UPPER BOUNDS OF THE BLEU ALGORITHM APPLIED TO ASSESSING STUDENT ESSAYS

DIANA PEREZ, ENRIQUE ALFONSECA AND PILAR RODRIGUEZ

UPPER BOUNDS OF THE BLEU ALGORITHM APPLIED TO ASSESSING STUDENT ESSAYS

Diana Perez, Enrique Alfonseca and Pilar Rodriguez
Computer Science Department
Universidad Autonoma de Madrid
28049 Madrid (SPAIN)
{diana.perez, enrique.alfonseca, pilar.rodriguez}@ii.uam.es

Abstract

In a previous work (Perez et al., 2004) we applied Bleu (Papineni et al., 2001) in assessing short essays written by students. We showed that it outperforms other keyword-based algorithms in assessing students' essays, and varied some of its internal parameters. In this study, we perform a global search on our sets of possible references, using genetic algorithms, in order to find out the highest accuracy that it can attain when assessing students' answers. For each question, the automatic search is done among over fifty possible references, so we can be confident that the correlation attained constitutes an upper bound for the Bleu algorithm. Finally, conclusions are drawn from a comparison between our system's scores and the teacher's, and we consider some possible extensions of BLEU.

1. Introduction

It has been noted that assessment based only in multiple-choice, fill-in-the-blank or yes/no questions is not accurate enough to measure the amount of knowledge the students have acquired, or whether they have understood the subject (Whittington and Hunt, 1999). We can assume that this holds both for traditional learning and for computer-based assessment. On the other hand, evaluating free-text answers automatically is not a trivial task to tackle, as it requires some degree of natural-language understanding of the student answers. This problem has attracted interest from the research community since the sixties (Page, 1968), and has become a flourishing research line in the last few years, giving rise to a couple of specialised conferences such as the Computer-Assisted Assessment Conference (Danson and Eabry, 2001; Danson, 2002).

There are currently several approaches to solve this problem:

- Combining keyword-based methods (e.g. the vector-space model) with deep analyses of texts (Burstein et al., 2001).
- Using pattern-matching techniques (Ming et al., 1999).
- Breaking the answers into concepts and their semantic dependencies (Callear et al., 2001).

- Combining several Machine Learning techniques (Rosé et al., 2003).
- Reducing the dimension space with Latent Semantic Analysis (LSA) (Foltz et al., 1999).
- Improving LSA with syntactic and semantic information (Wiemer-Hastings and Zipitria, 2001; Kanejiya and Prasad, 2003).

A classification of techniques to automatically assess free-text answers can be found in Mitchell et al. (2002), who distinguish three main kinds:

- **Keyword analysis,** the simplest method, which consists in looking for coincident keywords or n-grams. It has usually been considered a poor method, as it cannot deal with synonyms or with polysemous terms in the student answers.
- **Full natural-language processing,** which consists of a deep text parsing and semantic analysis in order to gather more information about the meaning of the student's answer. It may improve on the previous technique, but it is hard to accomplish and very difficult to port across languages.
- **Information Extraction techniques,** which offer an affordable and more robust approach, making use of NLP tools for searching the texts for some specific contents, and without doing an in-depth analysis.

An overview of tools that perform automatic evaluation of assessments can be found in Valenti et al. (2003).

In previous work (Perez et al., 2004), we applied BLEU (Papineni et al., 2001) to evaluate student answers. The method was originally conceived for ranking Machine Translation systems, by comparing a candidate translation to several reference texts (human-made translations). It has shown a good potential to be used in an e-learning environment, given that:

- The text to be translated can be considered the student's answer.
- The reference text is no longer written by a translator but by the teacher.

On the other hand, there are some obvious differences with respect to assessment evaluation:

- The structure of the student's answer is completely free, whereas a translation usually follows closely the rhetorical structure of the original text.
- BLEU is a ranking procedure, whose aim is to decide which system outperforms the rest, whilst we would like to mark student's answers, independently of each other.

