Evaluation of Machine Translation and its Evaluation

Joseph P. Turian, Luke Shen, and I. Dan Melamed

New York University
New York, New York
{turian,ls750,melamed}@cs.nyu.edu

Abstract

Evaluation of MT evaluation measures is limited by inconsistent human judgment data. Nonetheless, machine translation can be evaluated using the well-known measures precision, recall, and their average, the F-measure. The unigram-based F-measure has significantly higher correlation with human judgments than recently proposed alternatives. More importantly, this standard measure has an intuitive graphical interpretation, which can facilitate insight into how MT systems might be improved. The relevant software is publicly available from http://nlp.cs.nyu.edu/GTM/.

1 Introduction

In the early 1990s, the U.S. government sponsored a competition among machine translation (MT) systems. One of the valuable outcomes of that enterprise was a corpus of manually produced numerical judgments of MT quality, with respect to a set of reference translations (White *et al.*, 1993). The relatively high cost of producing such judgments and the benefits of objective evaluation have encouraged many researchers to seek reliable methods for estimating such measures automatically.

Most efforts have focused on strategies for computing some kind of similarity score between the output of an MT system and one or more reference translations. Early approaches to scoring a "candidate" text with respect to a reference text were based on the idea that the similarity score should be proportional to the number of matching words (e.g. Melamed, 1995). Another idea is that matching words in the right order should result in higher scores than matching words out of order (e.g. Brew & Thompson, 1994; Rajman & Hartley, 2001).

Perhaps the simplest version of the same idea is that a candidate text should be rewarded for containing longer contiguous subsequences of matching words. Papineni *et al.* (2002) recently reported that a particular version of this idea, which they call "BLEU," correlates very highly with human judgments. Doddington (2002) proposed another version of this idea, now commonly known as the "NIST" score. Although the BLEU and NIST measures might be useful for comparing the relative quality of different MT outputs, it is difficult to gain insight from such measures. What does a BLEU score of 0.016 mean?

In this paper, we show how MT can be evaluated in terms of the standard measures of precision and recall, as well as their composite F-measure. These measures have an intuitive graphical interpretation, which can facilitate insights into how MT systems might be improved. We present experiments showing that:

- The correlation between human judgments of MT quality is surprisingly low.
- Therefore, not surprisingly, the correlation between human judges and all automatic measures of MT quality is also quite low, contrary to Papineni *et al.* and Doddington.
- For the MT systems evaluated in the 2002 DARPA MTEval exercises, the unigram-based F-measure that follows from Melamed (1995) is more reliable than the more recently proposed BLEU and NIST measures.

2 Precision and Recall of MT

Precision and recall are widely used to evaluate NLP systems. When comparing a set of candidate items Y to a set of reference items X:

$$\text{precision}(Y|X) = \frac{|X \cap Y|}{|Y|}; \ \text{recall}(Y|X) = \frac{|X \cap Y|}{|X|} \quad (1)$$

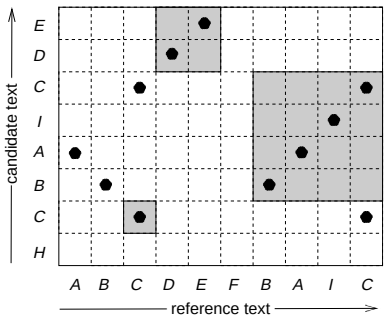


Figure 1: Computation of the maximum match size, using either unigrams or aligned blocks.

Both functions are proportional to $|X \cap Y|$, the size of the set intersection in their numerator. The main challenge in adopting these well-known measures for evaluation of MT systems is finding an appropriate definition for the intersection of a pair of texts.

2.1 Unigram-Based Measures

The intersection of two items is what they have in common. A bitext grid can show what two texts have in common. Figure 1 shows an hypothetical reference text on the X axis and an hypothetical candidate text on the Y axis. Whenever a cell in the grid co-ordinates two words that are identical, we place a bullet in it, and call it a **hit**.

As a first approximation, suppose we were not interested in giving more credit for correct word order. A naïve approach to computing $|X \cap Y|$ would be to count the number of hits in the grid. However, this algorithm runs the risk of double-counting, for example by awarding two hits for B in the reference in Figure 1.

