How Does Automatic Machine Translation Evaluation Correlate With Human Scoring as the Number of Reference Translations Increases?

Andrew FINCH
ATR Spoken Language Translation Reseach Laboratories
2-2-2 Hikaridai "Keihanna Science City"
Kyoto, 619-0288, Japan
Andrew.Finch@atr.jp

Yasuhiro AKIBA
ATR Spoken Language Translation Reseach Laboratories
2-2-2 Hikaridai "Keihanna Science City"
Kyoto, 619-0288, Japan
Yasuhiro.Akiba@atr.jp

Eiichiro SUMITA
ATR Spoken Language Translation Reseach Laboratories
2-2-2 Hikaridai "Keihanna Science City"
Kyoto, 619-0288, Japan
Eiichiro.Sumita@atr.jp

Abstract

Automatic machine translation evaluation is a very difficult task due to the wide diversity of valid output translations that may result from translating a single source sentence or textual segment. Recently a number of competing methods of automatic machine translation evaluation have been adopted by the research community, of these the some of the most utilized are BLEU, NIST, mWER and the F-measure. This work extends the work of others in the field looking at how closely these evaluation techniques match human performance at ranking the translation output. However, we focus on investigating how these systems scale up with increasing numbers of human-produced references. We measure the correlation of the automatic ranking of the output from nine different machine translation systems, with the ranking derived from the score assigned by nine human evaluators using up to sixteen references per sentence. Our results show that evaluation performance improves with increasing numbers of references for all of the scoring methods except NIST which only shows improvements with small numbers of references.

1. Introduction

Evaluation by human judges is perhaps the most trusted means of evaluating machine translation (MT) output in spite of the fact that the analysis is subjective, and scoring is often inconsistent between different annotators. The main drawback of this approach however, is that it is very expensive in terms of the time and cost involved in annotating the machine translation output. Moreover, during the development process of machine translation systems, many evaluations need to be performed making evaluation by human annotators prohibitively expensive. Therefore researchers have sought ways of evaluating their systems automatically during development.

All of the automatic machine translation evaluation systems commonly used at present work on the principle of measuring the closeness, in some sense, of the target sentence produced by the machine translation system being evaluated to a reference translation, or a set of reference translations. One of the main problems that needs to be addressed in this approach is caused by the fact that the number of ways in which it is possible to express the same meaning in a language is very large. The challenge for an automatic machine translation evaluation method is to cover as many of these possible cases as possible. One way to tackle this problem, proposed by Thompson (1991) is to compare the output of the machine translation system to a set of references rather than just a single reference, thereby covering more of the possible correct translations. Clearly therefore, having the ability to scale up well with an increasing number of references is an important attribute for any machine translation evaluation scheme. This paper investigates how a number of popular scoring systems behave on the output from a number of different machine translation systems when the number of reference translations is varied.

The composition of this paper is as follows: the first section describes the automatic evaluation methods (BLEU, NIST, mWER and the F-measure) that will be examined in this paper; the next section outlines the nine machine translation systems that were used to produce the output for our evaluation; Section 4 sets out the experimental methodology, data and analysis techniques; Section 5 presents the experimental results, and the final section concludes and offers some directions for future research.

2. Evaluation Methods

In this paper, we investigate the behavior of four of the most popular MT evaluation methods. Two of which, BLEU and NIST, are based on *n*-gram precision, that is, the proportion of *n*-grams in the output translation that match *n*-grams in the reference set. The third, mWER is based on edit distance, and the last, the F-measure being a unigram-based technique that measures the commonality of contiguous sequences of words in terms of both precision and recall. These methods are outlined briefly below.

BLEU[1]

The BLEU scoring system (Papineni *et al.*, 2001) scores the translation output by measuring the precision of the component *n*-grams (in this case, 1, 2, 3 and 4-grams) of the segments, with respect to a set of reference translations. The idea being that a good translation will share more *n*-grams in common with the reference segments than a bad one. The score is the geometric mean of the *n*-gram precision, multiplied by a brevity penally (BP) that penalizes the translation only if it is shorter than the reference. The score for a single candidate translation is given by:

1 The software used to derive the BLEU and NIST scores in these experiments is version 09c of the NIST MT evaluation kit and is available from the URL: http://www.nist.gov/speech/tests/mt/mt2001/resource/

$$BLEU = BP \cdot \exp\left[\sum_{n=1}^{N} \frac{1}{N} \cdot \ln(p_n)\right]$$

Where:

$$p_n = \frac{\sum_{w_1 \ldots w_n \in C} \text{Count}_{\text{clip}}(w_1 \ldots w_n)}{\sum_{w_1 \ldots w_n \in C} \text{Count}(w_1 \ldots w_n)}$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(1 - \frac{r}{c}) & \text{if } c \le r \end{cases}$$