- In the case of Machine Translation, it is to be expected that the system's automatic translation fully covers the answer. Therefore, it is more important to check that the output is correct (i.e. to measure its n-gram precision) than to check that it is complete. It is interesting to note that a similar procedure has been applied to evaluating automatic summarisation systems (Lin and Hovy, 2003). However, in this case, the focus is placed on finding whether the candidate summary contains all the relevant information, and therefore it calculates n-gram recall. We believe that, in the case of student answers, both precision and recall are to be measured.

The aim of this paper is double. Firstly, our previous experiments show that BLEU outperforms other keyword-based metrics, but these results are very dependent on the type of question, and in a few cases the correlation with teacher's scores is dramatically low. Therefore, we have performed a study for identifying the upper bounds of BLEU when applied to computer-based assessment. Secondly, we analyse the results obtained and make conclusions on the weakest points of the procedure and describe possible solutions.

The paper is organised as follows: in section 2 we depict the experimental settings. Section 3 explores a way in which to choose the reference texts so that BLEU scores correlate best with teachers' scores. Section 4 describes a variation of BLEU in which precision has been combined with recall in the same measure. Finally, conclusions are drawn and future work is described in Sections 5 and 6.

2. Experimental settings

2.1. Data sets

We have built six benchmark datasets. They all contain students answers from real exams about the topic of Computer Science, written in Spanish. The sets are described in Table 1.

SET	***NC***	***MC***	***NR***	***MR***	***LANG.***	***TYPE***
1	79	51	3	42	Spanish	Def.
2	96	44	4	30	Spanish	Def.
3	143	48	7	27	Spanish	A/D
4	295	56	8	55	Spanish	A/D
5	117	127	5	71	Spanish	Y/N
6	117	166	3	186	Spanish	A/D

Table 1: The MC column indicates the mean length of the candidate texts, NR the number of reference texts, MR the mean length of the reference texts, LANG. the texts language and TYPE the type of question: Def. for definitions, A/D for advantages/disadvantages and Y/N for yes/no and justification. More details in (Perez et al., 2004).

Each of the six sets contains all the student answers to a single question. We have classified the questions in one of the three following types:

- Definitions and descriptions, e.g. *What is an operating systems?; Describe how to encapsulate a class in C++.*
- Advantages and disadvantages, e.g. *Enumerate the advantages and disadvantages of the token ring algorithm.*
- Yes/No questions, e.g. *Is RPC appropriate for a chat server? (Justify your answer)*.

All the answers have been manually scored by at least two teachers.

2.2. Evaluation of the procedure

In all the experiments, we evaluate the goodness of BLEU with the correlation value between the automatic scores and the judge's. This is a common measure to find out the behaviour of computer-based assessment methods.

3. Upper bounds for BLEU's correlation value

3.1. The experiment

The BLEU algorithm is very dependent on the quality of the reference answers written by the teacher. We have observed that the results can dramatically change depending on the quality and number of these texts (Perez et al., 2004), and there is no way for the teachers to be sure that the references they have written are appropriate for the evaluation. Therefore, we present here a procedure that can be used for finding the best references amongst a large dataset.

The procedure consists in searching, from a large set of possible references, which selection of references produces the best correlation between BLEU's scores and the gold standard. Because we search among a very large set of possible reference answers, it is not probable that a teacher would write a larger amount of correct answers for his questions, and thus we expect that the best correlation obtained can be taken as the upper bound that BLEU will hardly exceed in this task.

Given that we have, for each question, a small set of references written by the teacher and a large set of student's answers already marked, we decided to choose the best references for BLEU among the union of those sets. The procedure is the following:

1. Join the set with the teacher's references and the student's answers.
2. Divide it into two subsets. One will be taken as reference texts, and the answers in the other subset will be scored with BLEU against those references.
3. Repeat 2 for each possible subset choice.

The possible sets of references out of a set with roughly 100 texts is the size of the superset, 2^{100}, and it would be computationally very expensive

to repeat step 2 as many times. This is the reason why we have chosen a genetic algorithm to guide the search. The results are shown in Table 2. As can be seen, we could always find a selection of reference texts for which the accuracy of BLEU, measured as the correlation with the teacher's scores, improved significantly with respect to our original sets of references.

