Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 117-120.
Association for Computational Linguistics.

Towards a Speech-to-Speech Machine Translation Quality Metric

Kurt Godden
Lockheed Martin Advanced Technology Laboratories
1 Federal Street
Camden, New Jersey 08102
kgodden@atl.lmco.com

Abstract

General characteristics of a pragmatic metric for the *production* evaluation of speech-to-speech translations are discussed. While these characteristics constrain the space of allowable metrics, infinite definition space remains from which to select and define any particular metric. The recommended characteistics are drawn from the author's experience as primary developer of a text-based translation quality metric used in a production environment. The primary contribution is that of strict category ordering and two meta-rules that reduce the variance in assignment of errors to categories.

1 Introduction

When we consider speech-to-speech (S2S) translation systems, several abstract models are possible.

In Model 1 (Figure 1) we treat the entire software system as a "black box," just recognizing that the input is a source language utterance (SLU) and the output is a target language utterance (TLU).

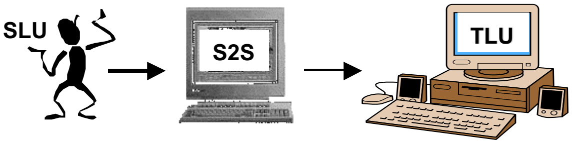

Figure 1. Model 1.

In Model 2 (Figure 2) we break the previous black box into several traditional components, reflecting typical language processing modules. The source language utterance is transformed to a source language text (SLT) by an automatic speech recognition (ASR) system. The SLT is then translated by a machine translation (MT) system to a target language text (TLT), which is in turn converted to the target language utterance by a text-to-speech (TTS) system.

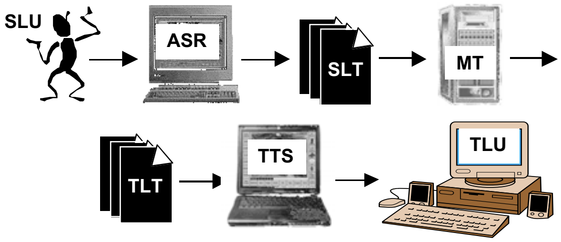

Figure 2. Model 2.

Model 3 (Figure 3) illustrates how the source language text and MT component may be replaced by a natural language generation (NLG) system, given a rich enough semantic representation. Other models are certainly possible, depending upon how the various processing tasks are subdivided.

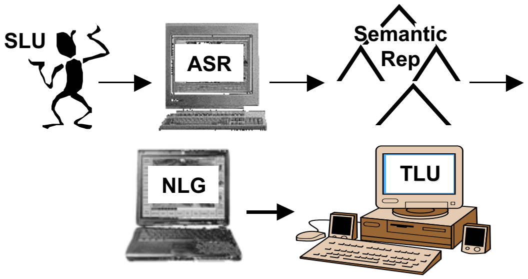

Figure 3. Model 3.

Regardless of how many levels and components there are in a given implementation, different metrics could be applied around any input-output pair of interest to help drive quality improvements. In Model 2 above, for example, we could have a metric around each processing module; that is, one metric for the mapping SLU to SLT, another metric for SLT to TLT, and a third from TLT to TLU. Each metric would be used to study the effectiveness of the system module of interest.

However, since the only guaranteed input-output pair regardless of the particular combination of technologies used would be SLU to TLU, and since all systems can be abstracted into Model 1 above, let us focus on an abstract metric which we will call the utterance-to-utterance (U2U) Metric. What do we require of our abstract U2U metric?

2 Metric Characteristics

I will not take a position on a particular metric, since metrics may vary depending upon the purpose of the people using them. However, I will take a position on various general characteristics of a metric for production use.

These arguments are based on my experience as principal author of the J2450 translation quality metric that has been adopted by the Society of Automotive Engineers as a *recommended practice* for evaluation of service information translations. [SAE]

In production environments—as opposed to system development—translation metrics are typically applied to random samples of source and target language translations. Metrics based on static reference translations for the automatic evaluation of system quality during system development [Doddington] are thus not in the domain of this discussion.