To avoid double-counting, we borrow the concept of "maximum matching" from graph theory (Cormen *et al.*, 2001, pg. 1051). A **matching** is a subset of the hits in the grid, such that no two hits are in the same row or column. The **match size** of a matching is the number of hits in the subset. A **maximum matching** is a matching of maximum possible size for a particular bitext.[1] The **maximum match size** (MMS) is the size of any maximum matching. For example, the hits that are in the shaded region of Figure 1 are a maximum matching, so the MMS is 7.

The MMS ranges from zero to the length of the shorter bitext axis. We can divide the MMS by the length of the candidate text (C) or the length of the reference text (R) to obtain the precision or the recall, respectively:

$$\text{precision}(C|R) = \frac{\text{MMS}(C,R)}{|C|} \quad (2)$$

$$\text{recall}(C|R) = \frac{\text{MMS}(C,R)}{|R|} \quad (3)$$

2.2 Rewards for Longer Matches

The unigram-based measures above can be extended to reward a candidate text for contiguous hits in the right order. Contiguous sequences of matching words appear in a bitext grid as diagonally adjacent hits, running parallel to the main diagonal. We shall refer to such sequences as **runs**. The unigram-based method for computing the MMS already rewards a candidate text proportionally to run length, but it produces the same MMS if the hits are not contiguous or are in the wrong order. To reward correct word order, it is necessary to reward runs *more* than linearly in their length. BLEU and NIST do so by double-counting all sub-runs. We propose to do so by generalizing the definition of match size.

We treat runs as atomic units. Each run's minimum enclosing square is one **aligned block**. A candidate text is rewarded in proportion to the *area* of non-conflicting aligned blocks, as illustrated by the shaded squares in Figure 1. Specifically, we define the **weight** of a run to be the square of the run length. We then generalize the definition of match size as follows:

$$\text{size}(\text{M}) = \sqrt{\sum_{r \in \text{M}} \text{length}(r)^2} \quad (4)$$

where each r is a run in the matching M. A maximum matching and its size are determined as before. For example, the size of the maximum matching in Figure 1 is $\sqrt{4^2 + 2^2 + 1^2} = \sqrt{21} \approx 4.6$.

When some run r_1 partially conflicts with a longer run r_2, the non-conflicting remainder of r_1 (which is itself a run) can still participate in the maximum matching. In particular, if individual hits are part of the maximum matching, they contribute a weight of $1^2 = 1$ to the MMS.

The purpose of the square root in Equation 4 is to normalize the MMS with respect to the lengths of the inputs. In the limiting case that a candidate text is

[1] There may be more than one maximum matching for a given bitext.

identical to the reference text, the entire bitext grid is covered by one aligned block, and precision = recall = 1.

Since precision and recall scores in isolation are "gameable",[2] they are typically combined into various other common measures. Their harmonic mean, the so-called "F-measure," (van Rijsbergen, 1979) has a particularly intuitive interpretation in the context of a bitext grid: It represents the (root of the) fraction of the grid covered by aligned blocks.

Measures based on Equation 4 heavily weight matching longer runs. We can adjust this weight by generalizing Equation 4 to arbitrary exponents:

$$\text{size}(\text{M}) = \sqrt[e]{\sum_{r \in \text{M}} \text{length}(r)^e} \qquad (5)$$

The special case where $e = 1$ follows from Melamed (1995).

We conjecture that when $e > 1$, computing the MMS is NP-hard. In practice, we use a greedy approximation that builds a matching by iteratively adding the largest non-conflicting aligned blocks. Simulations on the data described in Section 3.1 have shown that this approximation finds a true maximum matching 99% of the time. In the rare remaining cases, the size of the output matching is at least 80% of the maximum.

So far, we have described how to measure the similarity between two sentences.[3] We now extend our measures to score documents. For a candidate document C and a reference document R, each of which contain n sentences:
$|C| = \Sigma_{i=1}^{n}|C_i|$; $|R| = \Sigma_{i=1}^{n}|R_i|$; $|C \cap R| = \Sigma_{i=1}^{n}|C_i \cap R_i|$
The precision, recall, and F-measure are calculated as before, using these aggregate values.