Data set	***No. of references from GA***	***BLEU score, Original references***	***BLEU score, References from GA***
1	22	0.3609	0.6616
2	5	0.3694	0.4720
3	55	0.4159	0.8187
4	20	0.0209	0.4187
5	48	0.2102	0.7025
6	41	0.4172	0.7223

Therefore, we propose the following iterative procedure for any teacher who would like to apply BLEU in evaluating student's answers (either as a stand-alone procedure or as a help module to other existing e-learning system):

1. Write a few reference answers (five or six) for the question.
2. Obtain the set of answers from the students.
3. Mark them with BLEU, and supervise the markings.
4. Put together all th

3.2. Analysis of the results

Figure 1 plots BLEU's marks against the gold standard for data sets 1 and 3. As expected (because the correlation was positive), the regression line grows with the teacher's score.

We can draw some observations from the graphs. Firstly, we can observe that the teachers' marks are discrete and, in most cases, teachers simply mark the answers as right or wrong. Therefore, most of the points appear either at the leftmost or the rightmost side of the graph. This fact is very evident for set 3, where only a few dots are located at the middle of the image. Still, BLEU's scores correlate well with the teacher's scores. For that set, most of the wrong answers received a BLEU score below 0.5, and most of the right ones received a score above 0.5, so BLEU is distinguishing between right and wrong replies.

Figure 2 (left) shows the regression line for set 4. With this set, the correlation is rather poor, as the data points are evenly scattered around the regression line regardless of the teachers' score. This is probably the hardest data set, as the answers had to be an enumeration of advantages and disadvantages. We believe that in evaluating this question, it will be necessary to perform some other processing such as deciding whether each of the student's statements refers to an advantage or not. The correlation attained for this set has been the lowest among all sets.

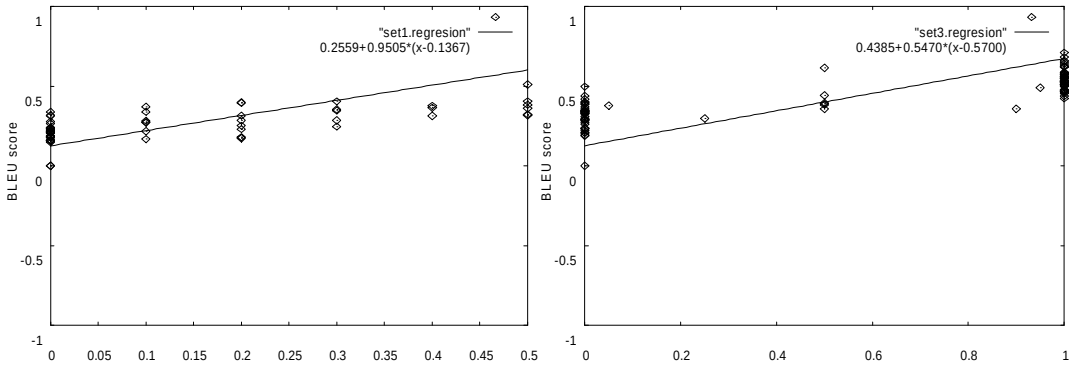

Figure 1: Regression lines between the teachers' scores and BLEU's marks for sets 1 and 3.

