Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results

Kishore Papineni Salim Roukos Todd Ward
IBM T. J. Watson Research Center
{papineni,roukos,toddward}@us.ibm.com

John Henderson Florence Reeder
MITRE
{jhndrsn.freeder} @mitre.org

ABSTRACT

We describe two metrics for automatic evaluation of machine translation quality. These metrics, BLEU and NEE, are compared to human judgment of quality of translation of Arabic, Chinese, French, and Spanish documents into English.

1. INTRODUCTION

Machine Translation (MT) evaluation has long been considered something of a black art. Therefore, MT evaluation (MTE) metrics have been historically conspicuous in their absence. Evaluations have tended to be holistic scoring by committees of raters on the basis of the somewhat vaguely specified parameters of intelligibility and fidelity ([8], [1], [4]). Intelligibility and fidelity have been called fluency and adequacy respectively. While these evaluations represent valid judgments of MT quality, their proper administration and application have been problematic and too expensive to support rapid development paradigms [3]. Furthermore, comparing successive evaluations and evaluations at different sites has often been a fool's errand. Rating committees change, corpora are different, and evaluation criteria are often underspecified and subjective at best. A typical question one encounters during an attempt to compare sets of evaluations could be as simple and as complex as "Does a score of 4 on a 1-5 scale mean the same thing as a 6 on a 1-7 scale?" On the other hand, human judgments are the final indicator of system acceptance and cannot be ignored.

Examination of past MTE strategies and the uses of MT and MTE has yielded some desired characteristics for current MT evaluations: MTE measures should be automated or extremely cheap; they must be replicable; they should correlate well with human judgments of quality; they should be predictive of the possible uses of imperfect MT output; and they should be diagnostic for the improvement of systems. Recently, a number of interesting, useful and rational metrics have appeared. Two of these metrics, BLEU and NEE, meet the desired criteria for MTE. By virtue of the electronic corpora used for their refinement, automation and replicability, we are able to look at both metrics and see what they tell us about different systems and MT. This paper examines these two metrics on four different bilingual corpora.

Proceedings of HLT 2002, Second International Conference on Human Language Technology Research, M. Marcus, ed., Morgan Kaufmann, San Francisco, 2002.

2. BACKGROUND

Before describing the experiments performed, we will describe each of the metrics. First is BLEU[7] and the second is the Named-Entity translation Evaluator, NEE (descended from [5]).

2.1 BLEU

The closer a machine translation is to a professional human translation, the better it is. This is the central idea behind BLEU. To judge the quality of a machine translation, one measures its closeness by a numerical metric to one or more reference human translations. Thus, BLEU requires two ingredients:

1. a numerical "translation closeness" metric
2. a corpus of good quality human reference translations

These reference translations can be reused over and over again and incur only a one-time startup expense. Each evaluation can be accomplished in seconds.

BLEU is fashioned after the highly successful *word error rate* metric used by the speech recognition community, appropriately modified for *multiple* reference translations and allowing for legitimate differences in word choice and word order. The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. The baseline BLEU metric uses two key concepts in deciding closeness. The first is a *modified n*-gram precision and the second is a recall on length of the translation. Both are reminiscent of the familiar precision and recall used in information retrieval literature, but have been adapted for *multiple* reference translations.

The intuition behind the recall on length is that a one-word candidate translation is bad if all the reference translations have ten or more words, for example. A good translation roughly matches the reference translations in length. Traditionally, precision has been paired with recall to overcome such length-related problems. However, BLEU considers *multiple* reference translations, each of which may use a different word choice to translate the same source word.

Furthermore, a good candidate translation will only use (recall) one of these possible choices, but not all. Indeed, recalling all choices leads to a bad translation. To account for recall on concepts among multiple reference translations with different word choice and word order, one could align the reference translation for concepts. Such alignment is complicated. Instead, the recall on length couples with the modified *n*-gram precision to achieve the same effect in a very simple way. The recall on length is implemented as a penalty on the mismatch between the lengths of the candidate translation and the reference translations.

Modified *n*-gram precision counters against cheating by repeating a phrase in the reference translation to match the expected length of the reference translations. Thus, to score high with BLEU the candidate translation must match a reference translations in length and then in word choice (determined by modified unigram precision) and in word order (determined by higher order modified n-gram precisions). BLEU uses a geometric average of *n*-gram precisions at various *n* (1 to 4) multiplied by the length penalty. The details are below.

