Application of the BLEU method for evaluating free-text answers in an e-learning environment

Diana Pérez, Enrique Alfonseca and Pilar Rodríguez

* Departamento de Ingeniería Informática, Universidad Autónoma de Madrid
28049 Madrid, Spain
{Diana.Perez,Enrique.Alfonseca,Pilar.Rodriguez}@ii.uam.es

Abstract

We have applied BLEU (Papineni et al., 2001), a method originally designed to evaluate automatic Machine Translation systems, in assessing short essays written by students. We study how much BLEU scores correlate to human scorings and other keyword-based evaluation metrics. We conclude that, although it is only applicable to a restricted category of questions, BLEU attains better results than other keyword-based procedures. Its simplicity and language-independence makes it a good candidate to be combined with other well-studied computer assessment scoring procedures.

1. Introduction

As has recently been noted, it has not been proved that the most common types of exercises in computer-based courses (e.g. fill-in-the-blank, multi-choice or yes/no questions) can always measure higher order cognitive skills (Whittington and Hunt, 1999). Therefore, there is an increasing interest for automatic assessment of free-text answers from the e-learning community. The success of the Computer-Assisted Assessment (CAA) conferences (Danson and Eabry, 2001; Danson, 2002), amongst others, has also motivated advances in the area. CAA can be applied in e-Learning environments to give feedback to the students, and as auxiliary tools for the teacher.

Mitchell et al. (2002) classifies traditional marking of free-text responses in two main kinds, *keyword analysis* and *full natural-language processing*, to which they add a third kind based on *Information Extraction techniques*. Keyword analysis has usually been considered a poor method, given that it is difficult to tackle problems such as synonymy or polysemy in the student answers; on the other hand, a full text parsing and semantic analysis is hard to accomplish, and very difficult to port across languages. Hence, Information Extraction offers an affordable and more robust approach, making use of NLP tools for searching the texts for the specific contents and without doing an in-depth analysis (Mitchell et al., 2002).

Other approaches work either by combining keyword-based methods (e.g. the vector-space model) with deep analyses of texts (Burstein et al., 2001); by using pattern-matching techniques (Ming et al., 1999); by breaking the answers into concepts and their semantic dependencies (Callear et al., 2001); by combining several Machine Learning techniques (Rosé et al., 2003); or by reducing the dimension space with Latent Semantic Analysis (LSA) (Foltz et al., 1999). Simple LSA has been improved with syntactic and semantic information (Wiemer-Hastings and Zipitria, 2001; Kanejiya and Prasad, 2003). Valenti et al. (2003) review several systems that performs automatic evaluation of assessments.

This paper presents an application of the BLEU algorithm (Papineni et al., 2001) for the assessment of students' short essays. This method has been applied for ranking Machine-Translation systems and, with some variations, for automatic evaluation of summarisation procedures (Lin and Hovy, 2003). We argue that this procedure, although it does not attain a correlation high enough to be used as a stand-alone assessment tool, improves other existing keyword-based procedures and it is a good candidate for replacing them in existing applications. It keeps all their advantages (language-independence and simplicity), and produces better results.

This paper is organised as follows: section 2. describes briefly the BLEU method; section 3. details the application to e-learning environments and the results obtained; and, finally, section 4. ends with the conclusions and future work.

2. The BLEU method

The BLEU method (Papineni et al., 2001) was proposed as a rapid way for evaluating and ranking Machine Translation systems. Its robustness stems from the fact that it works with several reference texts (human-made translations), against which it compares the candidate text. The procedure is the following:

1. Count how many N-grams from the candidate text appear in any of the reference text (for different values of N). The frequency of each N-gram is clipped with the maximum frequency with which it appears in any reference.

2. Combine the marks obtained for each value of N, as a weighted linear average.

3. Apply a brevity factor to penalise the short candidate texts (which may have many N-grams in common with the references, but may be incomplete).

The use of several references, made by different human translators, increases the probability that the particular words and their relative order, in the automatic translation, will appear in any reference. On the other hand, the procedure is very sensitive to the choice of the reference translations.

