Re-evaluating Machine Translation Results with Paraphrase Support

Liang Zhou, Chin-Yew Lin, and Eduard Hovy
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{liangz, cyl, hovy} @isi.edu

Abstract

In this paper, we present ParaEval, an automatic evaluation framework that uses paraphrases to improve the quality of machine translation evaluations. Previous work has focused on fixed *n-gram* evaluation metrics coupled with lexical identity matching. ParaEval addresses three important issues: support for paraphrase/synonym matching, recall measurement, and correlation with human judgments. We show that ParaEval correlates significantly better than BLEU with human assessment in measurements for both *fluency* and *adequacy*.

1 Introduction

The introduction of automated evaluation procedures, such as BLEU (Papineni et al., 2001) for machine translation (MT) and ROUGE (Lin and Hovy, 2003) for summarization, have prompted much progress and development in both of these areas of research in Natural Language Processing (NLP). Both evaluation tasks employ a comparison strategy for comparing textual units from machine-generated and gold-standard texts. Ideally, this comparison process would be performed manually, because of humans' abilities to infer, paraphrase, and use world knowledge to relate differently worded pieces of equivalent information. However, manual evaluations are time consuming and expensive, thus making them a bottleneck in system development cycles.

BLEU measures how close machine-generated translations are to professional human translations, and ROUGE does the same with respect to summaries. Both methods incorporate the comparison of a system-produced text to one or more corresponding reference texts. The closeness between texts is measured by the computation of a numeric score based on n-gram co-occurrence statistics. Although both methods have gained mainstream acceptance and have shown good correlations with human judgments, their deficiencies have become more evident and serious as research in MT and summarization progresses (Callison-Burch et al., 2006).

Text comparisons in MT and summarization evaluations are performed at different text granularity levels. Since most of the phrase-based, syntax-based, and rule-based MT systems translate one sentence at a time, the text comparison in the evaluation process is also performed at the single-sentence level. In summarization evaluations, there is no sentence-to-sentence correspondence between summary pairs—essentially a multi-sentence-to-multi-sentence comparison, making it more difficult and requiring a completely different implementation for matching strategies. In this paper, we focus on the intricacies involved in evaluating MT results and address two prominent problems associated with the BLEU-esque metrics, namely their lack of support for paraphrase matching and the absence of recall scoring. Our solution, ParaEval, utilizes a large collection of paraphrases acquired through an unsupervised process—identifying phrase sets that have the same translation in another language—using state-of-the-art statistical MT word alignment and phrase extraction methods. This collection facilitates paraphrase matching, additionally coupled with lexical identity matching which is designed for comparing text/sentence fragments that are not consumed by paraphrase matching. We adopt a unigram counting strategy for contents matched between sentences from peer and reference translations. This unweighted scoring scheme, for both precision and recall computations, allows us to directly examine both the power and limitations of

ParaEval. We show that ParaEval is a more stable and reliable comparison mechanism than BLEU, in both fluency and adequacy rankings.

This paper is organized in the following way: Section 2 shows an overview on BLEU and lexical identity n-gram statistics; we describe ParaEval's implementation in detail in Section 3; the evaluation of ParaEval is shown in Section 4; recall computation is discussed in Section 5; in Section 6, we discuss the differences between BLEU and ParaEval when the numbers of reference translations change; and we conclude and discuss future work in Section 7.

2 N-gram Co-occurrence Statistics

Being an $8 billion industry (Browner, 2006), MT calls for rapid development and the ability to differentiate good systems from less adequate ones. The evaluation process consists of comparing system-generated *peer translations* to human written *reference translations* and assigning a numeric score to each system. While human assessments are still the most reliable evaluation measurements, it is not practical to solicit manual evaluations repeatedly while making incremental system design changes that would only result in marginal performance gains. To overcome the monetary and time constraints associated with manual evaluations, automated procedures have been successful in delivering benchmarks for performance hill-climbing with little or no cost.