2.3 Multiple References

One of the main sources of variance in MT evaluation measures is the multitude of ways to express any given concept in natural language. A candidate translation can be perfectly correct but still very different from an equally correct reference translation. One approach to reducing this source of variance, thereby improving the reliability of MT evaluation, is to use multiple references (Thompson, 1991).

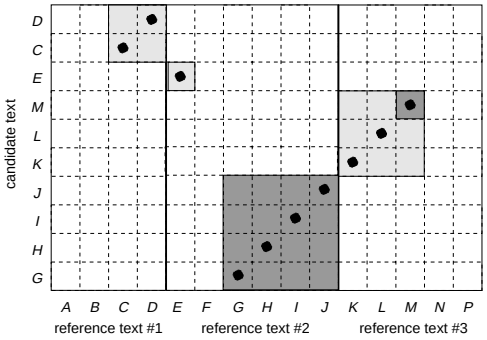


Figure 2: Using multiple references: The initial maximum matching (all shading) is capped by the mean reference length of 5, to arrive at the final matching (dark shading).

Figure 2 illustrates how to compute the MMS when multiple reference translations are available. Step 1 is to concatenate the relevant reference texts, in arbitrary order. Step 2 is to find a maximum matching in the resulting grid as before, except that a barrier between adjacent reference texts prevents runs from starting in one reference and finishing in another. Step 3 is to cap the MMS with respect to the lengths of the input texts.

Step 3 deserves more explanation. In the single-reference setting, the MMS is naturally limited by the candidate length and the reference length. By analogy, in the multiple-reference setting, we limit the MMS by the candidate length and the *mean* reference length. That is, we do not allow the number of hits in any matching to exceed the mean reference length. If there are excess hits in a maximum matching, we delete hits from the matching until the number of hits is equal to the mean reference length. Hits are deleted in the order that maximizes the size of the remaining matching, i.e. they are deleted from shorter runs first. Figure 2 illustrates hit deletion to cap the MMS. After the maximum matching has been pared in this manner, we normalize it as before.

3 Experimental Design

3.1 Data

We used two corpora, one comprising 10 English translations of 728 Arabic sentences and one comprising fourteen English translations of 878 Chinese sentences. Of the ten Arabic texts, six were ma-

[2] A system can inflate its precision and recall scores. Specifically:
precision = 1 if the candidate text contains only "the".
recall = 1 if the candidate text contains every word in the vocabulary.

[3] We use the term "sentence" loosely, to refer to any coherent segment of text.

chine ("candidate") translations and four were human ("reference") translations. The Arabic reference texts' sentences ranged in length from 1 to 95 words (mean 31.3, standard deviation 15.4). Of the fourteen Chinese texts, there were ten candidate translations and four reference translations. The Chinese reference texts' sentences ranged in length from 2 to 114 words (mean 30.8, standard deviation 16.8).

Human judges scored the candidate translations on Adequacy and Fluency, on a scale of 1-5.[4] Each judgment of each candidate sentence was made with respect to one particular reference translation. Although every candidate sentence received two or three scores from different judges, there were no sentences for which some judge evaluated every candidate translation. However, every sentence in a given document was evaluated by the same judge. As such, the human judges had access to information that automatic MT evaluation measures currently ignore.

3.2 Sampling the Corpora

Any MT evaluation measure is less reliable on shorter translations. But, reliability on shorter texts, as short as one sentence or even one phrase, is highly desirable because a reliable MT evaluation measure can greatly accelerate exploratory data analysis.

Consider how MT system developers would measure the effect of a system modification on a large development bitext. Typically, they would like to know not only whether the modification improved performance on some objective measure, but also why or why not. The fastest way to gain such insight is to compare the system's "before" and "after" output on some specific text sentences. The sentences that are most likely to highlight the qualitative effects of the modification to the MT system are those for which the objective evaluation measure changes the most. However, if the evaluation measure is not reliable, then the developer might need to examine many sentences before finding one that provides any intuition. Thus, unreliable measures can be a waste of time. A measure that is reliable only when averaged over a large corpus is not useful for exploratory data analysis.

