Re-evaluating the Role of BLEU in Machine Translation Research

Chris Callison-Burch Miles Osborne Philipp Koehn
School on Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
callison-burch@ed.ac.uk

Abstract

We argue that the machine translation community is overly reliant on the Bleu machine translation evaluation metric. We show that an improved Bleu score is neither necessary nor sufficient for achieving an actual improvement in translation quality, and give two significant counterexamples to Bleu's correlation with human judgments of quality. This offers new potential for research which was previously deemed unpromising by an inability to improve upon Bleu scores.

1 Introduction

Over the past five years progress in machine translation, and to a lesser extent progress in natural language generation tasks such as summarization, has been driven by optimizing against n-gram-based evaluation metrics such as Bleu (Papineni et al., 2002). The statistical machine translation community relies on the Bleu metric for the purposes of evaluating incremental system changes and optimizing systems through minimum error rate training (Och, 2003). Conference papers routinely claim improvements in translation quality by reporting improved Bleu scores, while neglecting to show any actual example translations. Workshops commonly compare systems using Bleu scores, often without confirming these rankings through manual evaluation. All these uses of Bleu are predicated on the assumption that it correlates with human judgments of translation quality, which has been shown to hold in many cases (Doddington, 2002; Coughlin, 2003).

However, there is a question as to whether minimizing the error rate with respect to Bleu does indeed guarantee genuine translation improvements. If Bleu's correlation with human judgments has been overestimated, then the field needs to ask itself whether it should continue to be driven by Bleu to the extent that it currently is. In this paper we give a number of counterexamples for Bleu's correlation with human judgments. We show that under some circumstances an improvement in Bleu is *not sufficient* to reflect a genuine improvement in translation quality, and in other circumstances that it is *not necessary* to improve Bleu in order to achieve a noticeable improvement in translation quality.

We argue that Bleu is insufficient by showing that Bleu admits a huge amount of variation for identically scored hypotheses. Typically there are millions of variations on a hypothesis translation that receive the same Bleu score. Because not all these variations are equally grammatically or semantically plausible there are translations which have the *same* Bleu score but a *worse* human evaluation. We further illustrate that in practice a higher Bleu score is not necessarily indicative of better translation quality by giving two substantial examples of Bleu vastly underestimating the translation quality of systems. Finally, we discuss appropriate uses for Bleu and suggest that for some research projects it may be preferable to use a focused, manual evaluation instead.

2 BLEU Detailed

The rationale behind the development of Bleu (Papineni et al., 2002) is that human evaluation of machine translation can be time consuming and expensive. An automatic evaluation metric, on the other hand, can be used for frequent tasks like monitoring incremental system changes during development, which are seemingly infeasible in a manual evaluation setting.

The way that Bleu and other automatic evaluation metrics work is to compare the output of a machine translation system against reference human translations. Machine translation evaluation metrics differ from other metrics that use a reference, like the word error rate metric that is used

Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Table 1: A set of four reference translations, and a hypothesis translation from the 2005 NIST MT Evaluation

in speech recognition, because translations have a degree of variation in terms of word choice and in terms of variant ordering of some phrases.

Bleu attempts to capture allowable variation in word choice through the use of multiple reference translations (as proposed in Thompson (1991)). In order to overcome the problem of variation in phrase order, Bleu uses *modified n-gram precision* instead of WER's more strict string edit distance.

Bleu's n-gram precision is modified to eliminate repetitions that occur across sentences. For example, even though the bigram "to Miami" is repeated across all four reference translations in Table 1, it is counted only once in a hypothesis translation. Table 2 shows the n-gram sets created from the reference translations.

Papineni et al. (2002) calculate their modified precision score, p_n, for each n-gram length by summing over the matches for every hypothesis sentence S in the complete corpus C as:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Counting punctuation marks as separate tokens, the hypothesis translation given in Table 1 has 15 unigram matches, 10 bigram matches, 5 trigram matches (these are shown in bold in Table 2), and three 4-gram matches (not shown). The hypothesis translation contains a total of 18 unigrams, 17 bigrams, 16 trigrams, and 15 4-grams. If the complete corpus consisted of this single sentence

