Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics

1 Introduction

Evaluation is recognized as an extremely helpful forcing function in Human Language Technology R&D. Unfortunately, evaluation has not been a very powerful tool in machine translation (MT) research because it requires human judgments and is thus expensive and time-consuming and not easily factored into the MT research agenda. However, at the July 2001 TIDES PI meeting in Philadelphia, IBM described an automatic MT evaluation technique that can provide immediate feedback and guidance in MT research. Their idea, which they call an "evaluation understudy", compares MT output with expert reference translations in terms of the statistics of short sequences of words (word N-grams). The more of these N-grams that a translation shares with the reference translations, the better the translation is judged to be. The idea is elegant in its simplicity. But far more important, IBM showed a strong correlation between these automatically generated scores and human judgments of translation quality.[1] As a result, DARPA commissioned NIST to develop an MT evaluation facility based on the IBM work. This utility is now available from NIST and serves as the primary evaluation measure for TIDES-sponsored MT research.[2]

2 N-gram Co-occurrence Scoring

Evaluation using N-gram co-occurrence statistics requires an evaluation corpus of source material along with one (or preferably more) high quality reference translations. Scoring may then be done by tabulating the fraction of N-grams in the test translation that also occur in the reference translations. The IBM algorithm scores MT quality in terms of a weighted sum of the counts of matching N-grams. The IBM algorithm also includes a penalty for translations whose length differs significantly from that of the reference translations. IBM's formula for calculating the score (which IBM has dubbed "BLEU"[1]) is

$$Score = \exp\left\{\sum_{n=1}^{N} w_n \log(p_n) \;-\; \max\left(\frac{L^*_{ref}}{L_{sys}} - 1,\;\; 0\right)\right\} \qquad \text{Eqn 1}$$

where

$$p_n = \frac{\sum_i \left(\begin{array}{l}\text{the number of } n\text{-grams in segment } i,\\ \text{in the translation being evaluated, with}\\ \text{a matching reference cooccurence in segment } i\end{array}\right)}{\sum_i \left(\begin{array}{l}\text{the number of } n\text{-grams in segment } i,\\ \text{in the translation being evaluated}\end{array}\right)}$$

$$w_n = N^{-1}$$

$$N = 4$$

and

L^*_{ref} = the number of words in the reference translation that is closest in length to the translation being scored

L_{sys} = the number of words in the translation being scored

N-gram co-occurrence scoring is typically performed segment-by-segment, where a segment is the minimum unit of translation coherence, usually one or a few sentences. The N-gram co-occurrence statistics, based on the sets of N-grams for the test and reference segments, are computed for each of these segments and then accumulated over all segments. It is intuitive that the smaller the segment, the better the co-occurrence statistics.

Before scoring, the translated text is conditioned to improve the efficacy of the scoring algorithm. This conditioning is applied both to the translation to be scored and to the reference translations. Here are the conditioning actions that are applied (for English):

- Case information is removed. All text is reduced to lower case.
- Numerical information (in terms of sequences of digits, commas and periods) is kept together as single words.
- Punctuation is tokenized into separate words (except for dashes and apostrophes).
- Adjacent non-ASCII words (which occur when source text is transferred to the output) are concatenated into single words.

3 Evaluation of N-gram Scoring

N-gram co-occurrence scoring is an extremely promising technique for efficient evaluation. But the technique needs to be validated and evaluated further with respect to its stability and its ability to predict human quality assessments reliably. In order to perform this validation, several translation corpora were assembled. These are summarized in Table 1.

3.1 Correlation with Human Assessments

The ability to predict human judgment of quality is the sine qua non of any automatic MT score. To this end, there exist human quality scores for each of the translated documents in the corpora listed in Table 1. These scores may then be averaged across documents to generate system-specific scores that indicate the translation quality of the systems. Human assessors were asked to judge translation quality along several different dimensions. For the 1994 corpora there were three dimensions, namely "Adequacy", "Fluency" and "Informativeness". For the 2001 corpus there were only two dimensions, namely "Adequacy" and "Fluency". Although the procedures used in 2001 differed somewhat from the procedures used in 1994[3], the judgments are basically the same:

- For "Adequacy", the translation being evaluated is compared with a high quality reference translation, segment by segment. Each evaluation segment is scored according to how well (how "adequately") the meaning conveyed by the reference translation is also conveyed by the evaluated segment.

[1] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001). "Bleu: a Method for Automatic Evaluation of Machine Translation". This report may be downloaded from URL http://domino.watson.ibm.com/library/CyberDig.nsf/home. (keyword = RC22176)

[2] Visit NIST's MT evaluation web site to download a copy of this utility. The URL is http://www.nist.gov/speech/tests/mt/

[3] The specification used by the LDC for the 2001 human assessment may be accessed from LDC's web site at the URL: www.ldc.upenn.edu/Projects/TIDES/Translation/TranAssessSpec.pdf

– For "Fluency", the translation being evaluated is judged according to how fluent it is. This is done segment by segment, with no reference to what the translation is supposed to convey.
– For "Informativeness", an assessor is asked to answer a set of questions about the content of each document after reading a translation of it. The Informativeness score is then the fraction of questions that are correctly answered.