While a variety of automatic evaluation methods have been introduced, the underlining comparison strategy is similar—matching based on lexical identity. The most prominent implementation of this type of matching is demonstrated in BLEU (Papineni et al, 2002). The remaining part of this section is devoted to an overview of BLEU, or the BLEU-esque philosophy.

2.1 The BLEU-esque Matching Philosophy

The primary task that a BLEU-esque procedure performs is to compare n-grams from the peer translation with the n-grams from one or more reference translations and count the number of matches. The more matches a peer translation gets, the better it is.

BLEU is a precision-based metric, which is the ratio of the number of n-grams from the peer translation that occurred in reference translations to the total number of n-grams in the peer translation. The notion of *Modified n-gram Precision* was introduced to detect and avoid rewarding false positives generated by translation systems. To gain high precision, systems could potentially over-generate "good" n-grams, which occur multiple times in multiple references. The solution to this problem was to adopt the policy that an n-gram, from both reference and peer translations, is considered exhausted after participating in a match. As a result, the maximum number of matches an n-gram from a peer translation can receive, when comparing to a set of reference translations, is the maximum number of times this n-gram occurred in any single reference translation. Papineni et al. (2002) called this capping technique *clipping*. Figure 1, taken from the original BLEU paper, demonstrates the computation of the modified unigram precision for a peer translation sentence.

Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.
Modified Unigram Precision = 2/7.[3]

Figure 1. Modified *n*-gram precision from BLEU.

To compute the modified n-gram precision, P_n, for a whole test set, including all translation segments (usually in sentences), the formula is:

$$P_n = \frac{\sum_{C \in \{peers\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{peers\}} \sum_{n-gram \in C} Count(n-gram)}$$

2.2 Lack of Paraphrasing Support

Humans are very good at finding creative ways to convey the same information. There is no one definitive reference translation in one language for a text written in another. Having acknowledged this phenomenon, however natural it is, human evaluations on system-generated translations are much more preferred and trusted. However, what humans can do with ease puts machines at a loss. BLEU-esque procedures recognize equivalence only when two n-grams exhibit the same surface-level representations, i.e. the same lexical identities. The BLEU implementation addresses its deficiency in measuring semantic closeness by incorporating the comparison with multiple reference translations. The rationale is that multiple references give a higher chance that the n-grams, assuming correct translations, appearing in the peer translation would be rewarded by one of the reference's n-grams. The more reference translations used, the better

the matching and overall evaluation quality. Ideally (and to an extreme), we would need to collect a large set of human-written translations to capture all possible combinations of verbalizing variations before the translation comparison procedure reaches its optimal matching ability.

One can argue that an infinite number of references are not needed in practice because any matching procedure would stabilize at a certain number of references. This is true if precision measure is the only metric computed. However, using precision scores alone unfairly rewards systems that "under-generate"—producing unreasonably short translations. Recall measurements would provide more balanced evaluations. When using multiple reference translations, if an n-gram match is made for the peer, this n-gram could appear in any of the references. The computation of recall becomes difficult, if not impossible. This problem can be reversed if there is crosschecking for phrases occurring across references—paraphrase recognition. BLEU uses the calculation of a brevity penalty to compensate the lack of recall computation problem. The brevity penalty is computed as follows:

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

Then, the BLEU score for a peer translation is computed as:

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

BLEU's adoption of the brevity penalty to offset the effect of not having a recall computation has drawn criticism on its crudeness in measuring translation quality. Callison-Burch et al. (2006) point out three prominent factors:

- ``Synonyms and paraphrases are only handled if they are in the set of multiple reference translations [available].
- 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."

With the introduction of ParaEval, we will address two of these three issues, namely the paraphrasing problem and providing a recall measure.

3 ParaEval for MT Evaluation

3.1 Overview

Reference translations are created from the same source text (written in the foreign language) to the target language. Ideally, they are supposed to be semantically equivalent, i.e. overlap completely. However, as shown in Figure 2, when matching based on lexical identity is used (indicated by links), only half (6 from the left and 5 from the right) of the 12 words from these two sentences are matched. Also, "to" was a mismatch. In applying paraphrase matching for MT evaluation from ParaEval, we aim to match all shaded words from both sentences.