[4] See `http://www.ldc.upenn.edu/TIDES/` for details about the corpora and the manual evaluation method.

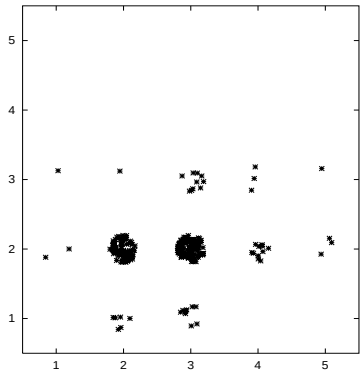


Figure 3: Paired Adequacy judgments for two of the judges over the 227 sentences that they both evaluated. For these pairs, the Spearman correlation coefficient is 0.019. A random skew of 0.20 was added to each point in this figure to show the density.

We measured the reliability of various MT evaluation measures on texts of different lengths. A pseudo-document of length *n* was created by concatenating *n* randomly chosen sentences. The human score for a candidate pseudo-document was computed by randomly selecting one of the human judgments for each sentence therein, and taking their mean. In order to ensure statistical significance, we created 1000 pseudo-docs of length *n* for each *n*.

3.3 Calculating Correlation Coefficients

The most important criterion for an automatic MT evaluation measure is that it rank MT systems the same way that a human judge would rank them. A measure that often misranks systems is less useful, even if it is otherwise good at predicting the absolute differences between systems' scores.[5] Therefore, we compared the automatic measures by how well their relative rankings of the candidates matched those of the human judges. Specifically, for a given pseudo-document containing *n* sentences, and a fixed set of *r* references, we computed the Spearman rank correlation between the human judgments and the automatic measure, for every machine translation of that pseudo-document.

As each candidate sentence was judged by several people, our sampling method enabled us to compute inter-judge correlations. Inter-judge correlation was poor. Figure 3 plots the paired Adequacy scores for

[5] In any case, the instructions for manual evaluation all but guaranteed that the judges' scores would not be on a linear scale, so linear regression is inappropriate for evaluating automatic measures.

the two judges with the lowest inter-judge correlation. To improve the rank correlation between human judgments, we performed a *z*-transform on each judge's scores, such that each judge's scores would have zero mean and unit variance.

4 Results

On advice from George Doddington (p.c.), we ran our first set of experiments on unstemmed text, with original case information retained. Figures 4(a) and 4(b) show the mean Spearman correlation with Adequacy and Fluency scores, respectively, on the Chinese corpus for several automatic MT evaluation measures—the F-measure with $e = 2$, the F-measure with $e = 1$, the BLEU score, and the unweighted NIST score—as well as the inter-judge correlation. These graphs reveal several interesting trends.

Our most important finding is that on shorter documents (where it counts the most), the mean inter-judge correlation is disappointingly low. This is partly attributable to the difficulty of comparing MT systems of similar quality, but partly to the design of the manual evaluation procedure. Melamed *et al.* (2003) did not encounter this problem because they dealt with fewer systems whose translation quality was much easier to distinguish. Low inter-judge correlation in the present experiment underscores how little the community understands about the MT evaluation problem. If the MT research community is serious about designing reliable automatic MT evaluation measures, then we must obtain human judgment data through more reliable means.

Automatic MT evaluation measures cannot be faulted for poor correlation with the human judges, as the judges do not correlate well with each other. Contrary to intuition, the automatic measures' correlations nonetheless surpass the inter-judge correlation in some instances. This happens because the human scores are rather inconsistent. So, there is more co-variance between human score pairs than between human scores and automatic scores.