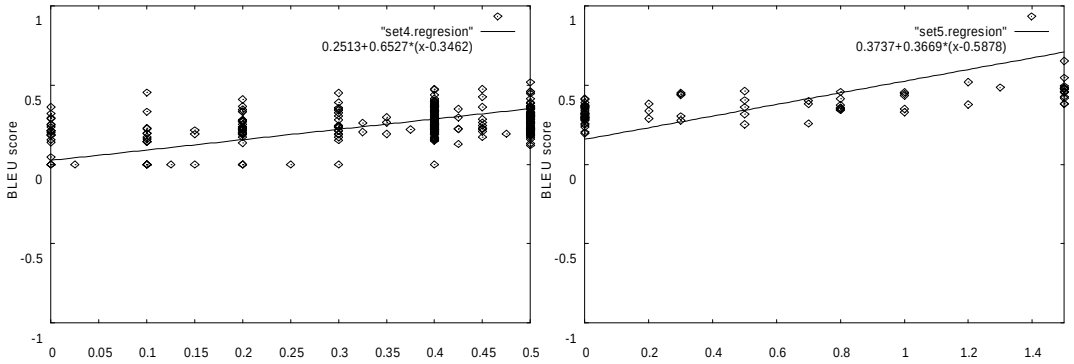

Figure 2: Regression lines between the teachers' scores and BLEU's marks for sets 4 and 5.

Figure 2 (right) displays the regression line for data sets 5. The results for this set, together with set 6, have been surprising for us. These are yes/no questions in which the students had to justify their answers, so we expected further pre-processing necessary in order to detect whether the answer was affirmative or negative. However, BLEU has been able to attain a 0.7 correlation on these sets.

The answer may lie in the gold standard. If we examine the student answers, we can see that the teachers have not marked the simple yes-no decision, but the student's reasoning supporting it. Many students do get a high mark in the gold-standard even though they have answered wrongly the question, because their reasoning is correct. BLEU acts in the same way, disregarding the yes-no, and evaluating the words in the student's discussion.

3.3. Comparison between BLEU and the teacher scores

BLEU always gives a score between 0 and 1, but that is not necessarily the scale used by the teacher. However, the regression lines can be used to obtain the score in the teacher's scale from the score in BLEU's scale. We just have consider BLEU's score as the independent variable, and obtain the expected teacher's score.

The next experiment performed is the following:

1. For each of the data sets, evaluate all the answers with BLEU.
2. Use the regression line to get the scores using the teacher's scale.
3. Calculate the deviation between BLEU's scores and the gold standard.
4. Generate the histogram of the deviations.

Figure 3 shows the histograms obtained when subtracting the gold standard's scores from the scores obtained automatically, for sets 1 and 2. The X axis represents this result multiplied tenfold. We can observe that most of the answers either receive the same score or 0.1 points more.

In Figure 4, we can see that the histograms for sets 4 and 6 the results are very similar, and follow a gaussian. Again, most of the times the difference is lower than 0.2.

Finally, Figure 5 shows the histograms obtained for Sets 3 and 5. We can see that here many answers have been very badly evaluated by BLEU, as the differences between the automatic score and the gold standard can be very high. In particular, just 1 answer from set 5 received the same score as the gold standard! We believe that the problem here is that, for these data sets, most of the answers are considered by the teacher completely right or wrong, receiving either the highest or the lowest mark. BLEU, on the other hand, scores them with many intermediate values, so the differences in the scores are high. Probably, for these questions, the automatic evaluation should just classify the answers as correct or incorrect, rather than trying to give them a quantitative score.