Evaluation is usually performed by a qualified translator with domain knowledge, who is generally employed by a translation agency, though client companies sometimes perform their own internal evaluations.

2.1 Primary Categories

First, a U2U metric for production use should consist of approximately seven categories of errors, plus or minus two. Seven categories are enough to provide adequate linguistic coverage yet are few enough to be usable by people in the real world. This is consistent with most metrics used by translation agencies.

For example, one category may refer to the appropriate mapping of source language words to target language words, where *appropriate* is defined by the category description. Another category may refer to word order. Given that we are discussing speech-to-speech systems, a category may be reserved for the intelligibility of the target language audio. Again, the particular categories will be dictated by the purpose of the people employing the metric. System application developers will have different interests from researchers working on the base technologies, and clients will have different interests than suppliers of the technologies.

With respect to categories, I would suggest that any particular U2U metric not include a primary category called *mistranslation*. While many translators use this term, and many translation agencies employ their own proprietary metrics that have an error category by this name, it is rarely, if ever, defined with any precision. The word *mistranslation* itself tends to evoke strong emotions in the translation industry, and for that reason alone it is not a good term to use for what is hoped will be a relatively objective, scientifically motivated metric.

2.2 Numeric Scores

The next characteristic of a U2U metric should be that it produces numeric scores for the utterance evaluations. That is, each category should itself produce a numeric score, and each category can be weighted or not, according to the goals of the metric's users. The presence and categorization of the errors are generally matters of human judgement, but once the error is recognized and categorized, a numeric result has numerous advantages. Given appropriate sample sizes it allows reasonable comparisons across different translation systems. It also allows easy use of statistical control charts for quality control [Godden 1996].

Employing a numeric score as the basis for a quality evaluation does not ipso facto allow the identification of a translation as good, bad or indifferent. Two different people may use the same metric, producing the same evaluation scores and yet define the notion of acceptable quality entirely differently. One person may define acceptable quality as a normalized quality score of .80 or higher, while another person may define acceptable quality only with a score of .90. The threshold of acceptance is independent of the metric, and is a business, not a technical decision.

2.3 Major and Minor Subcategories

Another necessary characteristic for a production metric encompasses the notion of a major vs. a minor error. An example will clarify both the concept and its utility. Suppose that a source language utterance contained the phrase *a door*, but that the S2S system translated it as *a window*. This is a lexical translation error that is major. An example of a lexical error that is minor would be a target language utterance of *an door*. It is ill-advised to penalize both errors with the same numeric score, which would happen if a 'wrong word' category always resulted in a single numeric value. Adding the major vs. minor distinction with different numeric scores allows the evaluator to penalize the first error more than the second.

Thus we have now constrained our metrics to include approximately seven primary error categories and two secondary categories (major versus minor) for a total of roughly fourteen distinct classifications of any given error. When an error in a translation is detected, the evaluator therefore has two assignments to make, the primary category and one of its two secondary categories. These primary and secondary category assignments are not always clear. Since translation quality judgements are generally human judgements, there will be evaluation variance across evaluators.

Is an incorrect gender on an article an example of an incorrect term or a syntax error? If the two categories have different penalty scores, then the category assignment can be a significant source of variance. Is the translation of a definite article as an indefinite article a minor error or a major one? That will depend upon the context of course, but it may also depend upon the person performing the evaluation.

2.4 Reducing Variance

To the extent that this human variance can be reduced, then the metric used by that evaluator will become more valuable.

The most effective way in which variance can be reduced is to give as precise a definition as possible of each error category, both primary and secondary. If a category of *wrong term* is to be used, then the notions of both *wrong* and *term* need to be defined precisely. Is "gas pedal" one term or two? Are function words regarded as terms? If the source language term is ambiguous, then what constitutes a wrong term in the target language? Definitions of error categories should be amply illustrated with examples.