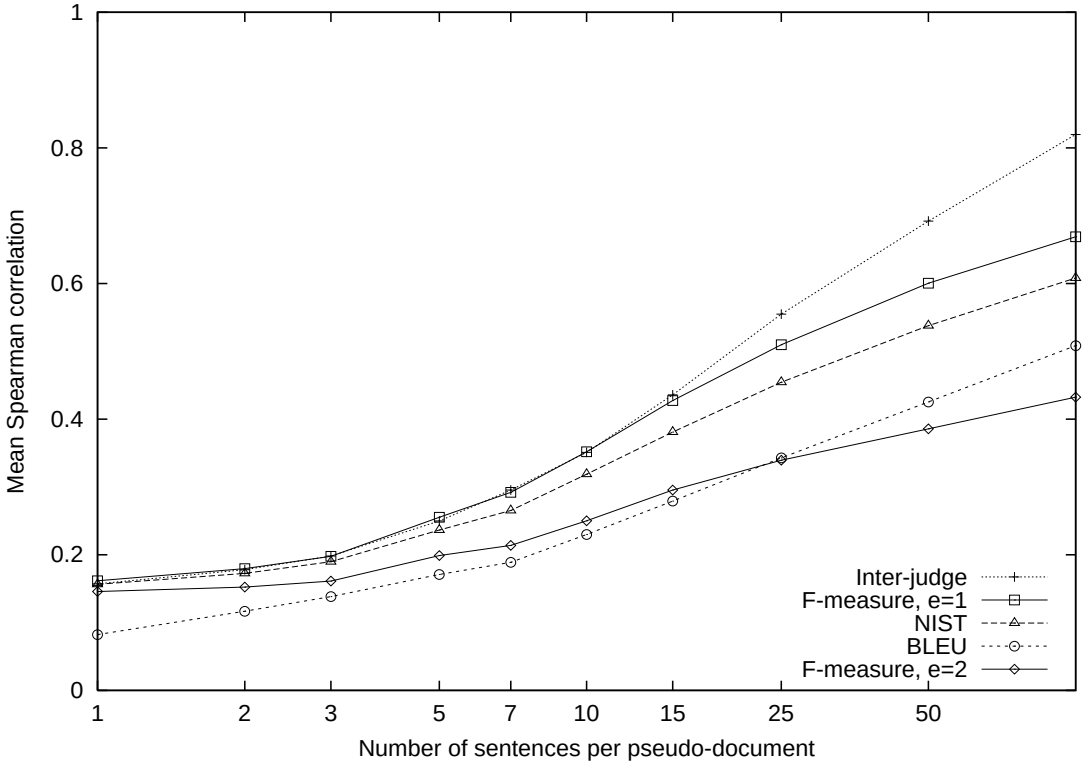
Our other main finding is that a simple unigram measure produces the most accurate rankings of MT systems on the Chinese corpus. A detailed analysis of the results revealed two complementary explanations. First, none of the MT systems involved in these experiments was very good at rendering English syntax correctly. More often than not, when several words appeared in a translation in the right order, the effect on human judgments of Adequacy was insignificant. Second, because of common *n*grams like "of the" and "Xinhua News Agency", automatic evaluation measures that placed heavier emphasis on matching longer *n*grams had higher co-variance with the human scores.

When we ran the same experiments on the Arabic corpus, we found very little difference between the various automatic MT evaluation measures, in terms of Spearman correlation with human judgments.[6] Exploratory data analysis revealed that the quality of an Arabic MT system correlates very highly with whether it outputs correct case information. Therefore, an automatic measure can perform well on this corpus simply by assigning high scores to candidate translations that match the case of their references. Since all the measures we compared are essentially based upon string matching, they are all good at measuring the quality of case matches. So, the differences between the automatic measures are overshadowed by case matching, and all other criteria are insignificant in comparison. We cannot conclude from the above, however, that all the automatic measures are equally good. Our results on Chinese prove otherwise. Without other MT evaluation corpora to analyze, we cannot be sure that the high predictive power of case information on this corpus is no more than coincidental.