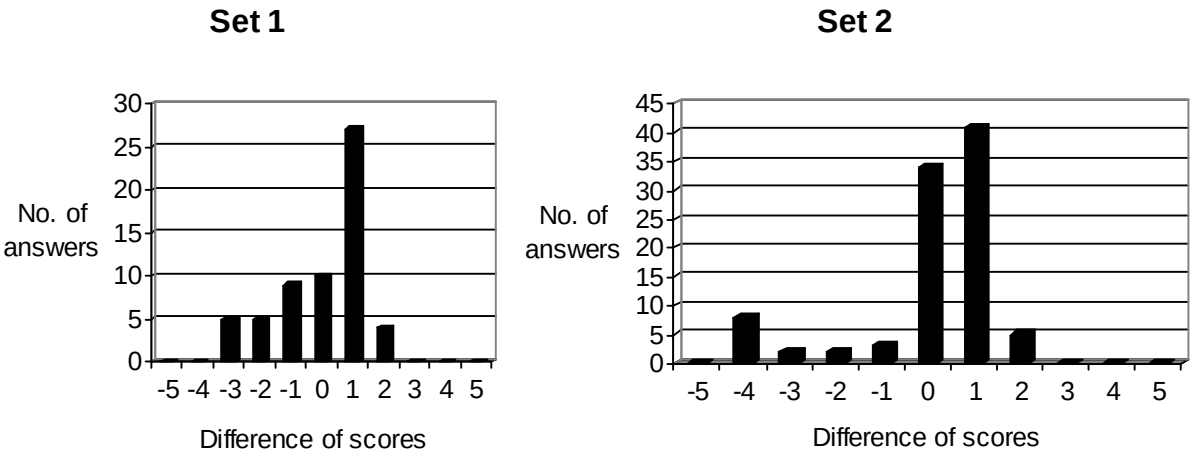

Figure 3: Histogram for the two first data sets (definitions)

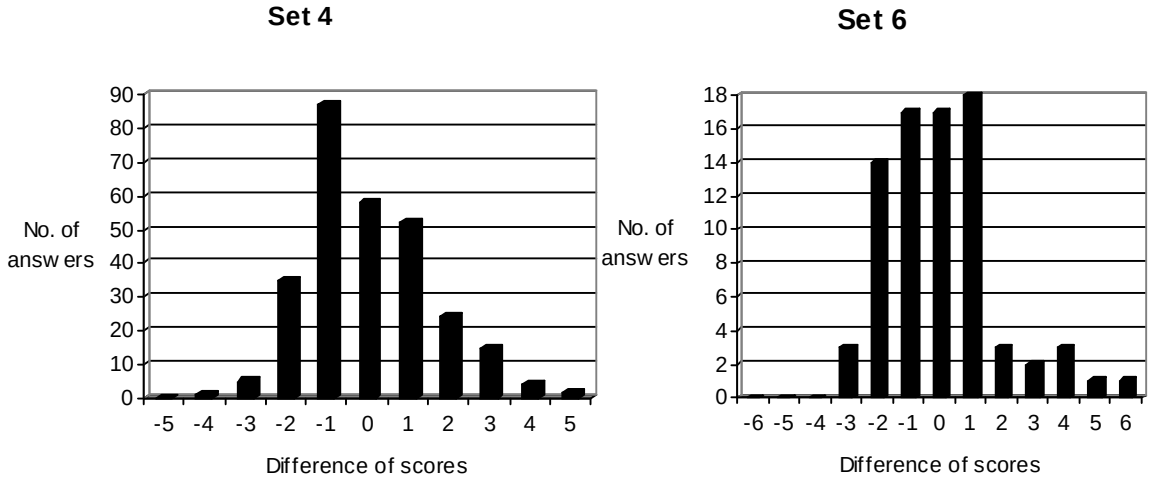

Figure 4: Histogram for the data sets 4 and 6.

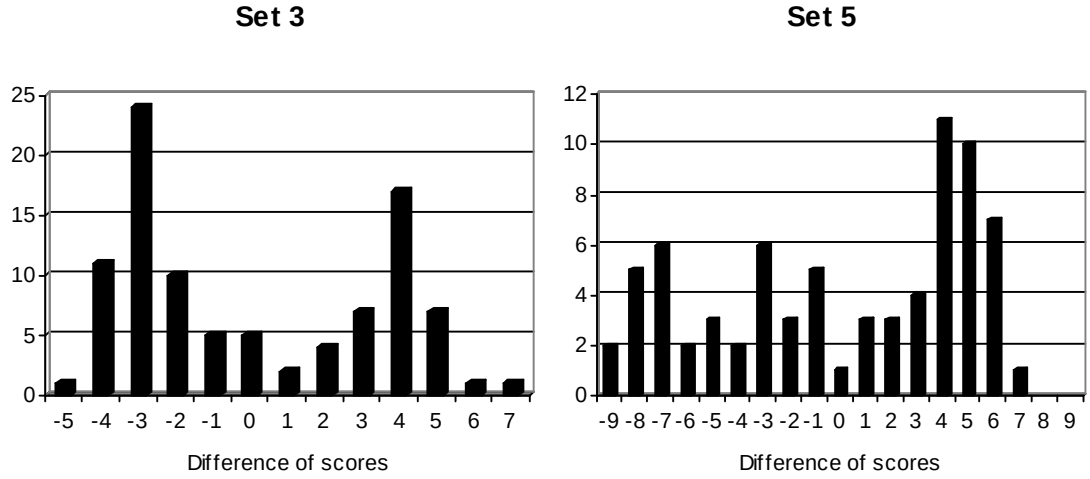

Figure 5 : Histograms for data sets 3 and 5.

