	.0075	23.49	.0656
WB	Unstructured Text	1.55	.0085	27.72	.0731
BN	Structured Audio	2.12	.0106	25.88	.0689
BC	Unstructured Audio	2.51	.0105	33.07	.0661

Table 1: Human translation correction rate (CR) and HTER for GALE evaluation data in 4 genres.

While CR and HTER are not parallel for all genres, investigating the areas of correlation offers some insight into common difficulties for human translation and MT.

This section will examine that correlation where it exists, and offer some hypotheses to explain the areas where correlation is absent. Section 3.5 will offer a more detailed breakdown of the genre-specific features contributing to these results.

3.1 Newswire

Newswire consistently demonstrates the lowest level of difficulty for both human and machine translators. The structured text, high-frequency vocabulary, lack of intermediary transcription/ASR, standard punctuation and capitalization, standard grammar, and straightforward syntax of newswire data are all likely factors contributing to the relatively high level of accuracy for humans and MT systems. For humans, newswire data is also a more common genre; translators' high level of familiarity with newswire is certainly a reason for the low CRs for this genre. Similarly, it has been observed that new genres lead to a significant drop-off in MT performance, since more standard-genre data has been made available to train MT systems (Zhao 2004).

Newswire also shows the most direct correlation between CR and HTER; the average scores are by far the lowest in each metric, and the known features of newswire make it the easiest genre for both humans and MT systems.

3.2 Broadcast conversation

Broadcast conversation, in sharp contrast to newswire, is the most challenging genre for both machine and human translators. BC is particularly challenging given of its dual status as unstructured and audio. Because GALE targets an end-to-end transcription, translation and distillation engine, MT performance for audio genres (BN and BC) is confounded by the effect of using automatic speech recognition output as input to the translation task; whereas human translations are generated using source audio plus high-quality manual transcripts as input. However, the difficulty of unstructured data remains for the human translator, and creating fluent, meaning-accurate translations is a significant challenge. For instance, 77% of all Chinese BC evaluation files contain some speech overlap. Filled pauses like um and uh are frequent, affecting approximately 20% of all segments. Speech disfluencies and partial words occur at a much higher rate than in structured broadcast news -- 1.76 per 1000 tokens for BC compared to .35 per 1000 tokens for BN. Finally, the number of unique speakers per story segment averages 2.3 for BC. This genre is uniformly difficult for HT and MT as a whole, and on a per-file basis. BC files accounted for 7 out of the 10 worst HTER scores, and 9 out 10 worst CR scores.

3.3 Web

Web data presents a different set of challenges to both human translators and MT systems, with middling levels of accuracy across the board. Human translators consistently express frustration with web data, which – unique among the four tested genres – is generally produced informally and by amateurs, without regard to the standards of television (BC/BN) or printed news (NW). This can result in unpredictable errors, especially in MT, where frequent misspellings, absent punctuation, non-standard syntax, abbreviations, made-up words and sentence fragments in the source all present formidable challenges. These source problems

make web data significantly harder than newswire, but human translators are better equipped than MT systems to use context to achieve correct translations of poorly written segments. While HT performance is relatively consistent across documents, MT performance is highly erratic, as evidenced by the high standard deviation of HTER for this genre.

3.4 Broadcast news

CR-HTER correlation is weakest the structured audio data coming from Broadcast News. Although BN scores were middling for both MT systems and human translators, MT performance on BN was closer to NW, while HT performance on BN was closer to BC. Two possible factors contributing to this slight discrepancy include increased segment length for BN compared to BC (ratio of 2:1 in average number of translation tokens per segment) and high levels of semantic ambiguity compared to BC (signaled by the higher rate of translation variants added to the gold standards). First, long segments have been shown to have a negligible effect on MT performance, degrading by less than 1% (on fluency and adequacy metrics) when segment length is increased by a factor of 10 (Doddington 2002). Conversely, longer segments are more challenging for human translators, particularly because it is much harder preserve the original syntactic structure when translating long segments (which often also involve multiple embeddings and complex clauses). Second, while calculation of CR excluded translation variants (since they were not targeted in the raw translation task), when variants are added at subsequent stages they signal ambiguity in a segment, or even in an entire document. The prevalence of translation variants in the BN data raises several questions. Since gold standard translations include variants, MT systems are not penalized for translations that are reasonable interpretations of ambiguity in the source text. Human translators, however, may be especially challenged by regions with consistent ambiguity.