We first explain the modified unigram precision. To compute modified unigram precision, we first count the maximum number of times a word occurs in any single reference translation. Next, we clip the total count of each candidate word by its maximum reference count:

$$Count_{clip} = min(Count; MaxRefCount).$$

In other words, we truncate each word's count, if necessary, to not exceed the largest count observed in any single reference for that word. We then add these clipped counts up, and divide by the total (undipped) number of candidate words. This gives us the modified unigram precision. Modified *n*-gram precision p_n for any *n* is defined similarly.

When there is only one sentence in the entire test corpus, the above procedure is unambiguous. How do we compute modified n-gram precision on a multi-sentence test set? Although one typically evaluates MT systems on a corpus of entire documents, our basic unit of evaluation is the sentence. A source sentence may translate to many target sentences, in which case we abuse terminology and refer to the corresponding target sentences as a "sentence."

We first compute the *n*-gram matches sentence by sentence. Next, we add the clipped *n*-gram counts for all the candidate sentences and divide by the number of candidate n-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$\frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count(n\text{-}gram)}.$$

In other words, we use a word-weighted average of the sentence-level modified precisions rather than a sentence-weighted average. As an example, we compute word matches at the sentence level, but the modified unigram precision is the fraction of words matched in the entire test corpus.

Candidate translations longer than their references are already penalized by the modified n-gram precision measure: there is no need to penalize them again. Consequently, we introduce a multiplicative *brevity penalty* factor that only penalizes candidates shorter than their reference translations. With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order. Note that neither this brevity penalty nor the modified *n*-gram precision length effect directly considers the source length; instead, they consider the range of reference translation lengths in the target language.

The brevity penalty is a multiplicative factor modifying the overall BLEU score. We wish to make the penalty 1 when the candidate's length is the same as any reference translation's length. For example, if there are three references with lengths 12, 15, and 17 words and the candidate translation is a terse 12 words, we want the brevity penalty to be 1. We call the closest reference sentence length the "best match length."

We compute the brevity penalty over the entire corpus to allow some freedom at the sentence level. We first compute the test corpus' effective reference length, *r*, by summing the best match lengths for each candidate sentence in the corpus. The brevity penalty is a decaying exponential in *r/c*, where *c* is the total length of the candidate translation corpus.

We take the geometric mean of the test corpus' modified precision scores and then multiply the result by an exponential brevity penalty factor.

We first compute the geometric average of the modified n-gram precisions, p_n, using *n*-grams up to length *N* and positive weights w_n summing to one.

Next, let *c* be the length of the candidate translation and *r* be the effective reference corpus length. We compute the brevity penalty BP,

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}.$$

Then,

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

In the log domain,

$$\log \text{BLEU} = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

As *n* increases, BLEU places more emphasis on longer *n*-grams than unigrams (since *n*-gram precision decreases logarithmically).

2.2 NEE

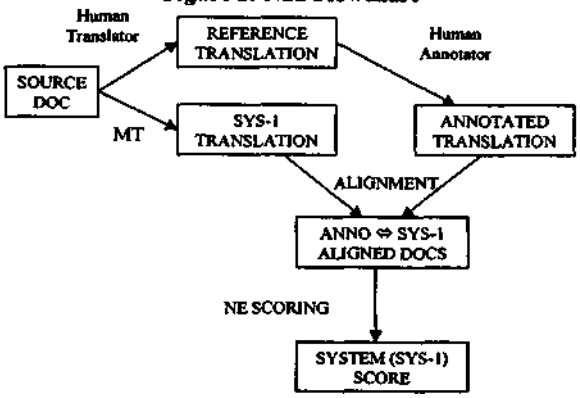
The Named Entity translation Evaluator NEE emerged from testing in named entity translation as an MT metric ([5], [9]). The score reflects the degree to which named entities [6] are translated correctly by a given MT system. The underlying assumptions are that a) named entities are important components in translation quality; b) named entities measurements are indicative of the kinds of downstream tasks for which MT output is suited; c) named entities metrics represent an actionable area for MT improvement . The premise that the NEE metric is a reasonable one for MTE comes from the fact that named entities are relatively objective measures with straightforward definitions and less variation than MT in the whole. Previously, named entities were identified as constrained, yet important nuggets in translation, particularly for TIDES applications [9]. The work presented here takes the prior work to the next level by applying the technique to larger corpora, more diverse language pairs and different levels of granularity.