1-grams: **American**, **Florida**, **Miami**, Orejuela, **appeared**, as, being, **calm**, carry, escorted, **he**, him, in, led, **plane**, quite, seemed, take, that, **the**, **to**, **to**, to, **was** , was, **which**, while, **will**, would, **,, .**
2-grams: **American plane**, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, **, Florida**
3-grams: American plane that, American plane which, **Miami , Florida**, Miami in Florida, Orejuela appeared calm, Orejuela seemed quite, appeared calm as, appeared calm while, as he was, being escorted to, being led to, calm as he, calm while being, carry him to, escorted to the, he was being, he was led, him to Miami, in Florida ., led to the, plane that was, plane that would, plane which will, quite calm as, seemed quite calm, take him to, that was to, that would take, **the American plane**, the plane that, **to Miami ,**, to Miami in, to carry him, **to the American**, to the plane, was being led, was led to, was to carry, which will take, while being escorted, will take him, would take him, **, Florida .**

Table 2: The n-grams extracted from the reference translations, with matches from the hypothesis translation in bold

then the modified precisions would be $p_1 = .83$, $p_2 = .59$, $p_3 = .31$, and $p_4 = .2$. Each p_n is combined and can be weighted by specifying a weight w_n. In practice each p_n is generally assigned an equal weight.

Because Bleu is precision based, and because recall is difficult to formulate over multiple reference translations, a *brevity penalty* is introduced to compensate for the possibility of proposing high-precision hypothesis translations which are too short. The brevity penalty is calculated as:

$$\text{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{1-r/c} & \text{if} \quad c \leq r \end{cases}$$

where c is the length of the corpus of hypothesis translations, and r is the effective reference corpus length.[1]

Thus, the Bleu score is calculated as

$$\text{Bleu} = \text{BP} \ * \exp(\sum_{n=1}^{N} w_n \log p_n)$$

A Bleu score can range from 0 to 1, where higher scores indicate closer matches to the reference translations, and where a score of 1 is assigned to a hypothesis translation which exactly

[1]The effective reference corpus length is calculated as the sum of the single reference translation from each set which is closest to the hypothesis translation.

matches one of the reference translations. A score of 1 is also assigned to a hypothesis translation which has matches for all its n-grams (up to the maximum n measured by Bleu) in the clipped reference n-grams, and which has no brevity penalty.

The primary reason that Bleu is viewed as a useful stand-in for manual evaluation is that it has been shown to correlate with human judgments of translation quality. Papineni et al. (2002) showed that Bleu correlated with human judgments in its rankings of five Chinese-to-English machine translation systems, and in its ability to distinguish between human and machine translations. Bleu's correlation with human judgments has been further tested in the annual NIST Machine Translation Evaluation exercise wherein Bleu's rankings of Arabic-to-English and Chinese-to-English systems is verified by manual evaluation.

In the next section we discuss theoretical reasons why Bleu may not always correlate with human judgments.

3 Variations Allowed By BLEU

While Bleu attempts to capture allowable variation in translation, it goes much further than it should. In order to allow some amount of variant order in phrases, Bleu places no explicit constraints on the order that matching n-grams occur in. To allow variation in word choice in translation Bleu uses multiple reference translations, but puts very few constraints on how n-gram matches can be drawn from the multiple reference translations. Because Bleu is underconstrained in these ways, it allows a tremendous amount of variation – far beyond what could reasonably be considered acceptable variation in translation.

In this section we examine various *permutations* and *substitutions* allowed by Bleu. We show that for an average hypothesis translation there are millions of possible variants that would each receive a similar Bleu score. We argue that because the number of translations that score the same is so large, it is unlikely that all of them will be judged to be identical in quality by human annotators. This means that it is possible to have items which receive identical Bleu scores but are judged by humans to be worse. It is also therefore possible to have a *higher* Bleu score *without* any genuine improvement in translation quality. In Sections 3.1 and 3.2 we examine ways of synthetically producing such variant translations.

3.1 Permuting phrases

One way in which variation can be introduced is by permuting phrases within a hypothesis translation. A simple way of estimating a lower bound on the number of ways that phrases in a hypothesis translation can be reordered is to examine bigram mismatches. Phrases that are bracketed by these bigram mismatch sites can be freely permuted because reordering a hypothesis translation at these points *will not reduce the number of matching n-grams* and thus will not reduce the overall Bleu score.

Here we denote bigram mismatches for the hypothesis translation given in Table 1 with vertical bars:

> Appeared calm | when | he was | taken | to the American plane | , | which will | to Miami , Florida .

We can randomly produce other hypothesis translations that have the same Bleu score but are radically different from each other. Because Bleu only takes order into account through rewarding matches of higher order n-grams, a hypothesis sentence may be freely permuted around these bigram mismatch sites and without reducing the Bleu score. Thus:

> which will | he was | , | when | taken | Appeared calm | to the American plane | to Miami , Florida .

receives an identical score to the hypothesis translation in Table 1.