3.5 Genre variation

The following table summarizes the features identified as probable factors in MT and HT variation and correlation on a per-genre basis. A score of +1 indicates a high incidence of a given factor, -1 indicates an absence of a given factor, and 0 indicates neutral or average incidence of a given factor. The most difficult genres will have the highest totals.

	NW	BN	BC	Web
Unstructured text	-1	-1	+1	+1
Variable vocabulary[4]	-1	0	0	+1
Intermediary ASR	-1	+1	+1	-1
Non-standard punctuation[5]	-1	-1	-1	+1
Non-standard grammar	-1	0	+1	+1
Complex syntax/fragments	-1	0	+1	+1
Misspellings	-1	-1	-1	+1
Speech overlap[6]	-1	-1	+1	-1
Partial words[7]	-1	0	+1	-1
Semantic ambiguity[8]	+1	0	-1	+1
Totals:	**-8**	**-3**	**+3**	**+4**

Table 2: Matrix of features contributing to human and machine translation error in four genres.

[4] This indicates high rates of invented and misused words, which are common on the web, infrequent in television news programs, and virtually absent from published newswire.

[5] BN and BC transcripts are created and validated by humans to ensure that they are properly punctuated. Since harvested NW data is printed in major news publications, punctuation rules are generally followed. Punctuation is often absent, erratic, or misused in web data, and the original punctuation left basically intact.

[6] 77% of BC files have overlapping speech segments from multiple speakers, compared to 0% of BN files.

[7] Partial words, which are only relevant for audio data, were 5 times more frequent in BC data than BN data.

[8] Variants were introduced 101 times in BC data, 131 times in BN data, 207 times in NW data, and 212 times in web data.

With each factor weighted equally, we see that these totals reflect some general trends that emerged from the CR and HTER scores. NW is by far the easiest, which was the most consistent and clear result of the CR/HTER analysis. Based on this list of factors alone, the totals for BC, BN, and web data closely follow HTERs but do not line up exactly. BC data, although its score is high, comes closer to web data in difficulty than HTER/CR scores would suggest, and further investigation is necessary to identify additional factors at play.

4 Conclusions and future work

Although no single feature can predict either MT or HT performance, this combination of factors can help explain MT performance as it correlates to HT performance on the genre level, but further details could be gleaned from an analysis at the document level. Once this feature-based analysis reaches a suitable level of granularity, data selection procedures can target challenging data more directly, and constrained, informed selection of training data becomes more feasible. The proposed rubric of features must be quantified more exactly and weighted based on correlation with past performance so that it can offer a more accurate metric for prediction.

This paper aims to begin a conversation on the value of analyzing correlation and lack of correlation in human and machine translation error rates, but a great deal of work remains to be done. While CR approximates a measure of difficulty for the human translator, it is necessary to undertake a larger experiment in post-editing that uses an original human translation in place of MT, and keeps the final gold standard translation as the reference. Such an experiment is currently underway, and results are forthcoming. The scores from this experiment will allow HTER-to-HTER comparison of human performance versus machine performance on a per-file level, and the CR approximation will no longer be needed.

While additional experiments are required to test the hypothesized model presented herein, this paper aims to be a starting point – to demonstrate (1) that an analysis of correlation in human-machine translation performance must be pursued, and (2) that a feature-based metric may be viable pending further research and testing.

5 References

Clifford, Ray et al. 2004. “The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean”. Proceedings of Language Resources and Evaluation Conference, Lisbon.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proc. ARPA Workshop on Human Language Technology.

LDC. 2006a. GALE Arabic Translation Guidelines V2.3
http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Arabic_translation_guidelines_v2.3.pdf

LDC. 2006b. GALE Chinese Translation Guidelines V2.3
http://projects.ldc.upenn.edu/gale/Translation/specs/GALE_Chinese_translation_guidelines_v2.3.pdf

NIST 2007a. GALE 2007 Phase 2 Evaluation Plan Version 3.1
http://www.nist.gov/speech/tests/gale/2007/doc/GALE07_evalplan-v3.1.pdf

NIST 2007b. Post Editing Guidelines Version 3.0.2http://www.nist.gov/speech/tests/gale/2007/doc/GALEpostedit_guidelines-3.0.2.pdf

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas.

Vanni, Michelle, & Florence Reeder. 2000. How are you doing? A Look at MT Evaluation. In Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000,

ed. by John S. White, Lecture Notes in Artificial Intelligence #1934, 109-116. Berlin: Springer.

Zhao, Bing, Matthias Eck, Stephan Vogel 2004. Language Model Adaptation for Statistical Machine Translation with structured query models. COLING-2004