Table 1 Primary characteristics of the corpora used to study the performance of N-gram co-occurrence based scoring of translation quality.

Description of Corpus	Source language	# of documents	# of human translations	# of MT systems
The 1994 DARPA corpus used to evaluate French-English MT	French	100	2	5
The 1994 DARPA corpus used to evaluate Japanese-English MT	Japanese	100	2	4
The 1994 DARPA corpus used to evaluate Spanish-English MT	Spanish	100	2	4
The 2001 DARPA corpus used for the Chinese-English dry run	Chinese	80	11	6[4]

The correlation between BLEU scores and human assessments of translation quality for the various systems evaluated in the DARPA 1994 and 2001 evaluations are listed in Table 2. In general, there is very strong correlation between human judgments and BLEU. Note however that the correlation for professional translators is much smaller than for machines. Not that the scores for professional translators aren't distinctly better than for machines. They are, as shown in Figure 1. Rather, the lower correlation means that the N-gram score distinctions between professional translations correlate less well with human judgments than those between different machine translations. A possible explanation for this difference in correlation is that differences between professional translators are far more subtle and thus less well characterized by N-gram statistics.

Other than the low correlation scores for the human translations, the correlations between human judgments and N-gram scores are above 90% for all of the comparisons, with the exception of the fluency score for Japanese. A possible explanation for this low correlation is simply that the Japanese systems seemed to be very similar in quality. Thus the uncorrelated differences account for more of the between-system variance.

Figure 2 shows a scatter-plot of N-gram scores versus human judgments of Adequacy and Fluency for the 6 commercial Chinese-to-English MT systems. Note that, while the correlation is quite high, there are some differences in judgment. Among them is one reversal in ranking with respect to Adequacy, albeit attributable to relatively minor differences in score.

Table 2 Correlation between IBM's BLEU scores and human assessments. The N-gram scores were produced using all (2) of the reference translations for the 1994 corpora MT systems and 8 reference translations for the 2001 Chinese corpus.

The Corpus	The Systems	Adequacy (%)	Fluency (%)	Informativeness (%)
1994 French Corpus	5 MT Systems	95.7	99.7	91.4
1994 Japanese Corpus	4 MT Systems	97.8	85.6	98.3
1994 Spanish Corpus	4 MT Systems	97.5	97.2	94.3
2001 Chinese Corpus	6 Commercial MT Systems	95.2	97.1	-
	7 Professional Translators	70.5	16.6	-

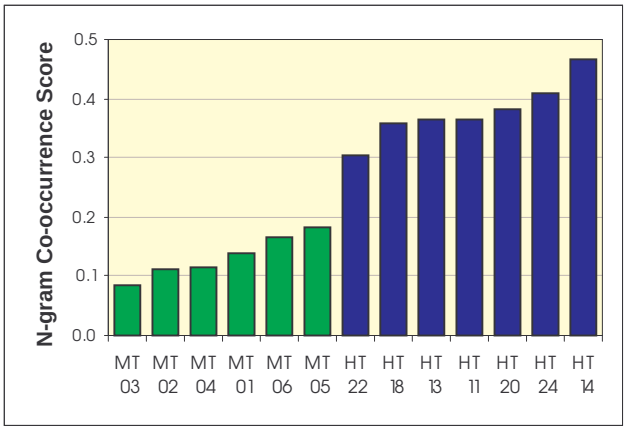

Figure 1 Rank-ordered N-gram co-occurrence scores for the 6 commercial MT systems and 7 professional translators in the 2001 Chinese-English dry run evaluation.

3.2 Sensitivity and Consistency

Ideally, a good score is both sensitive and consistent. That is, a good score will be able to distinguish between systems of similar performance, and this difference will be essentially unaffected by the selection of translations used for reference or documents used for scoring.[5] To measure the sensitivity and consistency of N-gram co-occurrence scoring, we examined the variability of system scores with respect to the choice of documents and the choice of reference translations used to compute the scores. To do this we used the F-ratio measure, namely the between-system score variance divided by within-system score variance. The between-

[4] These 6 systems are commercial MT systems. There were also 9 research MT systems included in the evaluation. The research systems were not included in the analysis, however, because human assessments were performed only on the output from commercial systems.

[5] For N-gram co-occurrence scoring, such reliable indication of performance can be expected only if the reference translations are all of high quality and the choice of documents is within the same distribution of genre and other relevant parameters.

system variance is the variance of average system scores across different systems, and the within-system variance is the variance of document scores for a given system, computed across different documents and different reference translations and then pooled over all systems. Thus the greater the F-ratio, the better the score.