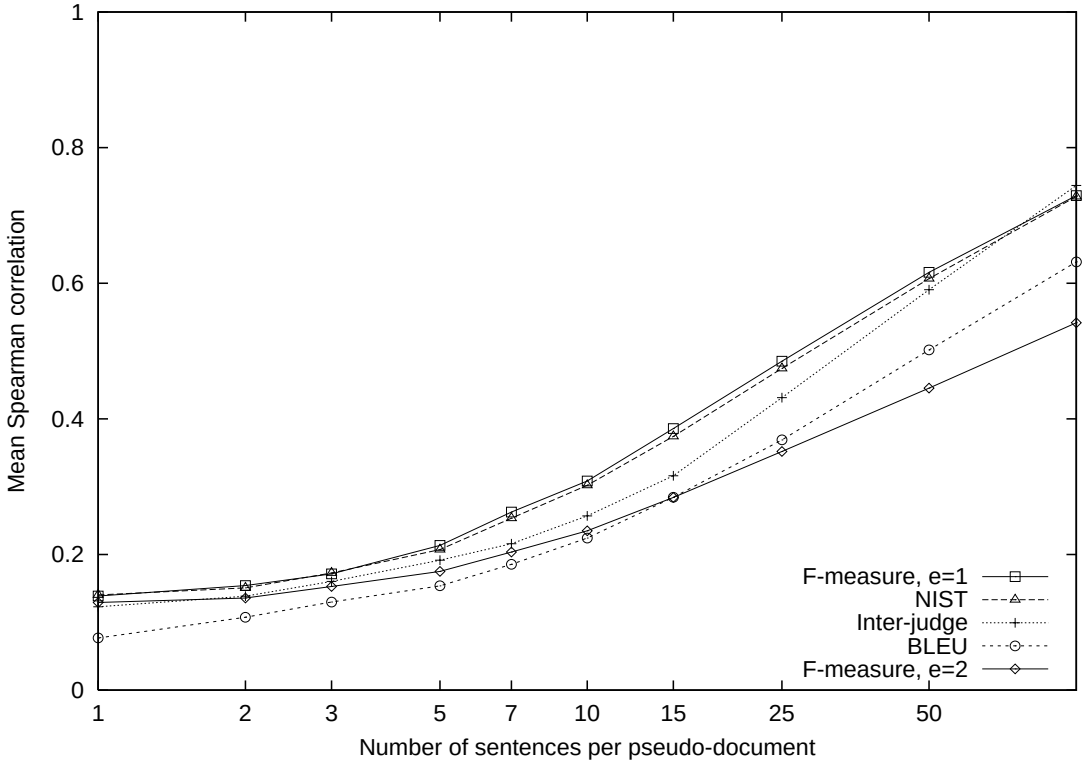
To gain additional insight from our Arabic corpus, we re-ran our experiments after lowercasing and stemming all the candidate and reference texts. The Spearman correlations with Adequacy of the various MT evaluation measures are shown in Figure 5. Figure 5(a) shows the correlations of the measures using a single reference, whereas Figure 5(b) shows the correlations of the measures both using a single reference and using three references. On Fluency, the measures have uniformly higher correlations and the same relative rankings. The relative reliabilities of the various automatic measures on the Arabic corpus largely concur with our results on the Chinese corpus.

Additional references generally improve correlation, but Figure 5(b) shows an anomaly: On longer documents, BLEU correlates worse against three

[6] However, BLEU was consistently much worse on shorter documents.