C	is the candidate translation.
Count(w_1...w_n)	is the number times the *n*-gram w_1...w_n occurs in the candidate translation.
$\text{Count}_{\text{clip}}$($w_1$...$w_n$)	is the number times w_1...w_n matches a reference *n*-gram, limited to the maximum number of times it has occurred in any of the references.
n	is the order of the *n*-gram.
N	is the maximum *n*-gram length.
c	is the length of C.
r	is the average reference length for this segment.

NIST[1]

The NIST scoring system (Doddington, 2002), a similar information theoretic approach, is again based on *n*-gram (in this case, 1, 2, 3, 4 and 5-grams) precision but it employs the arithmetic average of *n*-gram counts rather than a geometric average, and the *n*-grams in this case are weighted according to their information contribution, as opposed to just counting them as in BLEU. The score represents the average information per word, given by the *n*-grams in the translation that match an *n*-gram of a reference in the reference set. NIST's brevity penalty penalizes very short translations more heavily, and sentences close in length to the references less than the BLEU brevity penalty. The score for a single candidate translation is given by:

$$NIST = BP \cdot \sum_{n=1}^{N} \sum_{w_1 \ldots w_n \in C} \frac{\text{Info}(w_1 \ldots w_n)}{\text{Count}(w_1 \ldots w_n)}$$

Where:

$$\text{Info}(w_1 \ldots w_n) = \log_2\left[\frac{\text{number of occurences of } w_1 \ldots w_{n-1}}{\text{number of occurences of } w_1 \ldots w_n}\right]$$

The *n*-gram counts used to calculate these information weights are derived from the reference set.

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp(\beta \ln^2(\frac{c}{r})) & \text{if } c \le r \end{cases}$$

Here β is selected such that when $c = \frac{2r}{3}$ BP = 0.5

C, *n*, *N*, *c*, *r* and Count(w_1...w_n) are the same as for BLEU.

mWER

The multi-reference word error rate (mWER) (Nießen *et al.*, 2000) is based on edit distance: the minimum number of word insertion, deletion and substitution operations required to transform a translation into a reference sentence. When using multiple references the score generalizes to the minimum edit distance between the machine translation output segment and any of the segments in the reference set. The brevity penalty in this case being implicit in the method, penalizing sentences for being too long as well as too short.

F-measure[2]

The F-measure, proposed by Melamed *et al.*, (2003) is a unigram-based technique, extended to include longer contiguous word-sequence matches that is derived from the precision and recall scores often used in NLP evaluation. The score rewards longer matches that are contiguous and have the same word order as the reference. It also rewards longer runs of matching words in a way which is more than just linear in their run length. To do this, the measure incorporates a exponent parameter *e* that controls the weighting given to longer word sequence matches. For the purposes of these experiments, we used a value of *e*=1.

3. The Machine Translations Systems

The output from nine Japanese to English machine translation systems was used for this study, these consisted of three different releases spaced at 6-month intervals, of three different types of MT system. The training data for the first 2 releases remained constant, the systems differing only in their algorithms, or in the case of TDMT, the number of rules and translation dictionary content. For the final release, a training corpus approximately 3 times the size was used. The systems were:

- SMT (Statistical Machine Translation). Using the publicly domain GIZA++ software (Och and Ney, 2000) together with an in-house developed multi-stack decoder.
- TDMT (Transfer Driven Machine Translation). A pattern-based MT system using hand-coded syntactic transfer rules.
- D^3 (DP-match Driven Transducer). An example-based MT system using online-generated translation patterns.

4. Experimental Design

Overview

The output from nine different MT systems was first scored by human judges. We used this human scoring as the benchmark by which to judge the automatic evaluations. The same MT output was then evaluated using each of the four automatic scoring systems. The automatically scored segments were analyzed for Spearman Rank Correlation with the ranking defined by the categorical scores assigned by the human judges. An increase in correlation indicating

2 The software used to derive the F-measure scores in these experiments is version 1.1 of the General Text Matcher (GTM) software and is available from the URL: http://nlp.cs.nyu.edu/GTM/