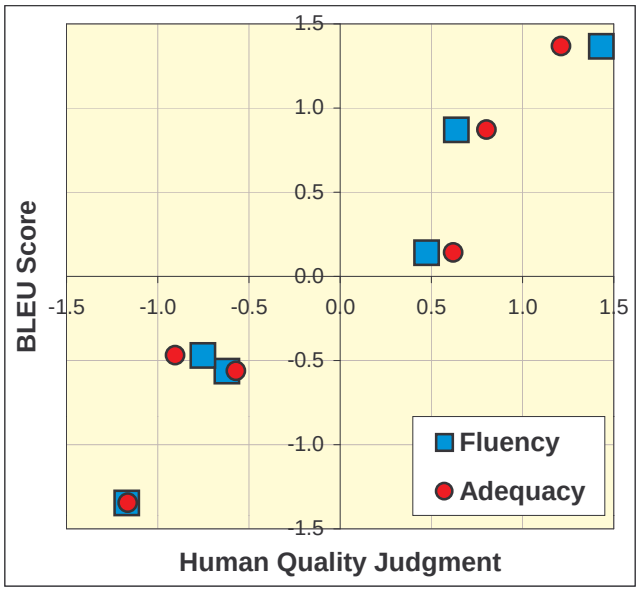

Figure 2 Scatter-plot of IBM's BLEU scores versus human judgments of Adequacy and Fluency for the 6 commercial Chinese-to-English MT systems. Scores were normalized to zero mean and unit variance before plotting.

Table 3 shows a comparison of F-ratios for human judgments and N-gram co-occurrence scores for all four corpora of this study. For purposes of cross-corpus comparison, the number of reference translations used to compute the co-occurrence score was held constant and equal to 2 for all of the corpora.

Note that in general the stability of the co-occurrence scores compares favorably to that of the human judgments. Note also that the F-ratios for the Japanese corpus are significantly poorer than for the French and Spanish 1994 corpora, for both human judgments as well as N-gram scores. By way of explanation, the Japanese MT systems were all quite close in quality, with a between-system score variance (of human scores) that was well over 4 times smaller than either French or Spanish. Also, note the relatively low correlation for Fluency for Japanese in Table 2. Nonetheless, the correlation for Adequacy remained high for Japanese.

On the other hand, note that the correlation between human and N-gram scores was very much smaller for human translations of Chinese than for machine translations. In this case, however, the spread of quality for human translations was comparable to the spread for machines, with between-human score variance (of human scores) being > 50% of N-gram score variance for Adequacy and > 80% of N-gram score variance for Fluency.

There are two sources of variance in N-gram co-occurrence scores shown in Table 3, namely variance due to the use of different sets of documents and variance due to the use of different reference translations. For judging relative translation quality, however, variance from the use of different reference translations may not be so important. This is because the variance due to choice of reference manifests itself primarily as a score offset that affects all systems similarly. Thus the relative ranking of systems remains largely unchanged, as illustrated in Figure 3.

Table 3 Comparison of F-ratios for human judgments versus IBM's BLEU scores.[6] F-ratios for reference variation are available only for the Chinese corpus because this is the only corpus with a number of reference translations that is large enough to support such analysis.

The Corpus	The Systems	F-ratios for Human Judgments			F-ratios for BLEU Scores	
		Adequacy	Fluency	Informativeness	Document variation	Reference variation
' 94 French Corpus	All MT Systems	86.7	82.4	36.1	213.4	-
' 94 Japanese Corpus	All MT Systems	8.4	14.2	2.8	45.5	-
' 94 Spanish Corpus	All MT Systems	62.5	61.5	34.3	226.0	-
2001 Chinese Corpus	Commercial MT Systems	53.7	44.6	-	42.5	45.1
	Professional Translators	19.8	39.5	-	26.5	2.6

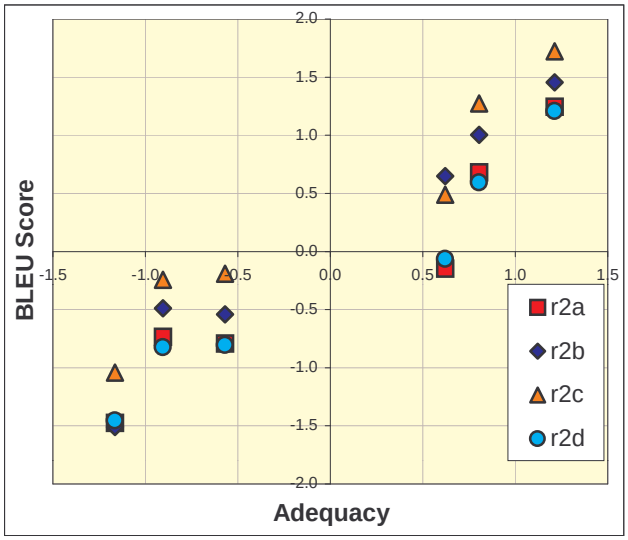

Figure 3 Scatter-plot of IBM's BLEU scores versus human Adequacy judgments for the 6 commercial Chinese-to-English MT systems. Four different sets of BLEU scores are shown, corresponding to the use of four different sets of 2 reference translations for each of four experiments. Scores were normalized to zero mean and unit variance (over all four experiments) before plotting.

[6] There were a total of 11 judges used for the 2001 Chinese corpus. The scores for each of the judges for this corpus were normalized to standard mean and variance individually for each judge. This normalization improved the F-ratios for human judgments by about a factor of 2.

4 The NIST Score Formulation

Several possible variations of N-gram scoring suggest themselves upon reflection on the characteristics of N-gram co-occurrence scores:

- First, note that the IBM BLEU formulation uses a geometric mean of co-occurrences over N. This makes the score equally sensitive to proportional differences in co-occurrence for all N. As a result, there exists the potential of counterproductive variance due to low co-occurrences for the larger values of N. An alternative would be to use an arithmetic average of N-gram counts rather than a geometric average.
- Second, note that it might be better to weight more heavily those N-grams that are more informative – i.e., to weight more heavily those N-grams that occur less frequently, according to their information value. This would, in addition, help to combat possible gaming of the scoring algorithm, since those N-grams that are most likely to (co-)occur would add less to the score than less likely N-grams.