The process for utilizing this metric is relatively straight-forward: a) identify the named entities within a given corpus; b) pull unique entities from the document[1]; c) find the entities in the system out-

[1] Experiments on French, Spanish showed that a strong correlation between paragraph and document level scores.

put text; and d) compare entities in the output text with those identified in the reference text. Figure 1 shows the flow. Identifying the named entities in the reference translation requires human annotation, and is the only stage of the process to do so. To prepare the corpora for evaluation, two expert annotators used the Alembic Workbench [2] annotation tool to tag occurrences of named entities according to the MUC [6] annotation guidelines. After the named entities are tagged in the reference translation (designated ANNO for our discussions), the metric can be applied.

Figure 1: NEE Flowchart



The next stage is to align the ANNO translation text with the evaluation text (SYS-1 for this discussion). This is performed, currently, at either the article (file) level or at the paragraph level within articles. The discussion of the most appropriate granularity follows in the results section. Once the ANNO and SYS-1 have been aligned, the aligned pairs are handed off to the scoring software.

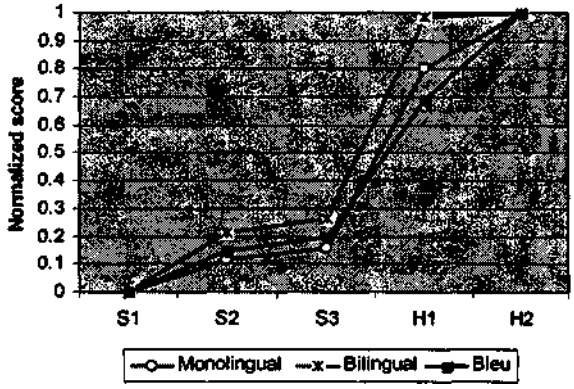
To score the translation, for each article in the aligned pair, the tagged named entities are pulled from the ANNO and a list of unique names for the comparison unit (paragraph or article) is prepared. This is followed by normalization, which becomes more important with increased divergence between language pairs (e.g. [10]). At this time, the normalization steps applied are: a) substitution of non-diacritic marked letters for the equivalent diacritic mark character for Romance languages[2]; b) down-casing[3]; c) the normalization of numeric quantities (particularly for numbers under 100) and d) the removal of possessives. Other normalization steps may be needed, as well as the incorporation of partial match scoring (see [9] for discussion of candidates). Once the named entity list and the SYS-1 tokens have been normalized, the search for named entities in the token lists is straight-forward. Only exact matches given the normalization steps described are considered at this time and all results here reflect this.

3. EXPERIMENTS

Different corpora were used for the development of the two metrics. The corpora were then exchanged for the purpose of running the different evaluators on them. Each corpus will now be described.

The Chinese-English corpus is a collection of documents from various sources including the internet. It consists of 40 documents and 500 sentences, 2 expert human reference translations, 3 system translations, and 2 additional human translations. A subset containing 50 sentences was randomly selected. Translations of these 50 sentence by the 3 systems and 2 humans (250 sentence pairs) were judged by 20 humans on a scale of 1 (very bad) to 5 (very good). Ten of the judges were monolingual native speakers of English and the rest were bilingual native speakers of Chinese. Figure 2 shows the correlation between human judgments and BLEU for this corpus. The correlation coefficient R of BLEU with monolingual human judgment is 0.99 while that with the bilingual judgments is 0.96. It may be noted that there is a high correlation between the monolingual and bilingual judgments. This suggests that we can use monolingual judgments to evaluate the metrics. We also show the *prediction error* (1 - R^2) expressed as percent in some of the graphs below.

Figure 2: BLEU vs Bilingual and Monolingual Judgments on Chinese to English



In addition to validating BLEU against human judgments on Chinese to-English translations, we compared BLEU with human judgments involving two other language pairs. We used the French-English human judgments from the DARPA-94 evaluation for 6 systems and the Spanish-English for 5 systems. BLEU was computed for each system against the one reference provided in the evaluation. One of the systems in each language pair was a human translator labeled "expert".

The French-English and Spanish-English corpora were both from the DARPA-94 evaluation. Each language set consists of 100 documents of newswire text, translated into English by 2 expert translators along with the machine translations from the original evaluation. BLEU was computed for the 100 test documents. Again, BLEU correlates highly with human judgments, particularly for adequacy and fluency. Figures 3 and 4 show the correlation of BLEU with human judgments on the French and Spanish translations respectively.