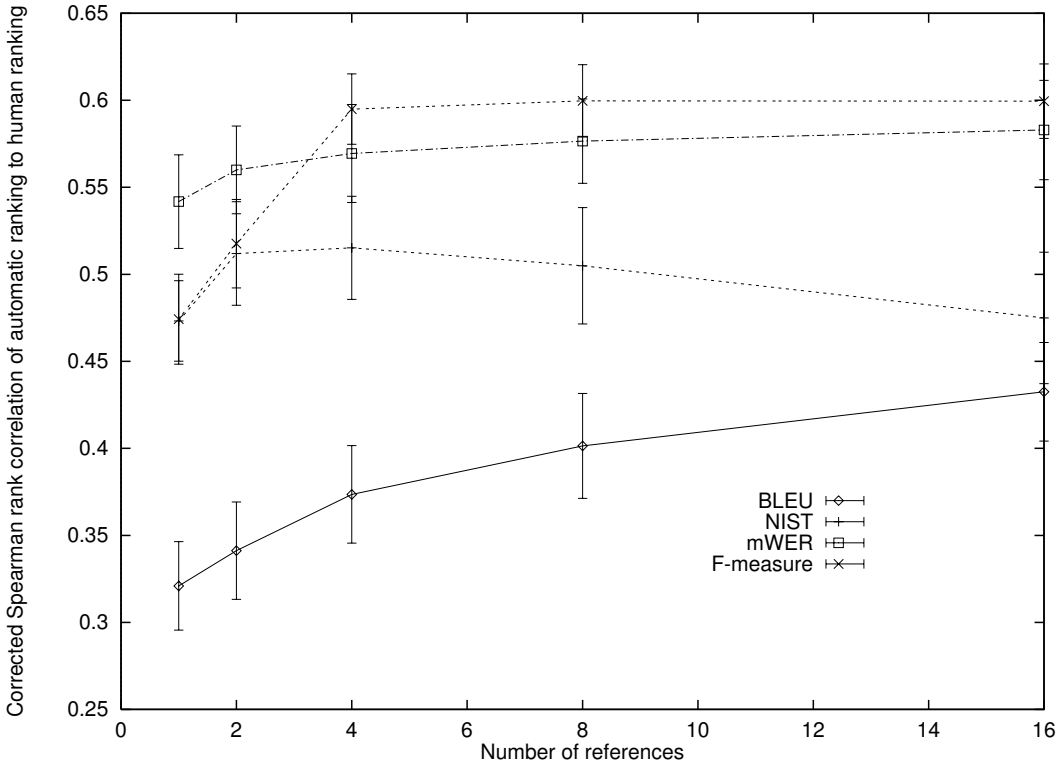

Figure 1 : The effect of adding more references on automatic MT scoring

that the automatic system is more similar to a human in ranking the MT output.

Experimental Data

Our test data consisted of a set of English sentences that have been translated from 345 Japanese by nine different machine translation systems (that is, 3105 machine translation output sentences in total). The Japanese source sentences were randomly selected from the Basic Travel Expression Corpus (BTEC) (Takezawa *et al.*, 2002). Each output sentence was scored by nine independent native English-speaking human evaluators who were also familiar with the source language. Every sentence was assigned a grade in accordance with the following five-point scale for both adequacy and grammaticality:

(S) Native English;

(A) Perfect: no problems in either information or grammar;

(B) Fair: easy-to-understand, with either some unimportant information missing or flawed grammar;

(C) Acceptable: broken, but understandable with effort;

(D) Nonsense: important information has been translated incorrectly.

A single grade was derived for each sentence by selecting the median grade from the nine grades assigned by the human judges.

Test sets of 1000 pseudo-documents were constructed by taking random samples of 30 sentences from the 345 test sentences in the same manner as Turian *et al.* (2003). This is because statistics based on short translations of a single sentence proved to be unreliable (see Turian *et al.*, (2003) for a more detailed exposition of this effect).

Correlation Analysis

Following Melamed *et al.* (2003) and Turian *et al.* (2003), we chose to use Spearman Rank Correlation to evaluate how similar our automatic MT evaluations were to human evaluation. Instead of correlating the absolute values of the scores themselves, the scored test data is ordered by score, and assigned a rank indicting its position in this ordering. The ranks themselves are then analyzed directly for correlation. By using this scheme we are placing importance on ensuring our automatic scoring system ranks translations in the same way that a human judge would rank them. We chose to use a variant called Corrected Spearman Rank Correlation which corrects for cases where tied ranks occur. Tied ranks can occur in the human grading since there are only 5 categories.