Information weights were computed using N-gram counts over the set of reference translations, according to the following equation:

$$Info(w_1 \ldots w_n) = \log_2\left(\frac{\text{the \# of occurrences of } w_1 \ldots w_{n-1}}{\text{the \# of occurrences of } w_1 \ldots w_n}\right) \quad \text{Eqn 2}$$

Table 4 compares F-ratios and Correlation values for individual N-gram co-occurrence scores for commercial translation systems evaluated on the 2001 Chinese-to-English corpus. Note that the information-weighted N-gram counts provide superior F-ratio and correlation performance for N = 1, about the same performance for N = 2, and poorer performance for N > 2. The poorer performance for the higher values of N may be due to poor estimation of N-gram likelihoods.[7] Note also that the F-ratios for single N-grams, both unweighted and information-weighted, are greater than the F-ratios for IBM's BLEU formulation for N = 1 and 2. Further, the single N-gram correlations also are comparable to the BLEU correlations for N = 1 and 2.

Table 4 F-ratios and Correlation values for individual N-gram co-occurrence scores for commercial translation systems for the 2001 Chinese-to-English corpus. Eight reference translations were used to compute these statistics.

	Unweighted			Information-weighted		
N-gram	F-ratio	Adequacy Correlation (%)	Fluency Correlation (%)	F-ratio	Adequacy Correlation (%)	Fluency Correlation (%)
1	98.6	97.7	97.6	149.2	99.0	97.3
2	94.5	97.1	98.4	97.5	96.1	97.7
3	46.1	94.8	96.3	39.9	84.5	90.4
4	22.4	93.0	95.0	19.5	87.8	92.7
5	9.5	94.7	95.7	5.5	87.6	91.9

[7] Large amounts of data are required to estimate N-gram statistics for N > 2. In the current implementation, however, the N-gram statistics are computed only from the reference translations for the evaluation corpus.

Based on the superior F-ratios of information-weighted counts and the comparable correlations, a modification of IBM's formulation of the score was chosen as the evaluation measure that NIST will use to provide automatic evaluation to support MT research. NIST's formula for calculating the score is

$$Score = \sum_{n=1}^{N}\left\{\frac{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{that co-occur}}} Info(w_1 \ldots w_n)}{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{in sys output}}} (1)}\right\} \cdot \exp\left\{\beta \log^2\left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1\right)\right]\right\} \quad \text{Eqn 3}$$

where

β is chosen to make the brevity penalty factor = 0.5 when the # of words in the system output is $2/3^{rds}$ of the average # of words in the reference translation,

$N = 5$

and

$\bar{L}_{ref}$ = the average number of words in a reference translation, averaged over all reference translations

L_{sys} = the number of words in the translation being scored

Notice that, in addition to the calculation of the co-occurrence score itself, a change was also made to the brevity penalty. This change was made to minimize the impact on the score of small variations in the length of a translation. This preserves the original motivation of including a brevity penalty (which is to help prevent gaming the evaluation measure) while reducing the contributions of length variations to the score for small variations. Figure 4 gives a comparison of the two brevity penalty factors.

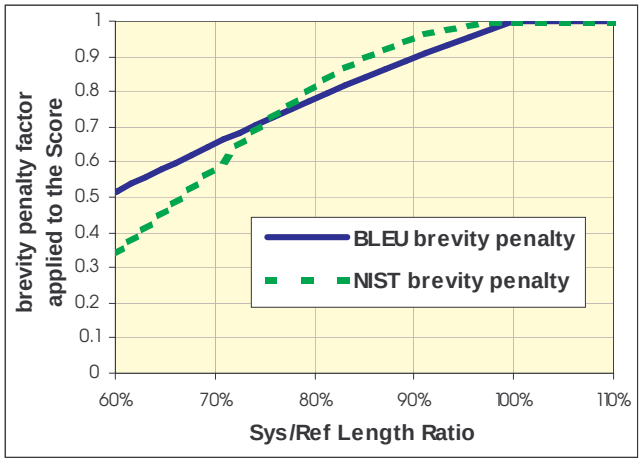

Figure 4 Comparison of the BLEU and NIST brevity penalty factors.

The NIST evaluation score is compared with IBM's original BLEU score in Figure 5 and Figure 6. Figure 5 demonstrates that the NIST score provides significant improvement in score stability and reliability for all four of the corpora studied. Figure 6 demonstrates that, for human judgments of Adequacy, the NIST score correlates better than the BLEU score on all of the corpora. For Fluency judgments, however, the NIST score correlates better than the BLEU score only on the Chinese corpus. This may be a mere random statistical difference between corpora. Or alternatively, this may be a consequence of different human judgment criteria or procedures. (The Chinese-to-English translations were judged at

LDC using a different procedure than that used by John White at PRC for the 1994 corpora.)