We summarize the correlation of BLEU with human judgments on DARPA-94 MT evaluation data below. It is remarkable that BLEU and fluency have a correlation of 0.9958 for French-English which we rounded to 1 in Table 1.

Table 1: Correlation of BLEU with Human Judgments of DARPA-94 MT Evaluation

Source	Adequacy	Fluency	Informativeness
French	0.94	1.00	0.86
Spanish	0.98	0.96	0.85

[2] For instance à becomes a.

[3] It has yet to be demonstrated that this is the correct normalization step, although it does increase human-human agreement.

Figure 3: BLEU Prediction Error on Adequacy, Fluency, and Informativeness (F-E)

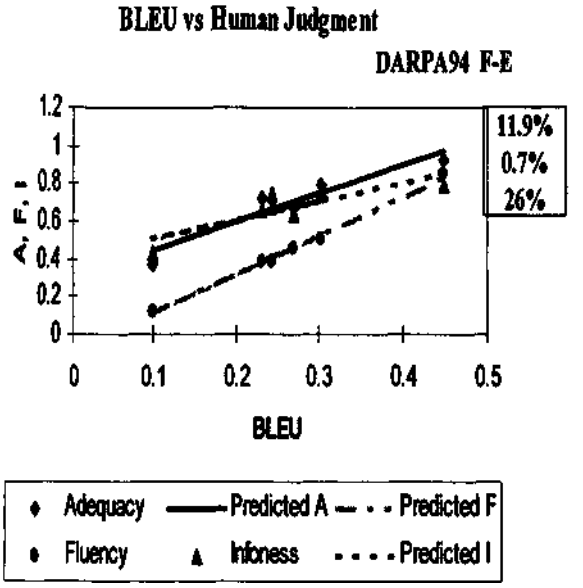
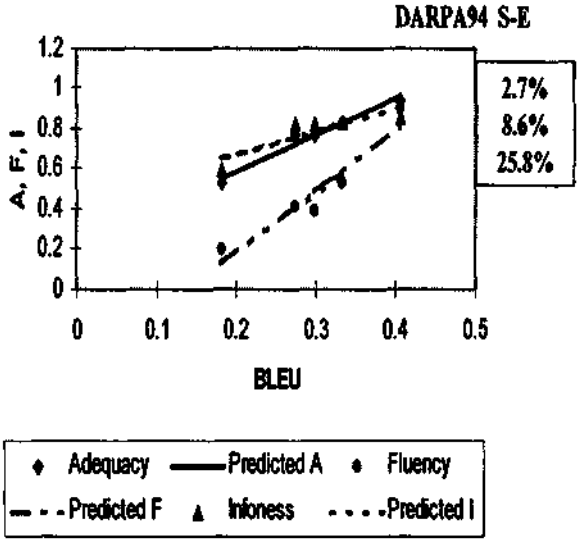


Figure 4: BLEU Prediction Error on A, F, I (S-E)

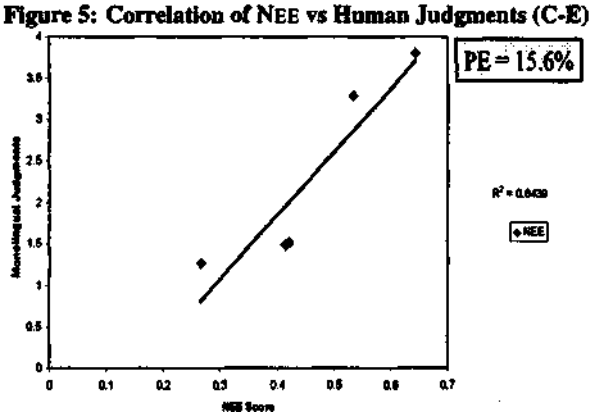


We also computed the correlation of NEE with human judgments of DARPA-94 MT evaluation. For the French-English NEE correlation, we only used 4 systems due to data-formatting issues. Figures 5, 6, 7 show the correlation of BLEU with human judgments on the Chinese, French, and Spanish translations respectively. Table 2 summarizes the results.