5. Results

Our experiments support the findings of others (Turian *et al.*, 2003; Doddington, 2002), showing that adding more references to the reference set improves the MT evaluation performance (Figure 1), except in the case of NIST where the performance improved gradually from with 1 to 4 references, with the addition of more then 4 references evaluation performance degraded. For NIST, 16 references offered a comparable level of performance to just a single reference. A similar effect was reported in Doddington (2002), where using up to 8 references offered little or no improvement over using a single reference. Our results also corroborate the overall findings of Turian *et al.* (2003) that

using Spearman Rank Correlation, the BLEU score correlates the lowest with human ranking, then the NIST score, with the F-measure having the strongest correlation. The F-measure with a single reference gave almost the same results as the NIST score. The F-measure however, benefits greatly from the addition of references, rapidly becoming the best of the scoring systems tested here when using four references or more. But the law of diminishing returns applied when increasing number of references above 4 references. We found that BLEU scoring improved well with increasing numbers of references, however, the overall correlation with human performance was lower than for the other systems, even with 16 additional references.

It is difficult to explain why increasing the number of references to the NIST score does not result in better correlation with the human ranking. One would expect that since the counts used in the estimates of information contribution in the score were based on more data, the estimate would become more accurate, thereby improving the score. One possible contributing factor might be the composition of the *n*-gram's influence. We observed that as the number of references increased, so did the proportion of the score that was contributed by higher-order *n*-grams. Furthermore, we also observed that decreasing the maximum *n*-gram size used by the scoring improved the correlation with human ranking on this data.

6. Conclusion

The experiments reported here compare the performance of several automatic machine translation evaluation systems with varying sizes of reference sets. Improvements in evaluation quality with increasing number of references is a natural and desirable characteristic of these systems. The production of references by human authors can be an expensive process however, and it is tempting to conclude that the focus should be on systems that perform well with small numbers of references. On the other hand, it is also possible to add much larger numbers of synthetic references to the reference set. Finch *et al.*, (2004), used a paraphraser to produce up to 100 paraphrases of the reference set, and then added these paraphrases to the reference set. Using this technique it proved possible to obtain a very large reference set very cheaply, although such a reference set contains erroneous sentences. Their results show that even with the noise from incorrect machine-generated references, it is possible to improve the evaluation performance over and above that obtained by using only human-produced references for BLEU and mWER. Adding paraphrases to the references when using NIST scoring however only made the evaluation performance worse. This is not surprising given that additional human-produced references made little improvement in evaluation performance in the experiments reported here. For applications such as this, the ability of the evaluation software to scale up well with the addition of large numbers of references is pivotal.

Acknowledgments

This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurance Statistics, *Human Language Technology: Notebook Proceedings* (pp. 128-132), San Diego, USA.

Finch, A., Watanabe, T., and Sumita, E. (2002). Paraphrasing by Statistical Machine Translation, *Proceedings of FIT-2002*, Tokyo, Japan.

Finch, A., Watanabe, T., and Sumita, E. (2003). Data-Oriented Paraphrasing, *Proceedings of RANLP-2003*, Borovets, Bulgaria.

Finch, A., Akiba, Y., and Sumita, E. (2004). Using a Paraphraser to Improve Machine Translation Evaluation, *Proceedings of IJCNLP-2004*, Hainan Island, China.

Melamed, I., Green R., and Turian, J. (2003). Precision and Recall of Machine Translation. *In Proceedings of the HLTNAACL 2003: Short Papers (pp. 61-63).* Edmonton, Canada.

Nießen, S., Och, F.J., Leusch, G., and Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research, *Proceedings of the LREC Conference*, Athens, Greece.

Och F.J. and Ney, H. (2000). Improved Statistical Alignment Models, *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (pp. 440-447)*, Hong Kong, China.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation, IBM Research Report rc22176 (w0109022), IBM Research Division, Thomas J. Watson Research Center.

Sugaya F., Takezawa, T. and Kikui, G. (2002). Proposal of a Very-large-corpus Acquisition method by Cellformed Registration, *Proceedings of the LREC Conference*, Las Palmas, Gran Canaria.

Takezawa, T., Sumita, E., Sugaya, F. and Yamamoto, H. (2002). Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World, *Proceedings of the LREC Conference*, Las Palmas, Gran Canaria.

Thompson, H. (1991). Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment . *In (ISSCO) Proceedings of the Evaluators Forum (pp. 215-223)*, Geneva, Switzerland.

Turian, J.P., Shen, L. and Dan Melamed, I. (2003). Evaluation of Machine Translation and its Evaluation, *Proceedings of MT Summit IX*, New Orleans, LA.

White, J., O Connell T., and Carlson L. (1993). Evaluation of machine translation. *In Human Language Technology: Proceedings of the Workshop (ARPA) (pp. 206 210).*