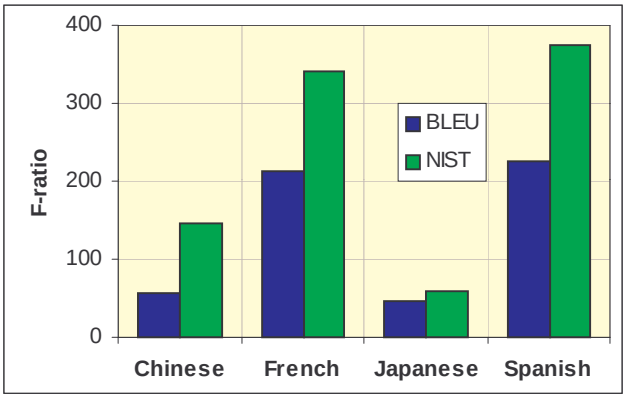

Figure 5 F-ratio comparison of the BLEU and NIST scores for document variance for the four corpora studied.

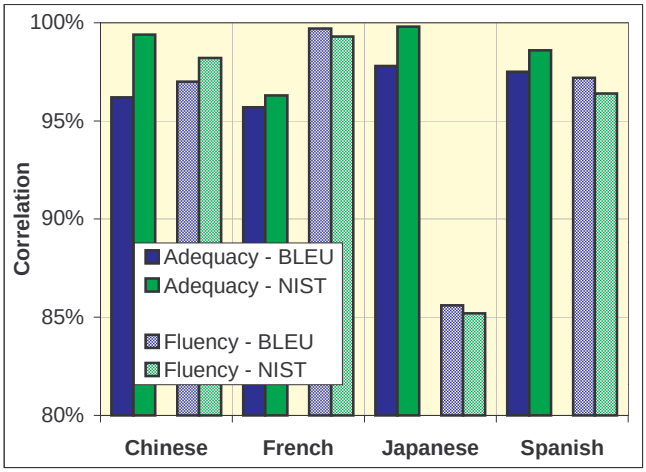

Figure 6 Comparison of the correlation of BLEU and NIST scores with human judgments for the four corpora studied.

5 Performance vs. Parameter Selection

In this section, the performance of the NIST scoring algorithm is analyzed as a function of several important parameters and conditions. Performance is analyzed in terms of the score's F-ratio the score's correlation with human judgment.

5.1 Performance as a function of source

The Chinese-to-English evaluation corpus included data from three sources, as shown in Table 5. Zaobao is a Chinese newswire from Singapore, and the Voice of America data comprises manual transcriptions of broadcasts in Mandarin. Since MT performance is sensitive to genre and style, human assessments of translation quality are broken out according to source and shown in Figure 7 both for professional and machine translations. From this figure it appears that the quality of professional translations of Voice of America transcripts is better than translations of newswire. This might be explained if VOA broadcasts were generally simpler language. The machine translations don't appear to exhibit marked differences between sources, although Fluency assessments of VOA broadcasts are poorer than those of newswire, this despite the better performance on professional translations.

Table 5 The three sources of data for the 2001 DARPA Chinese evaluation corpus.

Source	Number of Documents	Number of Words
Xinhua newswire	27	8411
Zaobao newswire	27	9083
Voice of America transcripts	26	6746

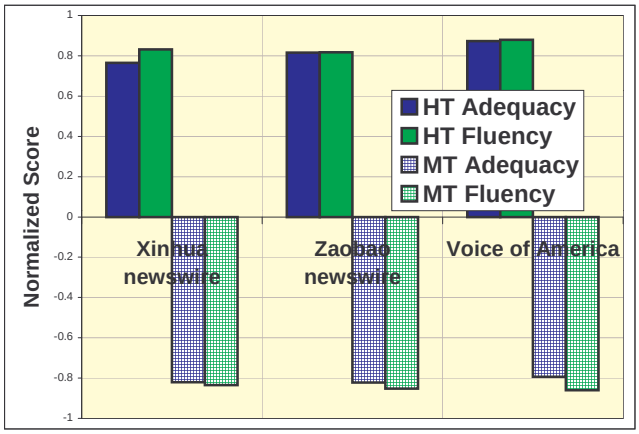

Figure 7 Average human assessment scores for 6 professional translations (denoted "HT") and 6 commercial off -the-shelf MT systems (denoted "MT") for the Chinese corpus, broken out according to source.

More interesting is the relative scoring of different MT systems on the different sources, shown in Figure 8. This figure is a scatter-plot of Adequacy scores for translations of Xinhua newswire and Voice of America transcripts versus Adequacy scores for Zaobao translations. This demonstrates that, while there is a loose agreement in the relative ranking of systems on different sources, the correlation between human assessments on the difference sources is much poorer than the correlation between human assessments and NIST scores, given the source.