If b is the number of bigram matches in a hypothesis translation, and k is its length, then there are

$$(k - b)! \quad (1)$$

possible ways to generate similarly scored items using only the words in the hypothesis translation.[2] Thus for the example hypothesis translation there are at least **40,320** different ways of permuting the sentence and receiving a similar Bleu score. The number of permutations varies with respect to sentence length and number of bigram mismatches. Therefore as a hypothesis translation approaches being an identical match to one of the reference translations, the amount of variance decreases significantly. So, as translations improve

[2]Note that in some cases randomly permuting the sentence in this way may actually result in a greater number of n-gram matches; however, one would not expect random permutation to increase the human evaluation.

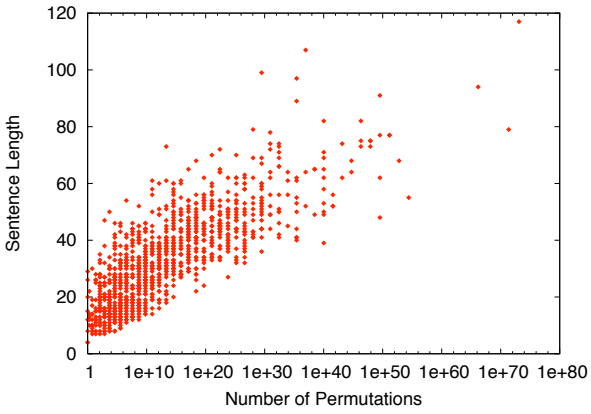

Figure 1: Scatterplot of the length of each translation against its number of possible permutations due to bigram mismatches for an entry in the 2005 NIST MT Eval

spurious variation goes down. However, at today's levels the amount of variation that Bleu admits is unacceptably high. Figure 1 gives a scatterplot of each of the hypothesis translations produced by the second best Bleu system from the 2005 NIST MT Evaluation. The number of possible permutations for some translations is greater than $\mathbf{10^{73}}$.

3.2 Drawing different items from the reference set

In addition to the factorial number of ways that similarly scored Bleu items can be generated by permuting phrases around bigram mismatch points, additional variation may be synthesized by drawing different items from the reference n-grams. For example, since the hypothesis translation from Table 1 has a length of 18 with 15 unigram matches, 10 bigram matches, 5 trigram matches, and three 4-gram matches, we can artificially construct an identically scored hypothesis by drawing an identical number of matching n-grams from the reference translations. Therefore the far less plausible:

> was being led to the | calm as he was | would take | carry him | seemed quite | when | taken

would receive the same Bleu score as the hypothesis translation from Table 1, even though human judges would assign it a much lower score.

This problem is made worse by the fact that Bleu equally weights all items in the reference sentences (Babych and Hartley, 2004). Therefore omitting content-bearing lexical items does not carry a greater penalty than omitting function words.

The problem is further exacerbated by Bleu not having any facilities for matching synonyms or lexical variants. Therefore words in the hypothesis that did not appear in the references (such as *when* and *taken* in the hypothesis from Table 1) can be substituted with arbitrary words because they do not contribute towards the Bleu score. Under Bleu, we could just as validly use the words *black* and *helicopters* as we could *when* and *taken*.

The lack of recall combined with naive token identity means that there can be overlap between similar items in the multiple reference translations. For example we can produce a translation which contains both the words *carry* and *take* even though they arise from the same source word. The chance of problems of this sort being introduced increases as we add more reference translations.

3.3 Implication: BLEU cannot guarantee correlation with human judgments

Bleu's inability to distinguish between randomly generated variations in translation hints that it may not correlate with human judgments of translation quality in some cases. As the number of identically scored variants goes up, the likelihood that they would all be judged equally plausible goes down. This is a theoretical point, and while the variants are artificially constructed, it does highlight the fact that Bleu is quite a crude measurement of translation quality.

A number of prominent factors contribute to Bleu's crudeness:

- Synonyms and paraphrases are only handled if they are in the set of multiple reference translations.
- The scores for words are equally weighted so missing out on content-bearing material brings no additional penalty.
- The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate recall.

Each of these failures contributes to an increased amount of inappropriately indistinguishable translations in the analysis presented above.