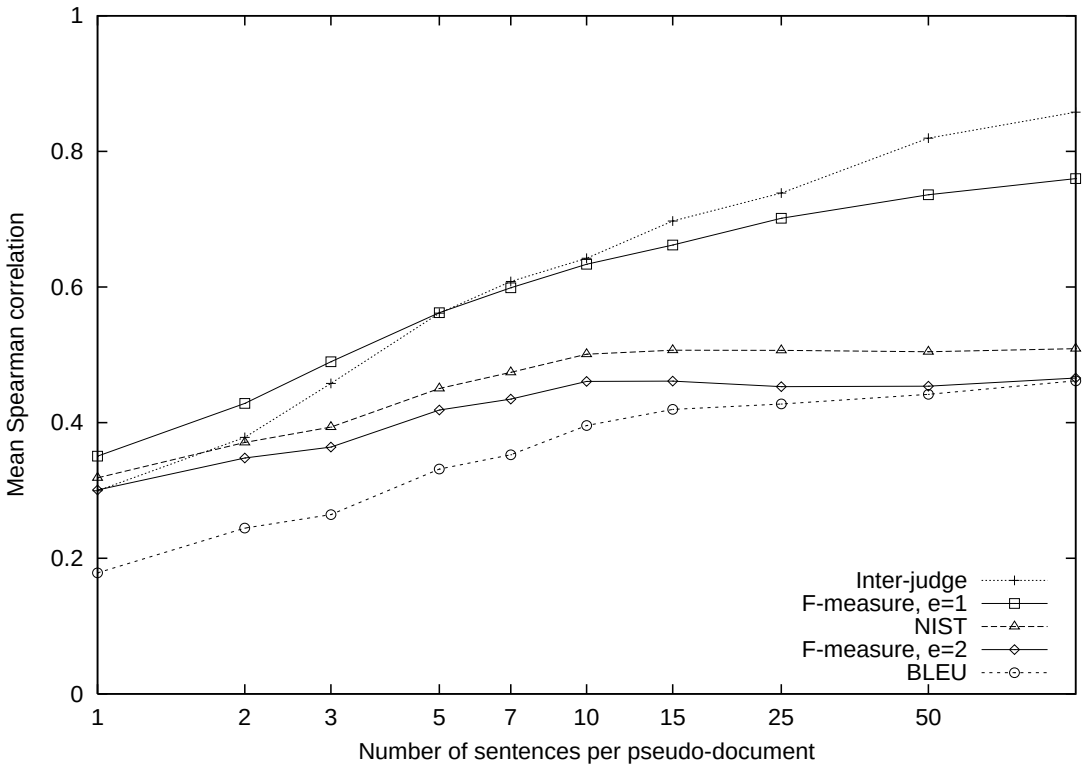


(a)

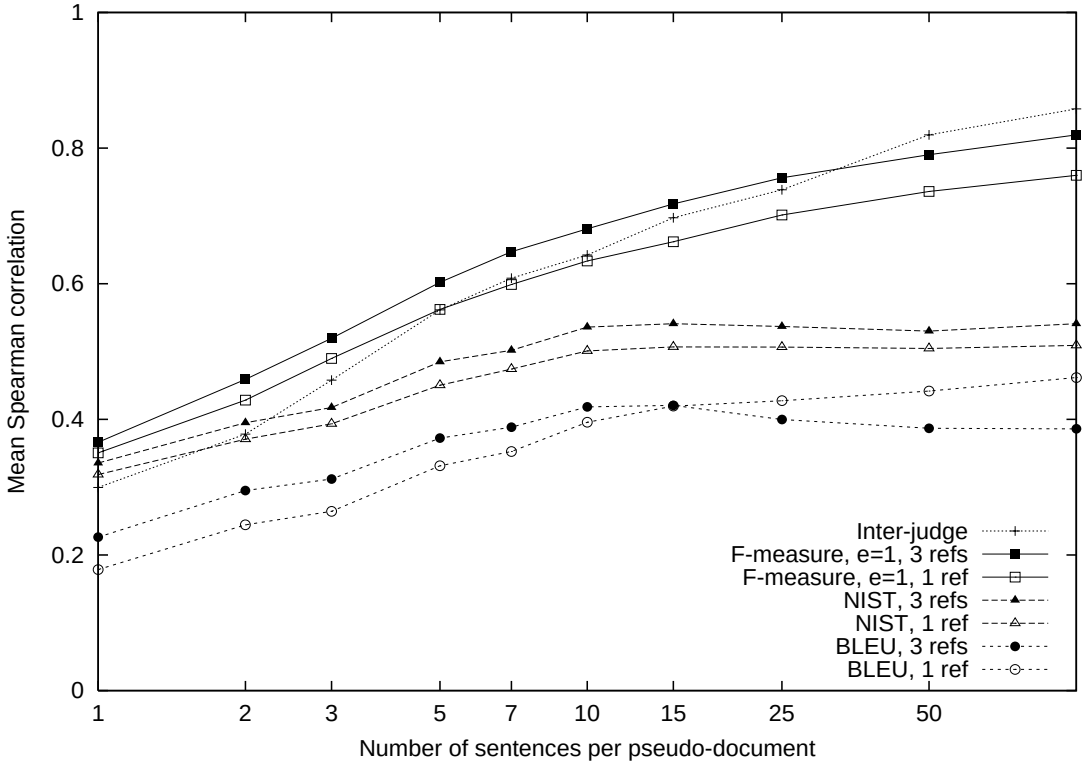


(b)

Figure 4: Spearman correlation on the (unstemmed, case preserved) Chinese corpus using a single reference with:
(a) Adequacy and (b) Fluency.
All correlation differences of 0.01 or more between the automatic evaluation measures are statistically significant using the Wilcoxon signed ranks test with $\alpha = 0.999$.