4. Extending BLEU with a recall metric

The original version of BLEU consists basically of a clipped n-gram recall: it calculates, for several values of N (typically from 1 to 4) the percentage of N-grams from the candidate text that appear in any reference. When evaluating student answers, we also need to make sure that the student has not left out relevant information. BLEU takes this into account with a simple Brevity Penalty factor that compares the lengths of the reference and the candidate, but we believed that this factor could be improved by calculating the percentage of information from the references that had not been considered, with the following procedure:

1. For N from a maximum (e.g. 10) down to 1, repeat:
    a. For each N-gram from the candidate text that has not yet been found in any reference text,
    b. If it appears in any reference, mark the words from the N-gram as found, both in the candidate and the reference.
2. For each reference text, count the number of words that are marked, and calculate the percentage of the reference that has been found.
3. The Modified Brevity Penalty factor (MBP) is the sum of all the percentage values.

The procedure is shown in Figure 6.

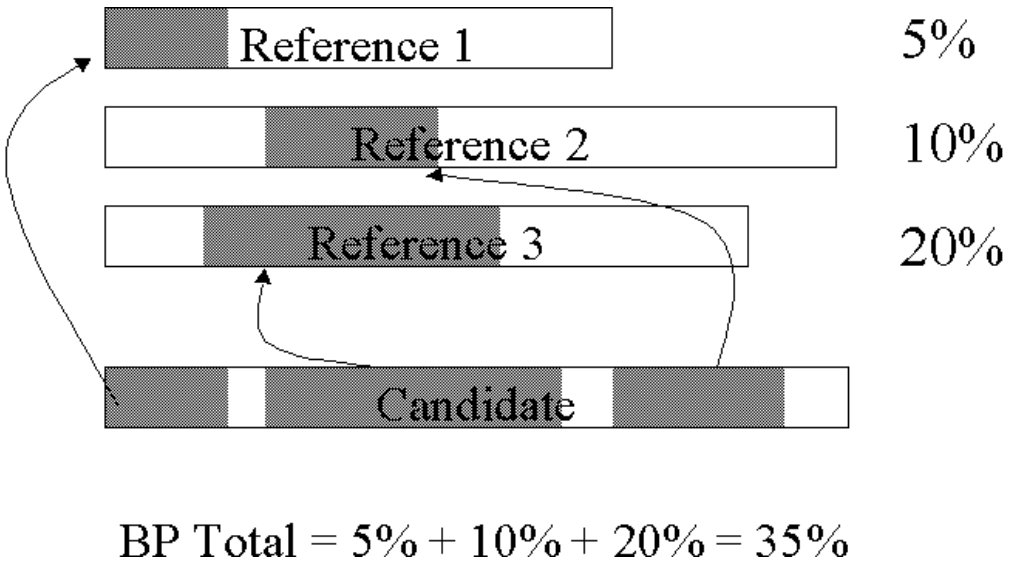

Figure 6: Procedure for calculating the Modified Brevity Penalty factor.

Table 3 (first two columns) shows the results of applying the new factor to the six data sets. It can be seen that for most cases this new Brevity Penalty factor improves the correlation. As we noted previously (Perez et al, 2004), the use of this factor has yet another consequence. The original version of BLEU produced the highest results using a combination of unigrams, bigrams and trigrams for calculating the N-gram precision. In contrast, we have noted that the addition of the new MBP factor produces the best results just by calculating the unigram precision. This last accuracy is shown in the last column.