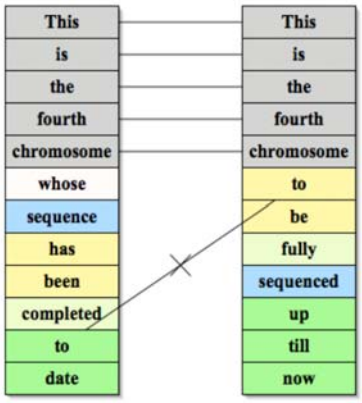

Figure 2. Two reference translations. Grey areas are matched by using BLEU.

3.2 Paraphrase Acquisition

The process of acquiring a large enough collection of paraphrases is not an easy task. Manual corpus analyses produce domain-specific collections that are used for text generation and are application-specific. But operating in multiple domains and for multiple tasks translates into multiple manual collection efforts, which could be very time-consuming and costly. In order to facilitate smooth paraphrase utilization across a variety of NLP applications, we need an unsupervised paraphrase collection mechanism that can be easily conducted, and produces paraphrases that are of adequate quality and can be readily used with minimal amount of adaptation effort.

Our method (Anonymous, 2006), also illustrated in (Bannard and Callison-Burch, 2005), to automatically construct a large domain-independent paraphrase collection is based on the assumption that two different phrases of the same meaning may have the same translation in a

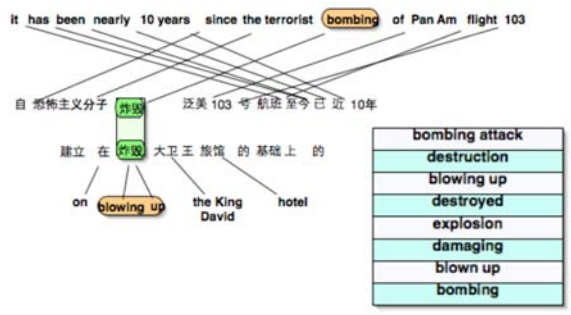

Figure 3. An example of the paraphrase extraction process.

foreign language. Phrase-based Statistical Machine Translation (SMT) systems analyze large quantities of bilingual parallel texts in order to learn translational alignments between pairs of words and phrases in two languages (Och and Ney, 2004). The sentence-based translation model makes word/phrase alignment decisions probabilistically by computing the optimal model parameters with application of the statistical estimation theory. This alignment process results in a corpus of word/phrase-aligned parallel sentences from which we can extract phrase pairs that are translations of each other. We ran the alignment algorithm from (Och and Ney, 2003) on a Chinese-English parallel corpus of 218 million English words, available from the Linguistic Data Consortium (LDC). Phrase pairs are extracted by following the method described in (Och and Ney, 2004) where all contiguous phrase pairs having consistent alignments are extraction candidates. Using these pairs we build paraphrase sets by joining together all English phrases that have the same Chinese translation. Figure 3 shows an example word/phrase alignment for two parallel sentence pairs from our corpus where the phrases "blowing up" and "bombing" have the same Chinese translation. On the right side of the figure we show the paraphrase set which contains these two phrases, which is typical in our collection of extracted paraphrases.

Although our paraphrase extraction method is similar to that of (Bannard and Callison-Burch, 2005), the paraphrases we extracted are for completely different applications, and have a broader definition for what constitutes a paraphrase. In (Bannard and Callison-Burch, 2005), a language model is used to make sure that the paraphrases extracted are direct substitutes, from the same syntactic categories, etc. So, using the example in Figure 3, the paraphrase table would contain only "bombing" and "bombing attack". Paraphrases that are direct substitutes of one another are useful when translating unknown phrases. For instance, if a MT system does not have the Chinese translation for the word "bombing", but has seen it in another set of parallel data (not involving Chinese) and has determined it to be a direct substitute of the phrase "bombing attack", then the Chinese translation of "bombing attack" would be used in place of the translation for "bombing". This substitution technique has shown some improvement in translation quality (Callison-Burch et al., 2006).