(a)



(b)

Figure 5: Spearman correlation on the (stemmed, lowercased) Arabic corpus with Adequacy.
All correlation differences of 0.004 or more between the automatic evaluation measures are statistically significant using the Wilcoxon signed ranks test with $\alpha = 0.999$.
For clarity, F-measure with $e = 2$ curves are omitted in (b); The F-measure with $e = 2$ 3ref curve falls between the NIST 1ref and 3ref curves.

references than against only one reference. This is because three references are more likely to include "distracting" *n*grams than a single reference.

We were also surprised to find that some of the automatic measures correlate less well with human judgments on longer documents. It turns out that the correlation estimates on short documents are slight overestimates. Our explanation is the same as for the instances of lower correlation using multiple references: Shorter documents are less likely to include any of the longer matching *n*grams that make the automatic measures diverge from the manual judgments. It is well-known that using more sentences and more references increases the reliability of MT evaluation. Our results show that the same is true for the reliability of the *evaluation of* MT evaluation measures.

5 Conclusions

Our research has raised more questions than it answered. There are many ways to evaluate MT and many ways to ascertain the reliability of automatic MT evaluation measures. More data and more rigorous analysis is necessary to pinpoint the salient variables. What works on one corpus might not work on another. MT evaluation research should be particularly wary of evaluation measures with parameters tuned to particular corpora. Such measures can overfit their objective function, and give misleading rankings on previously unseen corpora. On the other hand, the use of an unbiased language model could improve any of the metrics described herein. (Doddington, 2002)

Different measures might work better when MT systems improve. For example, on good translations the F-measure may do better with $e = 2$ than with with $e = 1$. However, there is no point in comparing MT systems on the correctness of the word order when all MT systems are equally disfluent. (We all hope that this will not always be the case.)

Our most important finding is that, even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgment. Nonetheless, we have shown that machine translation can be evaluated using well-known evaluation measures. In particular, on the data used for the 2002 DARPA MTEval exercises, the F-measure with $e = 1$ proved significantly more reliable than the BLEU and NIST measures. More importantly, the F-measure is easier to understand and to justify in terms familiar to practitioners and consumers of NLP. Our techniques can be used to compute standard evaluation measures for other NLP tasks where reference texts are available, such as text generation and summarization. GTM, the relevant software, is released under a BSD-style license and can be downloaded from `http://nlp.cs.nyu.edu/GTM/`.

Acknowledgment: This research was supported by the DARPA TIDES program, by an NSF CAREER award, and by a gift from Sun Microsystems.

References

C. Brew and H. Thompson (1994) "Automatic Evaluation of Computer Generated Text: A Progress Report on the TextEval Project". In *Human Language Technology: Proceedings of the Workshop (ARPA/ISTO)*:108–113.

T. Cormen, C. Leiserson, R. Rivest and C. Stein (2001), *Introduction to Algorithms*, 2nd Ed., MIT Press.

G. Doddington (2002) "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics". In *Human Language Technology: Notebook Proceedings*:128–132. San Diego.

I. Melamed (1995) "Automatic Evaluation and Uniform Filter Cascades for Inducing *N*-Best Translation Lexicons". In *Third Workshop on Very Large Corpora (WVLC3)*. Boston.

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

K. Papineni, S. Roukos, T. Ward, and W. Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation". In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*:311–318. Philadelphia.

M. Rajman and T. Hartley (2001) "Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores". In *Proceedings of the Workshop on Machine Translation Evaluation: "Who Did What To Whom"*:29–34. Santiago de Compostela, Spain.

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

C. van Rijsbergen (1979) *Information Retrieval*. Butterworths, London, 2nd edition.

J. White, T. O'Connell, and L. Carlson (1993) "Evaluation of machine translation". In *Human Language Technology: Proceedings of the Workshop (ARPA)*:206–210.