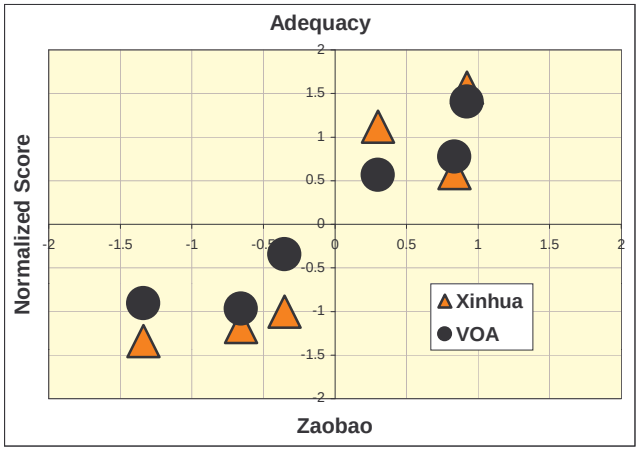

Figure 8 A scatter plot of average human Adequacy scores for 6 MT systems. Average scores for Xinhua and VOA are plotted versus average scores for Zaobao.

A scatter plot of NIST scores for the 6 commercial MT systems versus human Adequacy assessments is shown in Figure 9. Note that the correlation between the NIST score and human Adequacy assessment is much better than the correlation between human

Adequacy assessments between difference sources. This contrast is shown quantitatively in Table 6.

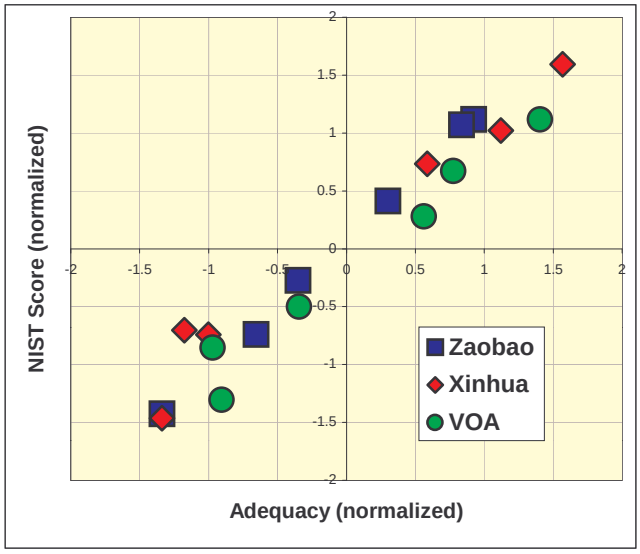

Figure 9 Scatter plot of NIST scores versus human Adequacy scores for the 6 commercial Chinese MT systems, plotted for each of the three different sources of data.

Table 6 Correlations (in percent) of human Adequacy scores for the three sources of data, compared with correlations between human Adequacy scores and NIST scores for each source, for the 6 commercial Chinese MT systems

Source	Xinhua	Zaobao	VOA	NIST score
Xinhua newswire	100.0	86.3	98.3	93.0
Zaobao newswire	-	100.0	91.5	99.8
Voice of America transcripts	-	-	100.0	93.9

5.2 Performance vs. number of references

Because of the wide variety of possible valid translations, the number of reference translations is generally regarded as an important factor in producing valid scores – the more reference translations, the better the performance of the co-occurrence score. However, as shown in Figure 10 and Figure 11, increasing the number of references appears to yield only modest improvements in evaluation performance. Specifically, there appears to be no significant improvement in the correlation with human judgments with the use of more than 1 reference translation. And the increase in F-ratio with increasing numbers of references is modest, at least for document variance. Although there is a great increase in F-ratio for the use of 4 references, this is quite likely an artifact attributable to the small sample of reference sets used in the experiment.[8]

[8] The experiment in which the number of reference translations was varied was structured as follows: A total of eight reference translations were used. These 8 references were divided into 8 sets of one reference, 4 sets of two references, 2 sets of four references, and 1 set of 8 references. This left only one degree of freedom for computing the variance for 4 references, and none at all for 8 references (which is why there is no bar shown for the 8 reference case).

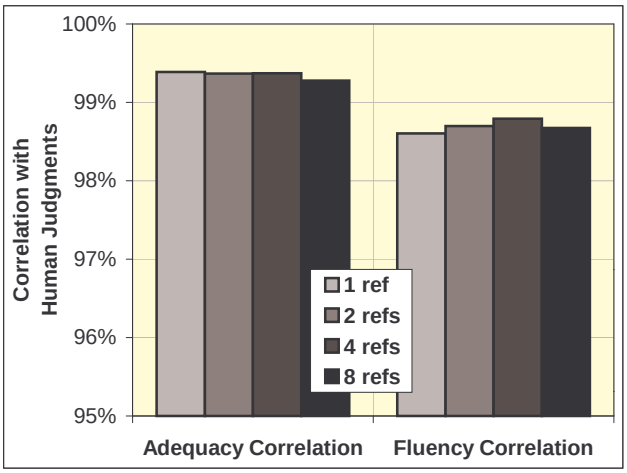

Figure 10 Adequacy and Fluency correlation statistics versus the number of reference translations used for scoring, for NIST scores for the 6 commercial Chinese-to-English MT systems.