The second most effective way to reduce variance is to provide training for evaluators. Sample utterances and translations with deliberate errors should be prepared in advance, offering several examples of each error category. Ideally, an entire training course would be designed around these examples and no person would perform working evaluations without taking the course. Also, evaluators should only be drawn from the ranks of qualified translators.

2.5 Ordering and Meta-Rules

But there is an additional way to reduce the variance in category assignment that can be incorporated into the metric itself. This can be done by employing rule ordering coupled with two meta-rules. The seven (plus or minus two) primary categories should be totally ordered by the numeric demerit penalty values referenced in the two subcategories.

For example, if primary category X has a major penalty demerit of five and a minor demerit of two, then it should be ordered before another primary category Y with major and minor demerits of four and three. If primary categories X and Y have major/minor demerits of three/four and three/five, respectively, then Y should be ordered before X.

Any potential ambiguities may be resolved by arbitrary sort order rules. The important concept is that each primary category be ordered with respect to every other primary category. Within a category, the major subtype is always ordered before the minor subtype.

Once the ordering is determined, then two meta-rules may be used to reduce error category assignment variance. The first meta-rule states that if the evaluator is unsure which primary category to assign to an error, then he or she should automatically assign it to that primary category highest in the sort order. Thus, if two evaluators are both unsure about which of two primary categories X or Y to assign to a given error, this first meta-rule forces them both to make the same decision. They will both select X, if X precedes Y in the sort order.

Similarly, the second meta-rule states that once the primary error category is assigned, if an evaluator is unsure whether the error constitutes a major or a minor instance of that error, then the evaluator should automatically regard it as a major error.

In this way, the metric itself—which now contains the two meta-rules—is removing some of the decision-making authority from the human evaluators, with the effect of reducing the variance in quality score demerit assignments. We must assume, of course, the honest and unbiased application of the metric by the evaluator. We also assume that both the metric definition as well as the training course and materials clearly emphasize the application and importance of the meta-rules.

The meta-rules impose a bias toward higher demerits, which is somewhat arbitrary. We could as easily have made the bias favor lower demerits. Any definition of acceptable quality, i.e., an acceptance threshold, based on the numeric scores can be adjusted up or down, according to the needs of the organization employing the metric. As previously stated, such acceptance criteria are business decisions, not technical ones. The important effect of the meta-rules is to reduce the variance in assignment of errors to categories.

3 Summary and Conclusions

The recommended U2U metric characteristics are summarized in Table 1.

Table 1. Characteristics of Metric.

Seven (+/- two) primary categories
Two secondary categories (major/minor) for each primary category.
Numeric demerits for each major and minor category.
Primary categories sorted by demerits.
Meta-Rule 1 removes ambiguity of primary category assignment.
Meta-Rule 2 removes ambiguity of secondary category assignment.

While I have discussed several characteristics of what I consider to be required elements of any adequate U2U translation quality metric for production use, these constraints still permit infinite variation in the definition of any particular metric. Primary categories may be drawn from any number of classifications that divide the error space, e.g. lexical, syntactic, semantic, phonetic, etc. Numeric demerits may be taken from any desired range, be it 0-1 or 1-1000.

Finally, let me say that just because a U2U metric conforms to the characteristics discussed in this paper, that metric does not automatically become a good metric. As previously discussed, the category definitions are of extreme importance, as are the examples used to illustrate the definitions and the training materials created for evaluators. Without clear, unambiguous and precise error definitions no metric will be of any practical value.

References

Doddington, G. 2002. “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Ocurrence Statistics.” *Notebook Proceedings.* Human Language Technology. San Diego, CA. pp. 128-132.

Godden, K. 1996. “Statistical Control Charts in Natural Language Processing.” *Proceedings.* Natural Language Processing and Industrial Applications. Moncton, NB, Canada. pp. 111-117.

SAE. 2001. “Translation Quality Metric.” Document Number J2450. Available through www.sae.org/technicalcommittees/j2450p1.htm