SET	NC	MC	NR	MR	Lang	Type	Description
1	38	67	4	130	En	Def.	"Operating System" definitions from "Google Glossary"
2	79	51	3	42	Sp	Def.	Exam question about Operating Systems
3	96	44	4	30	Sp	Def.	Exam question about Operating Systems
4	11	81	4	64	Sp	Def.	Exam question about Object-Oriented Programming
5	143	48	7	27	Sp	Def.	Exam question about Operating Systems
6	295	56	8	55	Sp	A/D	Exam question about Operating Systems
7	117	127	5	71	Sp	Y/N	Exam question about Operating Systems
8	117	166	3	186	Sp	A/D	Exam question about Operating Systems
9	14	118	3	108	Sp	Y/N	Exam question about Operating Systems
10	14	116	3	105	Sp	Def.	Exam question about Operating Systems

Table 1: Answer sets used in the evaluation. Columns indicate: set number; number of candidate texts (NC), mean length of the candidate texts (MC), number of reference texts (NR), mean length of the reference texts (MR), language (En, English; Sp, Spanish), question type (Def., definitions and descriptions; A/D, advantages and disadvantages; Y/N, Yes-No and justification of the answer), and a short description of the question.

3. Experiment: BLEU in e-learning

3.1. Overview

Computer-Assisted Assessment can be considered very related to MT assessment, as:

- The student answer can be treated as the candidate translation whose accuracy we want to score.
- The reference translation made by humans is the reference answer written by the instructors.

3.2. Training and test material

We have built nine different benchmark data from real exams, and an additional one with definitions obtained from Google Glossary[1]. The ten sets are described in Table 1. All the answers were marked by hand by two different human judges, who also wrote the reference texts.

We can classify the ten questions in three distinct categories:

- Definitions and descriptions, e.g. *What is an operative system?, 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 question, e.g. *Is RPC appropriate for a chat server? (Justify your answer)*.

3.3. Experiment performed

The algorithm has been evaluated, for each of the data sets, by comparing the N-grams from the student answers against the references, and obtaining the BLEU score for each candidate. The correlation value between the automatic scores and the judges' scores has been taken as the indicator of the goodness of this procedure.

We have varied the following parameters of BLEU:

- The number of reference texts used in the evaluation process.
- The length (N) of the maximum N-gram to look for coincidences in the reference texts.
- The brevity penalty factor to penalise short answers.

Set	B(1:1)	B(1:2)	B(1:3)	B(1:4)
1	0.4969	0.5409	**0.5886**	0.5090
2	0.3828	**0.4172**	0.3609	0.3029
3	0.2582	0.2733	**0.3694**	0.3170
4	0.6580	0.7286	**0.8220**	0.7226
5	0.3972	**0.4959**	0.4159	0.3169
6	**0.0740**	0.0641	0.0209	0.0305
7	0.2927	**0.3191**	0.2102	0.1293
8	0.2950	0.3592	**0.4172**	0.3403
9	**0.8256**	0.8174	0.7358	0.5948
10	**0.9240**	0.8759	0.7508	0.6131

Table 3: Scores of BLEU for a different choice of the type of N-grams chosen in the comparison, from only unigrams (in the first column) to N-grams with lengths from 1 to 4 (in the last column)

3.4. Results

Number of reference texts As said before, BLEU is very sensitive to the number and the quality of the reference texts available (Papineni et al., 2001). In our experiment, we have varied the number of references from 1 up to the maximum number available for each question (see Table 1, column NR). Table 2 shows the results for each of the data sets. As can be seen, in general, the results improved with the number of references, although in some cases the addition of a new reference having words in common with wrong answers decreased the accuracy of the marks.

Results for different N-grams We have varied the maximum size of the N-grams taken into consideration from 1 to 4. As Table 3 shows, in most cases, the addition of bigrams and trigrams improves the resulting correlation, which indicates that collocations are useful in the evaluation. Given that the students' answers are completely unrestricted, in most of the cases they do not contain enough sequences of four consecutive words from the references, and therefore using four-grams affects the results negatively.

Brevity penalty The BLEU procedure is basically a precision score, as it calculates the percentage of N-grams from the candidate answer which appear in any reference,

Set	No. of reference texts							
	1	2	3	4	5	6	7	8
1	0.3866	0.5738	0.5843	**0.5886**				
2	0.2996	0.3459	**0.3609**					
3	**0.3777**	0.1667	0.1750	0.3693				
4	0.3914	0.3685	0.5731	**0.8220**				
5	0.3430	0.3634	0.3383	0.3909	0.3986	0.4030	**0.4159**	
6	0.0427	0.0245	0.0257	0.0685	**0.0834**	0.0205	0.0014	0.0209
7	0.1256	0.1512	0.1876	0.1982	**0.2102**			
8	0.3615	0.41536	**0.4172**					
9	0.6909	**0.7949**	0.7358					
10	0.7174	**0.8006**	0.7508					