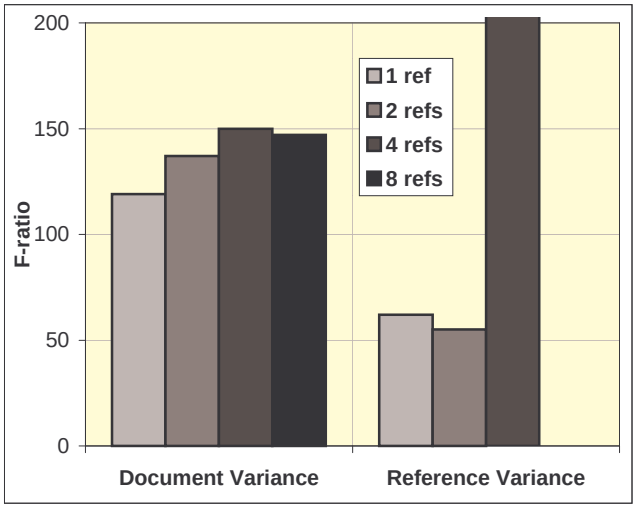

Figure 11 F-ratio statistics versus the number of reference translations used for scoring, for the NIST score on the Chinese-to-English evaluation corpus.

5.3 Performance versus segment size

Segment size is an important consideration. Intuitively, the shorter the segment over which co-occurrence is restricted, the better an N-gram co-occurrence score will perform. But the smaller the segments are made, the more work there is in establishing and maintaining the segments. More importantly, restricting the translation to be synchronous with the segmentation is an unnatural constraint that becomes more onerous as the segments become shorter. Obviously, segments should be no less than one sentence in length. And it would be ideal if the scoring algorithm performed well with no document-internal segmentation at all.

The effect of segmentation was studied by joining each adjacent pair of segments into single segment, thus effectively doubling the size of a segment. (Final odd segments at the end of a document were left as is.) This was done multiple times for the 2001 Chinese-to-English corpus until each document contained only a single segment. These modified document sets were then scored. The results are shown in Figure 12 and Figure 13. It is encouraging to see that correlation performance degrades only slightly, even at

271 words per segment, which corresponds to one segment per document. The decline in F-ratio is more pronounced, but still remains above 100 at 1 segment per document. Of course, using only one segment per document must be expected to yield progressively poorer performance as the average number of words in a document increases.

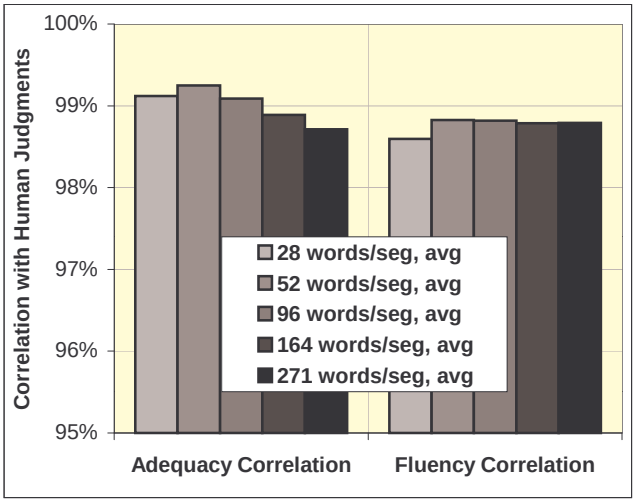

Figure 12 Adequacy and Fluency correlations statistics versus segment size, for NIST scores for 6 commercial Chinese-to-English MT systems.

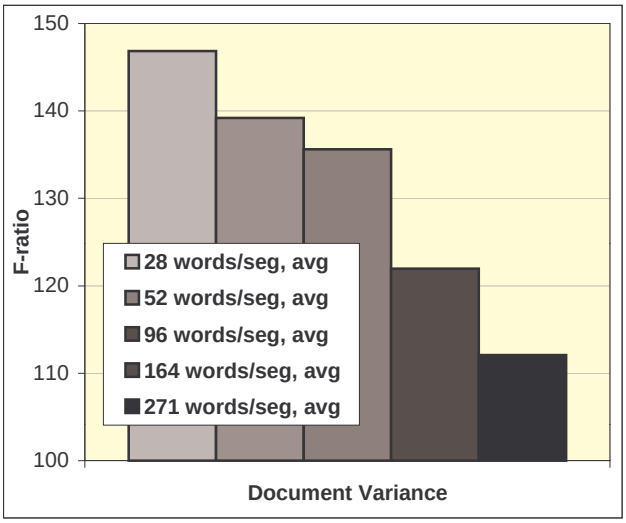

Figure 13 F-ratio versus segment size, for NIST scores for 6 commercial Chinese-to-English MT systems.

5.4 Performance with more language training

Table 4 shows that, while information-weighted N-gram counts are superior to unweighted counts for unigrams, information-weighted counts perform less well for N > 1. This may be attributable to poor information estimates that arise from using only the reference translations as a corpus to estimate N-gram likelihoods. To obtain reasonably accurate estimates, a much larger corpus would be required. To see if more accurate estimates of likelihoods might improve score performance, an auxiliary database comprising the entire English language subset of both the TDT2 and TDT3 corpora[9] was used to estimate N-gram likelihoods. Table 7 show the equivocal results of this experiment. While using the TDT corpus to estimate N-gram likelihoods yields minor (probably insignificant) improvements in the correlation of the NIST score with both Adequacy and Fluency judgments, this is accompanied by a (probably significant) decline in the F-ratio. Regarding individual N-grams, the table shows that there is minor improvement in the F-ratio for all N-grams except for N = 1 where there is a significant reduction in F-ratio. And while the correlation with human judgments is better for N = 2 and 3, it is worse for N = 4 and 5. (Even the TDT corpora may be inadequate to supply meaningful likelihood estimates for N > 3, especially considering the change in topics when switching from the TDT sources to the Chinese MT sources.)