3.3 The ParaEval Evaluation Procedure

We adopt a two-tier matching strategy for MT evaluation in ParaEval. At the top tier, a paraphrase match is performed on system-translated sentences and corresponding reference sentences. Then, unigram matching is performed on the words not matched by paraphrases. Precision is measured as the ratio of the total number of words matched to the total number of words in the peer translation.

Running our system on the example in Figure 2, the paraphrase-matching phase consumes the words marked in grey and aligns "have been" and "to be", "completed" and "fully", "to date" and "up till now", and "sequence" and "sequenced". The subsequent unigram-matching aligns words based on lexical identity.

We maintain the computation of *modified unigram precision*, defined by the BLEU-esque Philosophy, in principle. In addition to clipping individual candidate *words* with their corresponding maximum reference counts (only for words not matched by paraphrases), we clip candidate *paraphrases* by their maximum reference paraphrase counts. So two completely different phrases in a reference sentence can be counted as two occurrences of one phrase. For example in Figure 4, candidate phrases "blown up" and "bombing" matched with three phrases from the references, namely "bombing" and two instances of "explosion". Treating these two candidate phrases as one (paraphrase match), we can see its clip is 2 (from Ref 1, where "bombing" and "explosion" are counted as two occurrences of a single phrase). The only word that was matched by its lexical identity is "was". The modified unigram precision calculated by our method is 4/5, where as BLEU gives 2/5.

Candidate: [blown up] [bombing] was happening

Ref 1: the [bombing] resulted in an [explosion]

Ref 2: there was an [explosion]

Modified Unigram Precision

$$= \frac{\text{paraphrase match + lexical match}}{\text{total number of words in peer}}$$

$$= \frac{(2+1)+1}{5} = \frac{4}{5}$$

Figure 4. ParaEval's matching process.

4 Evaluating ParaEval

To be effective in MT evaluations, an automated procedure should be capable of distinguishing good translation systems from bad ones, human translations from systems', and human translations of differing quality. For a particular evaluation exercise, an evaluation system produces a ranking for system and human translations, and compares this ranking with one created by human judges (Turian et al., 2003). The closer a system's ranking is to the human's, the better the evaluation system is.

4.1 Validating ParaEval

To test ParaEval's ability, NIST 2003 Chinese MT evaluation results were used (NIST 2003). This collection consists of 100 source documents in Chinese, translations from eight individual translation systems, reference translations from four humans, and human assessments (on fluency and adequacy). The Spearman rank-order coefficient is computed as an indicator of how close a system ranking is to gold-standard human ranking. It should be noted that the 2003 MT data is separate from the corpus that we extracted paraphrases from.

For comparison purposes, BLEU[1] was also run. Table 1 shows the correlation figures for the two automatic systems with the NIST rankings on fluency and adequacy. The lower and higher 95% confidence intervals are labeled as "L-CI" and "H-CI". To estimate the significance of the rank-order correlation figures, we applied bootstrap resampling to calculate the confidence intervals. In each of 1000 runs, systems were ranked based on their translations of 100 randomly selected documents. Each ranking was compared with the NIST ranking, producing a correlation score for each run. A t-test was then performed on the 1000 correlation scores. In both *fluency* and *adequacy* measurements, ParaEval correlates significantly better than BLEU. The ParaEval scores used were precision scores. In addition to distinguishing the quality of MT systems, a reliable evaluation procedure must be able to distinguish system translations from humans' (Lin and Och, 2004). Figure 5 shows the overall system and human ranking. In the upper left corner, human translators are grouped together, significantly separated from the automatic MT systems clustered into the lower right corner.

	BLEU	ParaEval
Fluency	0.6978	0.7575
95% L-CI	0.6967	0.7553
95% H-CI	0.6989	0.7596
Adequacy	0.6108	0.6918
95% L-CI	0.6083	0.6895
95% H-CI	0.6133	0.694

Table 1. Ranking correlations with human assessments.