Table 2: Correlation of NEE with Human Judgments of DARPA-94 MT Evaluation

Source	Adequacy	Fluency	Informativeness
French	0.86	0.90	0.70
Spanish	0.76	0.93	0.49

Figure 5: Correlation of NEE vs Human Judgments (C-E)



Both BLEU and NEE are single number metrics, and cannot be expected to track different quantities such as Adequacy and Fluency (unless the quantities are highly correlated). Since BLEU is parametric in the size of *n*-grams that are matched, we can look at the effect of the *n*-gram size (phrase length) on prediction error of BLEU on Adequacy and Fluency. The effect is shown in Figures 8 and 9 for French and Spanish corpora respectively.

BLEU with shorter phrases correlates better with Adequacy whereas with longer phrases it correlates better with Fluency. In fact, $n = 1$ seems the best predictor of Adequacy. This is expected since adequacy is more about getting the words right, and BLEU with unigrams measures just that. This also suggests that as translation quality of the tested systems gets better, we should use BLEU with bigger n for better correlation. Conversely, when testing systems of poor quality, BLEU with smaller n may be better. Interestingly, on the Spanish-English test corpus we could go as far as matching 16-grams. Perhaps Spanish is an easier language to translate into English than French is.

We also tested these metrics on an Arabic-English corpus of 19 documents, containing about 5000 words. We evaluated 5 "systems" out of which two were human translators. We obtained human about 4000 sentence-level judgments from 16 monolingual native speakers of English. Again, we used only one reference translation. Figure 10 shows the correlation of BLEU with human judgments on this corpus.

The correlation coefficient of BLEU with monolingual human judgment is 0.98. This was a surprise to us: BLEU is a simple metric based on counting *n*-grams and with simple weights for different *n*-gram precisions. The same BLEU that was developed on the Chinese-English pilot corpus worked well on different test corpora spanning many source languages and systems of varying quality.

Figure 11 shows the correlation of NEE with human judgments on the Arabic-English corpus.

NEE showed for all languages that human translators are completely consistent even in named entity translation only about 80% of the time for Romance languages. For languages such as Arabic and Chinese, the scores drop to around 60% match on human translation agreement. Further study of this and of partial match scoring are indicated.

Acknowledgments This work was partially supported by the Defense Advanced Research Projects Agency and monitored by SPAWAR under contract No. N66001-99-2-8916. The views and

Figure 6: Correlation of NEE vs Human Judgments (F-E)

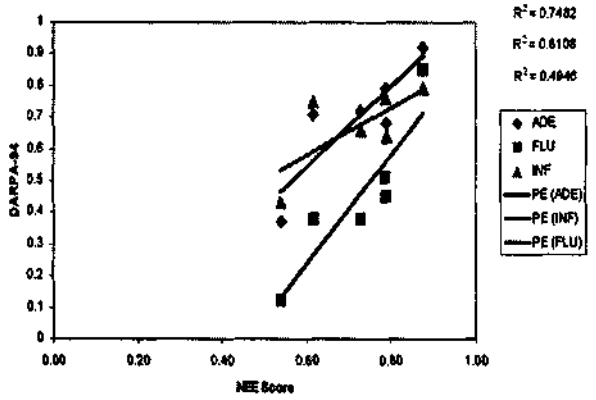
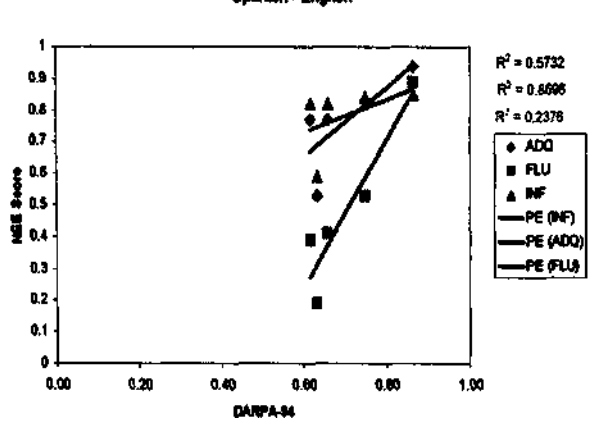


Figure 7: Correlation of NEE vs Human Judgments (S-E)



findings contained in this material are those of the authors and do not necessarily reflect the position of policy of the Government and no official endorsement should be inferred.

4. REFERENCES

[1] ALPAC. 1966. Language and machines: Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee. Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

[2] Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., & Vilain, M. 1997. Mixed-initiative Development of Language Processing Systems. Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLP.

[3] Doyon, J., Taylor, K., & White, J. 1998. The DARPA Machine Translation Evaluation Methodology: Past and Present. Proceedings of AMTA-98. Philadelphia, PA.