Given that Bleu can theoretically assign equal scoring to translations of obvious different quality, it is logical that a higher Bleu score may not

Fluency
How do you judge the fluency of this translation?
5 = Flawless English
4 = Good English
3 = Non-native English
2 = Disfluent English
1 = Incomprehensible
Adequacy
How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?
5 = All
4 = Most
3 = Much
2 = Little
1 = None

Table 3: The scales for manually assigned adequacy and fluency scores

Iran has already stated that Kharazi's statements to the conference because of the Jordanian King Abdullah II in which he stood accused Iran of interfering in Iraqi affairs.
n-gram matches: 27 unigrams, 20 bigrams, 15 trigrams, and ten 4-grams **human scores:** Adequacy:3,2 Fluency:3,2
Iran already announced that Kharrazi will not attend the conference because of the statements made by the Jordanian Monarch Abdullah II who has accused Iran of interfering in Iraqi affairs.
n-gram matches: 24 unigrams, 19 bigrams, 15 trigrams, and 12 4-grams **human scores:** Adequacy:5,4 Fluency:5,4
Reference: Iran had already announced Kharazi would boycott the conference after Jordan's King Abdullah II accused Iran of meddling in Iraq's affairs.

Table 4: Two hypothesis translations with similar Bleu scores but different human scores, and one of four reference translations

necessarily be indicative of a genuine improvement in translation quality. This begs the question as to whether this is only a theoretical concern or whether Bleu's inadequacies can come into play in practice. In the next section we give two significant examples that show that Bleu can indeed fail to correlate with human judgments in practice.

4 Failures in Practice: the 2005 NIST MT Eval, and Systran v. SMT

The NIST Machine Translation Evaluation exercise has run annually for the past five years as part of DARPA's TIDES program. The quality of Chinese-to-English and Arabic-to-English translation systems is evaluated both by using Bleu score and by conducting a manual evaluation. As such, the NIST MT Eval provides an excellent source of data that allows Bleu's correlation with human judgments to be verified. Last year's evaluation exercise (Lee and Przybocki, 2005) was startling in that Bleu's rankings of the Arabic-English translation systems failed to fully correspond to the manual evaluation. In particular, the entry that was ranked *1st* in the human evaluation was ranked *6th* by Bleu. In this section we examine Bleu's failure to correctly rank this entry.

The manual evaluation conducted for the NIST MT Eval is done by English speakers without reference to the original Arabic or Chinese documents. Two judges assigned each sentence in the hypothesis translations a subjective 1–5 score along two axes: adequacy and fluency (LDC, 2005). Table 3 gives the interpretations of the scores. When first evaluating fluency, the judges are shown only the hypothesis translation. They are then shown a reference translation and are asked to judge the adequacy of the hypothesis sentences.

Table 4 gives a comparison between the output of the system that was ranked 2nd by Bleu[3] (top) and of the entry that was ranked 6th in Bleu but 1st in the human evaluation (bottom). The example is interesting because the number of matching n-grams for the two hypothesis translations is roughly similar but the human scores are quite different. The first hypothesis is less adequate because it fails to indicated that Kharazi is boycotting the conference, and because it inserts the word *stood* before *accused* which makes the Abdullah's actions less clear. The second hypothesis contains all of the information of the reference, but uses some synonyms and paraphrases which would not picked up on by Bleu: *will not attend* for *would boycott* and *interfering* for *meddling*.

[3]The output of the system that was ranked 1st by Bleu is not publicly available.

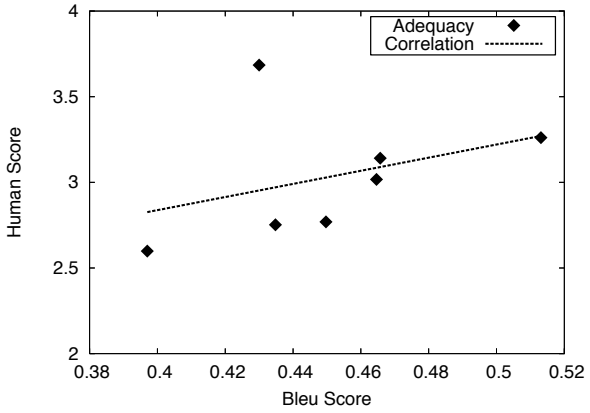

Figure 2: Bleu scores plotted against human judgments of adequacy, with $R^2 = 0.14$ when the outlier entry is included