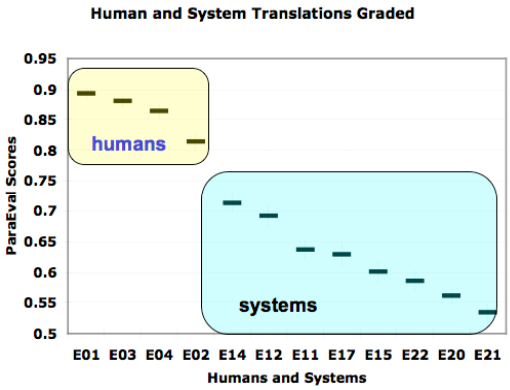

Figure 5. Overall system and human ranking.

4.2 Implications to Word-alignment

We experimented with restricting the paraphrases being matched to various lengths. When allowing only paraphrases of three or more words to match, the correlation figures become stabilized and ParaEval achieves even higher correlation with *fluency* measurement to 0.7619 on the Spearman ranking coefficient.

This phenomenon indicates to us that the bigram and unigram paraphrases extracted using SMT word-alignment and phrase extraction programs are not reliable enough to be applied to evaluation tasks. We speculate that word pairs extracted from (Liang et al., 2006), where a bidirectional discriminative training method was used to achieve consensus for word-alignment

[1] Results shown are from BLEU v.11 (NIST).

(mostly lower n-grams), would help to elevate the level of correlation by ParaEval.

4.3 Implications to Evaluating Paraphrase Quality

Utilizing paraphrases in MT evaluations is also a realistic way to measure the quality of paraphrases acquired through unsupervised channels. If a comparison strategy, coupled with paraphrase matching, distinguishes good and bad MT and summarization systems in close accordance with what human judges do, then this strategy and the paraphrases used are of sufficient quality. Since our underlining comparison strategy is that of BLEU-1 for MT evaluation, and BLEU has been proven to be a good metric for their respective evaluation tasks, the performance of the overall comparison is directly and mainly affected by the paraphrase collection.

5 ParaEval's Support for Recall Computation

Due to the use of multiple references and allowing an n-gram from the peer translation to be matched with its corresponding n-gram from any of the reference translations, BLEU cannot be used to compute recall scores, which are conventionally paired with precision to detect length-related problems from systems under evaluation.

5.1 Using Single References for Recall

The primary goal in using multiple references is to overcome the limitation in matching on lexical identity. More translation choices give more variations in verbalization, which could lead to more matches between peer and reference translations. Since MT results are generated and evaluated at a sentence-to-sentence level (or a segment level, where each segment may contain a small number of sentences) and no text condensation is employed, the number of different and correct ways to state the same sentence is small. This is in comparison to writing generic multi-document summaries, each of which contains multiple sentences and requires significant amount of "rewriting". When using a large collection of paraphrases while evaluating, we are provided with the alternative verbalizations needed. This property allows us to use single references to evaluate MT results and compute recall measurements.

5.2 Recall and Adequacy Correlations

When validating the computed recall scores for MT systems, we correlate with human assessments on *adequacy* only. The reason is that according to the definition of recall, the content coverage in references, and not the fluency reflected from the peers, is being measured. Table 2 shows ParaEval's recall correlation with NIST 2003 Chinese MT evaluation results on systems ranking. We see that ParaEval's correlation with *adequacy* has improved significantly when using recall scores to rank than using precision scores.

	BLEU	ParaEval
Adequacy	0.6108	0.7373
95% L-CI	0.6083	0.7368
95% H-CI	0.6133	0.7377

Table 2. ParaEval's recall ranking correlation.

5.3 Not All Single References are Created Equal

Human-written translations differ not only in word choice, but also in other idiosyncrasies that cannot be captured with paraphrase recognition. So it would be presumptuous to declare that using paraphrases from ParaEval is enough to allow using just one reference translation to evaluate. Using multiple references allow more paraphrase sets to be explored in matching.