Table 7 F-ratios and Correlation values for individual N-grams and the overall NIST score given different information weighting sources. Values are for commercial translation systems for the 2001 Chinese-to-English corpus. Eight reference translations were used to compute these statistics.

N-gram	Information weights computed from the evaluation corpus			Information weights computed from TDT2 and TDT3		
	F-ratio	Adequacy Correlation (%)	Fluency Correlation (%)	F-ratio	Adequacy Correlation (%)	Fluency Correlation (%)
1	149.2	99.0	97.3	115.4	98.3	96.0
2	97.5	96.1	97.7	105.4	99.2	98.8
3	39.9	84.5	90.4	48.1	92.0	94.9
4	19.5	87.8	92.7	21.2	84.8	89.3
5	5.5	87.6	91.9	5.8	82.2	87.6
NIST	**146.8**	**99.3**	**98.7**	**121.5**	**99.5**	**98.8**

In using the corpus-based likelihoods and resultant information calculations, it often happens that higher order N-grams don't contribute to the score. This occurs whenever the N-1 gram predicts the N-gram without error – i.e., whenever there are the same number of occurrences of both, usually one occurrence. In this case there is no (additional) information conveyed by the N^{th} word in the N-gram and the information is zero. Since individual N-grams appear to perform better unweighted than weighted, it is possible to force a minimum information contribution for all N-gram tokens by adding a certain minimum number of occurrences to the N-1 gram in Eqn 2. This was attempted for a number of values for the minimum number of occurrences of the N-1 gram. Unfortunately, and rather surprisingly, the performance of the score was virtually unaffected by such changes.

5.5 Performance with preservation of case

The assumption has been that removing case information would provide better N-gram scoring. This is not necessarily true, however. Furthermore, there are languages (other than English) where an argument can be made that case information might be more important than for English. With this in mind, an experiment was conducted to compare scoring performance with and without case information preserved in the translation. The results of this comparison are shown in Table 8. This table shows clearly that

[9] http://www.ldc.upenn.edu/Catalog/TDT.html

there is very little difference in scoring performance, whether case information is preserved or removed.

5.6 Performance with reference normalization

The score variance attributable to choice of reference translations appears to be an offset that applies roughly equally to all systems. Thus it might be the case that this offset might be at least partially mitigated by dividing the system score by the average reference score. However, when this normalization was attempted, the F-ratio remained essentially unchanged. (Correlation of system scores with human assessments is unaffected by this normalization, because the normalization applies to all system scores equally.)

Table 8 A comparison of F-ratio and of Adequacy/Fluency correlations with and with case information, computed for the 6 commercial MT systems on the Chinese corpus using 8 reference translations.

	F-ratio	Adequacy Correlation (%)	Fluency Correlation (%)
Case Info Removed	147	99.3	98.7
Case Info Preserved	148	99.0	98.9

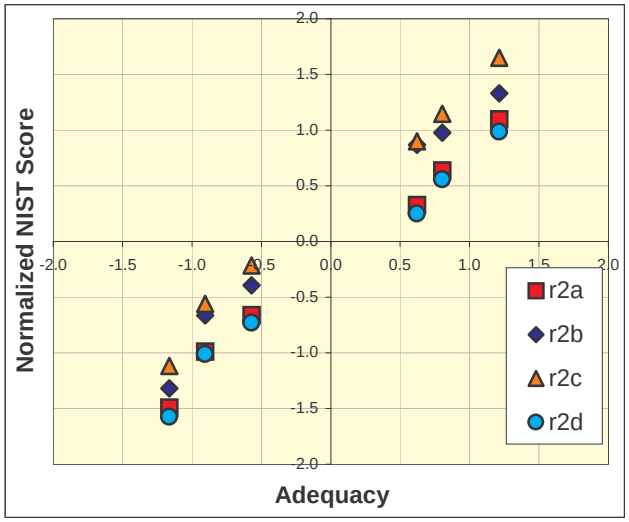

Figure 14 Scatter-plot of NIST scores versus human Adequacy judgments for the 6 commercial Chinese-to-English MT systems. Four different sets of NIST scores are shown, corresponding to the use of four different sets of 2 reference translations for each of four experiments. Scores were normalized to zero mean and unit variance (over all four experiments) before plotting.

6 The NIST MT Evaluation Facility

NIST now provides an evaluation facility that may be used to support MT research for translating various languages into English. This facility includes an N-gram co-occurrence scoring utility, which may be downloaded and used as desired by research sites. This utility requires a corpus of source documents and a corresponding set of one or more reference translations of each source document. The LDC offers corpus support for some source languages, and a research site's own corpora may be used, of course. In addition, formal evaluations of technology are supported with an email-based automatic evaluation utility. In this case, no reference translations are provided. Instead, each participating site receives the source documents, translates the documents, and then sends the translations to be evaluated to NIST via email. NIST then automatically scores the proffered translations and returns the results by email. Details of procedures and data formats are available from the NIST MT web site.[10]

[10] http://www.nist.gov/speech/tests/mt