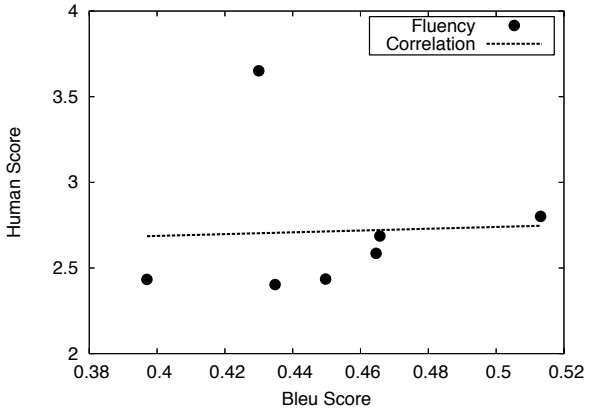

Figure 3: Bleu scores plotted against human judgments of fluency, with $R^2 = 0.002$ when the outlier entry is included

Figures 2 and 3 plot the average human score for each of the seven NIST entries against its Bleu score. It is notable that one entry received a much higher human score than would be anticipated from its low Bleu score. The offending entry was unusual in that it was not fully automatic machine translation; instead the entry was aided by monolingual English speakers selecting among alternative automatic translations of phrases in the Arabic source sentences and post-editing the result (Callison-Burch, 2005). The remaining six entries were all fully automatic machine translation systems; in fact, they were all phrase-based statistical machine translation system that had been trained on the same parallel corpus and most used Bleu-based minimum error rate training (Och, 2003) to optimize the weights of their log linear models' feature functions (Och and Ney, 2002).

This opens the possibility that in order for Bleu to be valid only sufficiently similar systems should be compared with one another. For instance, when measuring correlation using Pearson's we get a very low correlation of $R^2 = 0.14$ when the outlier in Figure 2 is included, but a strong $R^2 = 0.87$ when it is excluded. Similarly Figure 3 goes from $R^2 = 0.002$ to a much stronger $R^2 = 0.742$. Systems which explore different areas of translation space may produce output which has differing characteristics, and might end up in different regions of the human scores / Bleu score graph.

We investigated this by performing a manual evaluation comparing the output of two statistical machine translation systems with a rule-based machine translation, and seeing whether Bleu correctly ranked the systems. We used Systran for the rule-based system, and used the French-English portion of the Europarl corpus (Koehn, 2005) to train the SMT systems and to evaluate all three systems. We built the first phrase-based SMT system with the complete set of Europarl data (14-15 million words per language), and optimized its feature functions using minimum error rate training in the standard way (Koehn, 2004). We evaluated it and the Systran system with Bleu using a set of 2,000 held out sentence pairs, using the same normalization and tokenization schemes on both systems' output. We then built a number of SMT systems with various portions of the training corpus, and selected one that was trained with $\frac{1}{64}$ of the data, which had a Bleu score that was close to, but still higher than that for the rule-based system.

We then performed a manual evaluation where we had three judges assign fluency and adequacy ratings for the English translations of 300 French sentences for each of the three systems. These scores are plotted against the systems' Bleu scores in Figure 4. The graph shows that the Bleu score for the rule-based system (Systran) vastly underestimates its actual quality. This serves as another significant counter-example to Bleu's correlation with human judgments of translation quality, and further increases the concern that Bleu may not be appropriate for comparing systems which employ different translation strategies.

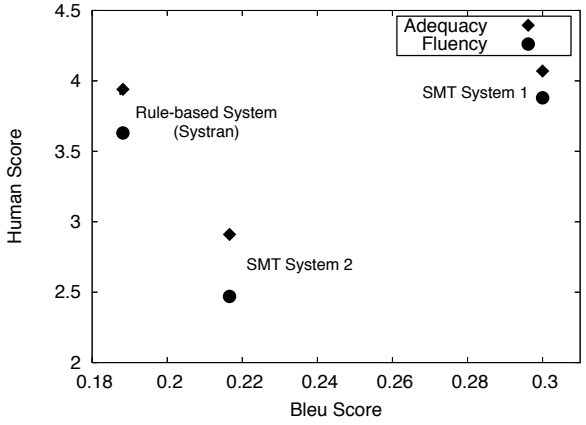

Figure 4: Bleu scores plotted against human judgments of fluency and adequacy, showing that Bleu vastly underestimates the quality of a non-statistical system

5 Related Work

A number of projects in the past have looked into ways of extending and improving the Bleu metric. Doddington (2002) suggested changing Bleu's weighted geometric average of n-gram matches to an arithmetic average, and calculating the brevity penalty in a slightly different manner. Hovy and Ravichandra (2003) suggested increasing Bleu's sensitivity to inappropriate phrase movement by matching part-of-speech tag sequences against reference translations in addition to Bleu's n-gram matches. Babych and Hartley (2004) extend Bleu by adding frequency weighting to lexical items through TF/IDF as a way of placing greater emphasis on content-bearing words and phrases.