In Table 3, we show ParaEval's correlation figures when using single reference translations. E01–E04 indicate the sets of human translations used correspondingly.

	E01	E02	E03	E04
Fluency	**0.683**	0.6501	**0.7284**	0.6192
95% L-CI	0.6795	0.6482	0.7267	0.6172
95% H-CI	0.6864	0.6519	0.73	0.6208
Adequacy	**0.6308**	0.5741	**0.6688**	0.5858
95% L-CI	0.6266	0.5705	0.665	0.5821
95% H-CI	0.635	0.5777	0.6727	0.5895

Table 3. ParaEval's correlation (precision) while using only single references.

Notice that the correlation figures vary a great deal depending on the set of single references used. How do we differentiate human translations and know which set of references to use? It is difficult to quantify the quality that a human written translation reflects. We can only define "good" human translations as translations that are written not very differently from what other humans would write, and "bad" translations as the ones that are written in an unconventional fashion. Table 4 shows the differences between the four sets of reference translations when com-

paring one set of references to the other three. The scores here are the raw ParaEval precision scores. E01 and E03 are better, which explains the higher correlations ParaEval has using these two sets of references individually, shown in Table 3.

	ParaEval	95% L-CI	95% H-CI
E01	**0.8086**	**0.8**	**0.8172**
E02	0.7383	0.7268	0.7497
E03	**0.7839**	**0.7754**	**0.7923**
E04	0.7742	0.7617	0.7866

Table 4. Differences among reference translations (raw ParaEval precision scores).

6 Observation of Change in Number of References

When matching on lexical identity, it is the general consensus that using more reference translations would increase the reliability of the MT evaluation (Turian et al., 2003). It is expected that we see an improvement in ranking correlations when moving from using one reference translation to more. However, when running BLEU for the NIST 2003 Chinese MT evaluation, this trend is inverted, and using single reference translation gave higher correlation than using all four references, as illustrated in Table 5.

BLEU	E01	E02	E03	E04	4 refs
Fluency	0.7114	0.701	0.7084	0.7192	0.6978
95% L-CI	0.7099	0.6993	0.7065	0.7177	0.6967
95% H-CI	0.7129	0.7026	0.7102	0.7208	0.6989
Adequacy	0.644	0.6238	0.6535	0.675	0.6108
95% L-CI	0.6404	0.6202	0.6496	0.6714	0.6083
95% H-CI	0.6476	0.6274	0.6574	0.6786	0.6133

Table 5. BLEU's correlating behavior with multi- and single-reference.

Turian et al. (2003) reports the same peculiar behavior from BLEU on Arabic MT evaluations in Figure 5b of their paper. When using three reference translations, as the number of segments (sentences usually) increases, BLEU correlates worse than using single references.

Since the matching and underlining counting mechanisms of ParaEval are built upon the fundamentals of BLEU, we were keen to find out the differences, other than paraphrase matching, between the two methods when the number of reference translation changes. By following the description from the original BLEU paper, three incremental steps were set up for duplicating its implementation, namely modified unigram precision (MUP), geometric mean of MUP (GM), and multiplying brevity penalty with GM to get the final score (BP-BLEU). At each step, correlations were computed for both using single- and multi- references, shown in Table 6a, b, and c.

MUP	E01	E02	E03	E04	4 refs
Fluency	0.6597	0.6216	0.6923	0.4912	0.692
95% L-CI	0.6568	0.6189	0.6917	0.4863	0.6915
95% H-CI	0.6626	0.6243	0.6929	0.496	0.6925
Adequacy	0.5818	0.5459	0.6141	0.4602	0.6165
95% L-CI	0.5788	0.5432	0.6132	0.4566	0.6156
95% H-CI	0.5847	0.5486	0.6151	0.4638	0.6174

6(a). System-ranking correlation when using modified unigram precision (MUP) scores.

