IMPROVED METRICS FOR MACHINE TRANSLATION EVALUATION

Viren Jain
Department of Computer & Information Science
University of Pennsylvania
viren@seas.upenn.edu

Supervision: **Aravind Joshi**
Department of Computer & Information Science
Institute for Research in Cognitive Science
University of Pennsylvania
joshi@linc.cis.upenn.edu

Abstract. We will investigate novel metrics for automated evaluation of machine translation quality. Current methods have proven useful over the past few years, but may no longer be sufficient to discern state of the art MT quality from superior human translations. We will consider the use of consensus and syntax based metrics that correlate well with human judgements. Furthermore, we wish to combine diverse metrics through a statistical framework trained from past human evaluations.

1. Introduction and Motivation

Automated machine translation (MT) evaluation has made it possible to measure the overall progress of the MT community as well as reliably compare the success of varying translation systems without relying on expensive and slow human evaluations. Yet despite the success of metrics such as the IBM BLEU or NIST score, state of the art performance in machine translation may be beginning to expose the weakness in such measures. Developers of the BLEU metric noted the existence of subtleties in translation output that their metric would fail to capture, but suggested that these subtleties would lead to relatively small effects as compared with other MT phenomena [Papineni 2000]. While this may have been true in the past, recent investigation and results suggests that the quality of state the art MT output exceeds the ability of current automated methods to capture discrepancies between machine output and qualitatively superior human translations.

An experiment performed during the 2003 Johns Hopkins CLSP Summer Workshop showed that sentences with the highest BLEU score from a list of hypothesis translations (ie, oracle-best sentence) had an average 105% relative performance as compared to humans. A quick survey of these sentences indicates that these sentences are much less fluent and comprehensible than even the worst reference translation. Additionally, the most recent TIDES evaluation showed that the best Arabic to English MT system had a relative 89% score to humans for translations whose fluency would suggest a greater than 10% relative discrepancy.

Problems with automated evaluation metrics have become even more important due to recent success in directly optimizing statistical MT systems to specific evaluation metrics (BLEU, word error rate, etc). Failure to provide an appropriate metric then directly limits the quality of translation and the ability to successfully incorporate improvements in these systems that deal with more subtle fluency issues such as syntactic well-formedness.

2. Proposed Work

As an extension of work done during the Johns Hopkins 2003 Summer Workshop on language engineering, we plan to systematically examine weaknesses in the current automated metrics and then investigate alternative and additional criteria for evaluation. Both of these goals would be pursued via highly empirical means that seek to establish relationships between automated output and human judgement of relative translation quality.

Two main considerations guide this proposal. Firstly, we seek robust **sentence-level** evaluation. Prior metrics such as the BLEU or NIST score are *corpus* level scores that are well-formed and reliable only in the case where one is comparing translations of an entire document or set of documents. This poses many problems, particularly when one seeks to optimize a translation system on a per-sentence basis using such metrics.

This suggests our second primary consideration, that of evaluating our metric on an **end-to-end** basis; we plan to optimize the discriminative error training on the translation system developed by Franz Och/ISI directly to the new metrics. This will provide us with an alternative set of translations, based simply on changing the evaluation measure. These translations can then be examined, on an initially anecdotal basis, to see whether translation quality has improved [Och 2003]. A particularly interesting experiment will be to see whether reranking features developed

during the 2003 Workshop have higher discriminability within this new evaluation framework. The negligible impact of many of the features developed during that time was, in part, believed to be due to the insensitivity of BLEU scores to syntactic improvement.

3. Methods and Procedure

We propose to investigate two primary types of evaluation metrics: *consensus* based lexical evaluation, and *syntactically* based fluency judgement.

Cross-lingual corpuses specifically made for machine translation research most often include multiple reference translations (i.e., each source sentence has multiple, human generated target sentences). Metrics such as BLEU deal with multiple reference translations most often by simply checking n-grams against all translations, or simply using the best score of all the reference translations. We believe there may be more information in this set of reference translations than is currently exploited.

By determining the common lexical items across all reference translations, we should be able to determine those items which represent important content words and proper nouns that any good translation should contain. Conversely, we can then also identify rare lexical items within the reference translations, denoting words that may be esoteric in use but not important to the content of the sentence. In this way, we hope to develop a robust measure that indicates whether a given translation "gets the important things right." The primary goal and challenge in this area of research will be to identify experimentally sound weighting parameters and formulas.

A more sophisticated line of work involves integrating the use of syntactic tools and measures into an evaluation framework. We propose to make use of shallow syntactic information given by part of speech taggers and chunkers. Part of speech taggers assign a surface level categorization to words in a sentence, denoting the basic linguistic role of a word in a given sentence. Modern taggers can achieve accuracy in excess of 96% (Ratnaparkhi 1996). Simple counts of tags or even full-blown n-gram analysis of tag sequences can be computed to exploit this information. Chunkers provide a flat representation of the syntax of a sentence that would normally be given by roughly the first level of a parse tree above part of speech tags (e.g., verb phrase, noun phrase, etc). This information can be similarly exploited. We plan to use easily and freely available taggers; if our metric proves useful we would like to make a straightforward software release of the system.

In order to combine these two types of metrics, we will introduce a machine learning system that treats each of these metrics (and possibly others) as features within a statistical framework. The system would be trained on recent TIDES MT evaluation results, older ARPA evaluations, and other resources such as data from the VerbMobil project at University of Aachen; these resources provide a collection of evaluated translations that will serve as a basis for comparison to human judgements and explicit training for the machine learning system. For the machine learning layer, we plan to use "off the shelf" software that implement techniques such as multi-layer perceptrons or log-linear models.

As suggested before, the new measure will be tested "end to end" by Franz Och's discriminative training. This is a highly effective decoding technique in which instead of optimizing a maximum likelihood citerion, an arbitrary error measure is

directly optimized. This has been proven effective in speech recognition, among other tasks (such as MT itself). Several interesting results are expected to emerge from this: the discrepancy between the best scoring machine translation and the score of any human translation, the "human-judged" quality of the best scoring translation under the new automated measure, and the effect of previously ineffectual reranking features on this new score. This will provide an interesting disambiguation of results and issues within the 2003 Summer Workshop on Syntax for Machine Translation.

4. Relation to Prior Work

Two major automated metrics have dominated most machine translation work: the BLEU and NIST measure. While there are a number of differences between the two, both are essentially n-gram corpus-level measures. BLEU heavily rewards large n-gram matches between the source and target; while a useful characteristic, this can often unnecessarily penalize syntactically valid but slightly altered translations with low n-gram matches. Furthermore, both of these measures are ill-formed at the sentence-level.

Both BLEU and NIST have in the past proven fairly effective in measuring overall translation quality. However, the limits of these measures have recently become clear. Our work distinguishes itself from these metrics in several ways: emphasis on *sentence-level* evaluation, full exploitation of *multiple* reference translations, and direct integration of *syntactic* information.

5. References

Och, Franz Josef. "Minimum Error Rate Training for Statistical Machine Translation". In "ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics", Japan, Sapporo, July 2003.

Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J. "Bleu: a method for automatic evaluation of machine translation". Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001.

Ratnaparkhi, Adwait. "A Maximum Entropy Part-Of-Speech Tagger". In "Proceedings of the Empirical Methods in Natural Language Processing Conference", University of Pennsylvania, May 1996.