Data set	***Original BLEU***	***BLEU+MBP(1:3)***	***BLEU+MBP(1:1)***
1	0.3609	0.4249	**0.5262**
2	**0.3694**	0.3615	0.3546
3	0.4159	0.6135	**0.6420**
4	0.0209	0.1223	**0.1756**
5	0.2102	0.2750	**0.4247**
6	0.4172	0.4106	**0.4308**

Table 3: The correlation values obtained while using the original BLEU, BLEU with the modified brevity penalty factor (MBP) and a combination of unigrams, bigrams and trigrams, and BLEU with MBP and only unigrams.

It must be noted that MBP is a little bit sensitive to the order of the reference texts. For instance, let us suppose that the candidate text has a length L, and it appears as a whole in two references, of lengths L and 2L. If the first reference was chosen, MBP would have a value of 100%, because there is a reference which is wholly covered. On the other hand, if we choose the second reference, MBP will have a value of 50%. Even so, in practice, we have only observed small variations of the results with respect to the order of the references.

5. Conclusions

We have previously applied BLEU for evaluating student essays (Perez et al., 2004). From those experiments we had concluded that it was good enough to substitute other keyword-based methods in existing applications, but the correlation was still not as good as to be used in practice as a stand-alone application. Some of its drawbacks are its dependency on the number and quality of the reference texts, its unsuitability for every type of questions, and the fact that it has not been designed to measure recall. On the other hand, its main advantages are its simplicity and language-independence.

One aim of this paper was to find the upper bounds of this method. We have estimated them using a large set of possible reference texts, and using genetic algorithms to find the combination of references which maximises the correlation. We are confident that, in practice, the results can be considered BLEU's upper bounds when evaluating our data sets, because it is unlikely that someone will use larger sets of references from which to choose the best ones. We have also provided a procedure for a teacher to train the system using the student's answers.

Secondly, by analysing the results, we have observed that there is much variation in the way the scoring algorithms perform, depending on the question. From our study, we conclude that some questions should may

P. W. Foltz, D. Laham, and T. K. Landauer. 1999. “Automated essay scoring: Applications to educational technology”. In *Proceedings of EdMedia’99*.

D. Kanejiya and S. Prasad. 2003. “Automatic evaluation of students’ answers using syntactically enhanced LSA”. In *Proceedings Of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, pages 53–60.

C.Y. Lin and E. H. Hovy. 2003. “Automatic evaluation of summaries using n-gram co-occurrence statistics”. In *Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada.

Y. Ming, A. Mikhailov, and T. Lay Kuan. 1999. “Intelligent essay marking system”. In *Educational Technology Conference*.

T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge. 2002. “Towards robust computerised marking of free-text responses”. In *Proceedings of the 6th International Computer Assisted Assessment (CAA) Conference*, Loughborough, UK.

E. B. Page. 1968. “The use of computer in analyzing student essays”. *International review of education*, 14, 210-224.

K. Papineni, S. Roukos, T. Ward and W. Zhu, “BLEU: a method for automatic evaluation of machine translation”, Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center (2001).

D. Perez, E. Alfonseca and P. Rodriguez, “Application of the Bleu method for evaluating free-text answers in an e-learning environment”, to appear in *Proceedings of LREC-2004*, Lisbon, 2004.

C. P. Rosé, A. Roque, D. Bhembe, and K. VanLehn. 2003. “A hybrid text classification approach for analysis of students essays”. In *Building Educational Applications Using Natural Language Processing*, pages 68–75.

S. Valenti, F. Neri, and A. Cucchiarelli. 2003. “An overview of current research on automated essay grading”. *Journal of Information Technology Education*, 2:319–330.

D. Whittington and H. Hunt. 1999. “Approaches to the computerized assessment of free text responses”. In *M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.

P.Wiemer-Hastings and I. Zipitria. 2001. “Rules for syntax, vectors for semantics”. In *Proceedings of the 23rd annual Conf. of the Cognitive Science Society*, Mawhwah. N.J. Erlbaum.