GM	E01	E02	E03	E04	4 refs
Fluency	0.6633	0.6228	0.6925	0.4911	0.6922
95% L-CI	0.6604	0.6201	0.692	0.4862	0.6918
95% H-CI	0.6662	0.6255	0.6931	0.4961	0.6929
Adequacy	0.5817	0.548	0.615	0.4641	0.6159
95% L-CI	0.5813	0.5453	0.614	0.4606	0.615
95% H-CI	0.5871	0.5508	0.616	0.4676	0.6169

6(b). System-ranking correlation when using geometric mean (GM) of MUPs.

BP-BLEU	E01	E02	E03	E04	4 refs
Fluency	0.6637	0.6227	0.6921	0.4947	0.5743
95% L-CI	0.6608	0.62	0.6916	0.4899	0.5699
95% H-CI	0.6666	0.6254	0.6927	0.4996	0.5786
Adequacy	0.5812	0.5486	0.5486	0.5486	0.6671
95% L-CI	0.5782	0.5481	0.5458	0.5458	0.6645
95% H-CI	0.5842	0.5514	0.5514	0.5514	0.6697

6(c). System-ranking correlation when multiplying the brevity penalty with GM.

Table 6. Incremental implementation of BLEU and the correlation behavior at the three steps: MUP, GM, and BP-BLEU.

Given that many small changes have been made to the original BLEU design, our replication would not produce the same scores from the current version of BLEU. Nevertheless, the inverted behavior was observed in *fluency* correlations at the BP-BLEU step, not at MUP and GM. This indicates to us that the multiplication of the brevity penalty to balance precision scores is problematic. According to (Turian et al., 2003), correlation scores computed from using fewer references are inflated because the comparisons exclude the longer n-gram matches that make automatic evaluation procedures diverge from the human judgments. Using a large collection of paraphrases in comparisons allows those longer n-gram matches to happen even if single references are used. This collection also allows ParaEval to directly compute recall scores, avoiding an approximation of recall that is problematic.

7 Conclusion and Future Work

In this paper, we have described ParaEval, an automatic evaluation framework for measuring machine translation results. A large collection of paraphrases, extracted through an unsupervised fashion using SMT methods, is used to improve the quality of the evaluations. We addressed three important issues, the paraphrasing support, the computation of recall measurement, and providing high correlations with human judgments.

Having seen that using paraphrases helps a great deal in evaluation tasks, naturally the next task is to explore the possibility in paraphrase induction. The question becomes how to use contextual information to calculate semantic closeness between two phrases. Can we expand the identification of paraphrases to longer ones, ideally sentences?

The problem in which content bearing words carry the same weights as the non-content bearing ones is not addressed. From examining the paraphrase extraction process, it is unclear how to relate translation probabilities and confidences with semantic closeness. We plan to explore the parallels between the two to enable a weighted implementation of ParaEval.

Reference

Anonymous. 2006. Complete citation omitted due to the blind review process.

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. *Proceedings of ACL-2005*.

Browner, J. 2006. The translator's blues. http://www.slate.com/id/2133922/.

Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT/NAACL-2006*.

Callison-Burch, C., M. Osborne, and P. Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL-2006*.

Inkpen, D. Z. and G. Hirst. 2003. Near-synonym choice in natural language generation. *Proceedings of RANLP-2003*.

Leusch, G., N. Ueffing, and H. Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*.

Liang, P., B. Taskar, and D. Klein. Consensus of simple unsupervised models for word alignment. In *Proceedings in HLT/NAACL-2006*.

Lin, C.Y. and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the HLT-2003*.

Lin, C.Y. and F. J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of ACL-2004*.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19—51, 2003.

Och, F. J. and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4), 2004.

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. 2002. IBM research report Bleu: a method for automatic evaluation of machine translation *IBM Research Division Technical Report*, RC22176, 2001.

Turian, J. P., L. Shen, and I. D. Melamed. 2003. Evaluation of machine translation and its evaluation. *Proceedings of MT Summit IX*.