[4] JEIDA, 1989. A Japanese View of Machine Translation in the Light of the Considerations and Recommendations Reported by ALPAC, USA. Japan Electronic Industry Development Association, Tokyo.

[5] Hirschman, L., Reeder, F, Burger, J., & Miller, K. 2000. Name Translation as a Machine Translation Evaluation Task. Proceedings of the Workshop on Machine Translation Evaluation, LREC-2000.

[6] MUC-7.1998. Proceedings of the Seventh Message Understanding Conference (MUC-7). http://www.muc.saic.com/proceedings/muc_7_toc.html

[7] Papineni, K., Roukos, S., Ward, R. T, Zhu, W.-Z. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, PA.

[8] White, J. S., T. A. O'Connell, & F. E. O'Mara. 1994. Advanced Research Projects Agency Machine Translation Program: 3Q94. Proceedings of the November 1994 Meeting.

[9] Reeder, F., Miller, K., Doyon, J., White, J. 2001. The Naming of Things and the Confusion of Tongues: An MT Metric. MT-Summit Workshop on MT Evaluation, September.

[10] Knight, K. & J. Graehl. 1998. Machine Transliteration. Computational Linguistics, 24(4), 598-612.

Figure 8: Prediction Error of BLEU vs n-gram length (F-E)

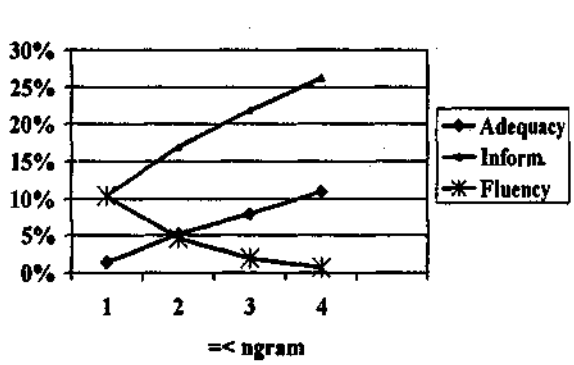


Figure 9: Prediction Error of BLEU vs n-gram length (S-E)

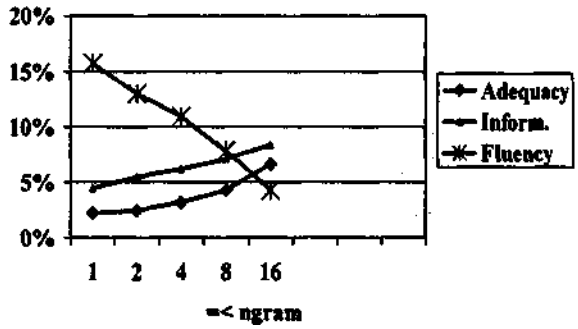


Figure 10: Correlation of BLEU vs Human Judgments (A-E)

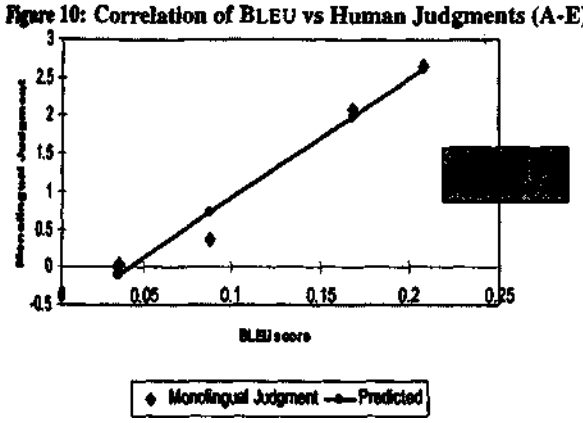
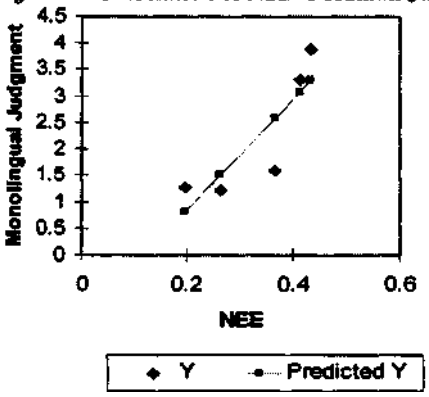


Figure 11: Correlation of NEE vs Human Judgments (A-E)



PE = 27%