Two alternative automatic translation evaluation metrics do a much better job at incorporating recall than Bleu does. Melamed et al. (2003) formulate a metric which measures translation accuracy in terms of precision and recall directly rather than precision and a brevity penalty. Banerjee and Lavie (2005) introduce the Meteor metric, which also incorporates recall on the unigram level and further provides facilities incorporating stemming, and WordNet synonyms as a more flexible match.

Lin and Hovy (2003) as well as Soricut and Brill (2004) present ways of extending the notion of n-gram co-occurrence statistics over multiple references, such as those used in Bleu, to other natural language generation tasks such as summarization. Both these approaches potentially suffer from the same weaknesses that Bleu has in machine translation evaluation.

Coughlin (2003) performs a large-scale investigation of Bleu's correlation with human judgments, and finds one example that fails to correlate. Her future work section suggests that she has preliminary evidence that statistical machine translation systems receive a higher Bleu score than their non-n-gram-based counterparts.

6 Conclusions

In this paper we have shown theoretical and practical evidence that Bleu may not correlate with human judgment to the degree that it is currently believed to do. We have shown that Bleu's rather coarse model of allowable variation in translation can mean that an improved Bleu score is not sufficient to reflect a genuine improvement in translation quality. We have further shown that it is not necessary to receive a higher Bleu score in order to be judged to have better translation quality by human subjects, as illustrated in the 2005 NIST Machine Translation Evaluation and our experiment manually evaluating Systran and SMT translations.

What conclusions can we draw from this? Should we give up on using Bleu entirely? We think that the advantages of Bleu are still very strong; automatic evaluation metrics *are* inexpensive, and *do* allow many tasks to be performed that would otherwise be impossible. The important thing therefore is to recognize which uses of Bleu are appropriate and which uses are not.

Appropriate uses for Bleu include tracking broad, incremental changes to a single system, comparing systems which employ similar translation strategies (such as comparing phrase-based statistical machine translation systems with other phrase-based statistical machine translation systems), and using Bleu as an objective function to optimize the values of parameters such as feature weights in log linear translation models, until a better metric has been proposed.

Inappropriate uses for Bleu include comparing systems which employ radically different strategies (especially comparing phrase-based statistical machine translation systems against systems that do not employ similar n-gram-based approaches), trying to detect improvements for aspects of translation that are not modeled well by Bleu, and monitoring improvements that occur infrequently within a test corpus.

These comments do not apply solely to Bleu.

Meteor (Banerjee and Lavie, 2005), Precision and Recall (Melamed et al., 2003), and other such automatic metrics may also be affected to a greater or lesser degree because they are all quite rough measures of translation similarity, and have inexact models of allowable variation in translation.

Finally, that the fact that Bleu's correlation with human judgments has been drawn into question may warrant a re-examination of past work which failed to show improvements in Bleu. For example, work which failed to detect improvements in translation quality with the integration of word sense disambiguation (Carpuat and Wu, 2005), or work which attempted to integrate syntactic information but which failed to improve Bleu (Charniak et al., 2003; Och et al., 2004) may deserve a second look with a more targeted manual evaluation.

Acknowledgments

The authors are grateful to Amittai Axelrod, Frank Keller, Beata Kouchnir, Jean Senellart, and Matthew Stone for their feedback on drafts of this paper, and to Systran for providing translations of the Europarl test set.

References

Bogdan Babych and Anthony Hartley. 2004. Extending the Bleu MT evaluation method with frequency weightings. In *Proceedings of ACL*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, Michigan.

Chris Callison-Burch. 2005. Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of ACL*.

Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for machine translation. In *Proceedings of MT Summit IX*.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, San Diego.

Eduard Hovy and Deepak Ravichandra. 2003. Holy and unholy grails. Panel Discussion at MT Summit IX.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*.

Philipp Koehn. 2005. A parallel corpus for statistical machine translation. In *Proceedings of MT-Summit*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Audrey Lee and Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. Official release of automatic evaluation scores for all submissions, August.

Chin-Yew Lin and Ed Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*.

Dan Melamed, Ryan Green, and Jospeh P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of HLT/NAACL*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of NAACL-04*, Boston.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Radu Soricut and Eric Brill. 2004. A unified framework for automatic evaluation using n-gram co-occurrence statistics. In *Proceedings of ACL*.

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