Table 2: Scores of BLEU for a varying number of reference texts

Data set	Original	Modified
1	**0.5886**	0.5525
2	0.3609	**0.4249**
3	**0.3694**	0.3615
4	**0.8220**	0.7674
5	0.4159	**0.6135**
6	0.0209	**0.1223**
7	0.2102	**0.2750**
8	**0.4172**	0.4106
9	**0.7358**	0.7012
10	**0.7508**	0.6357

Table 4: Comparison of the accuracy of BLEU using the original brevity factor and the modified one.

Set	B(1:1)	B(1:2)	B(1:3)	B(1:4)
1	**0.5976**	0.5392	0.5525	0.4707
2	0.5262	**0.5329**	0.4249	0.3117
3	0.3546	0.3247	**0.3615**	0.2463
4	**0.8064**	0.7014	0.7674	0.6802
5	0.6420	**0.6815**	0.6135	0.4901
6	**0.1756**	0.1730	0.1223	0.1077
7	**0.4247**	0.3609	0.2750	0.1870
8	**0.4308**	0.3887	0.4106	0.3531
9	0.6484	**0.7817**	0.7012	0.5776
10	**0.7645**	0.7564	0.6357	0.5707

Table 5: Scores of BLEU, for a different choice of N-gram lengths, with the modified BP.

but it does not check the recall of the answer. Therefore, it applies a brevity penalty so as to penalise very short answers which do not convey the complete information. In its original form (Papineni et al., 2001), BLEU compares the length of the candidate answer against the length of the reference with the most similar length. We have also tested the following brevity penalty:

1. For each reference text, calculate the percentage of its words covered in the candidate.

2. $BP = $ Add up all the percentages.

3. Multiply the basic BLEU score and BP.

As shown in Table 4 the new factor improves the correlation particularly with the data sets with more answers, although a t-test shows that the improvement is not statistically significant at 0.95 confidence.

Different N-grams with modified brevity penalty As shown in Table 5, in contrast to Table 3, if we use the modified Brevity Penalty, then the best results are obtained with just unigrams, a result which is very similar to that obtained by Lin and Hovy (2003) for evaluating summaries.

Comparison with other methods We have implemented two other scoring algorithms as baseline:
- **Keywords**, consisting in calculating the proportion of words which appear in any of the reference texts.

SET	BLEU	Keywords	VSM
1	0.5886	0.0723	0.3100
2	0.3609	0.2390	0.0960
3	0.3694	0.1972	0.2489
4	0.8220	0.5712	-
5	0.4159	0.5705	0.5238
6	0.0209	-0.0551	0.0518
7	0.2102	0.3289	0.1778
8	0.4172	0.2283	0.1789
9	0.7358	0.2487	-
10	0.7508	0.0900	-

Table 6: Comparison of BLEU with two other keyword-based methods. Because of the kind of evaluation performed, those with very few answers couldn't be evaluated with VSM.

- **VSM**, using a vectorial representation of the answers. In this case, we cannot implement it with reference texts, but by calculating similarities between answers. We have done a five-fold cross-evaluation, in which 20% of the candidate texts are taken as training set for calculating tf.idf weights for each term. The rest of the answers are assigned the score of the text in the training set which is most similar to it.

The results obtained are listed in Table 6. The improvement using BLEU is significant with 0.95 confidence.

4. Conclusions

We have described here an application of the BLEU algorithm for evaluating student answers with a shallow procedure. The main advantages of this approach are that

- It is very simple to program (just a few hours).
- It is language-independent, as the only processing done to the text is tokenisation.
- It can be integrated with other techniques and resources, such as thesauri, deep parsers etc., or it could be used in substitution of other keyword-based procedures in more complex systems.

On the other hand, as has sometimes been noted, BLEU is very dependent on the choice of the reference texts, so that leaves a high responsibility for the professors, who have to write them. Secondly, it is not suitable for all kinds of questions, such as those where the order of the sentences is important or, as we have seen, for an enumeration of advantages and disadvantages. Further processing would be necessary for scoring these questions.

We observe that, for evaluating student answers, a new brevity penalty that measures directly the recall of the student answer improves the results, and makes it less necessary to use higher-order N-grams.

Type of question

- The questions about advantages and disadvantages (data sets 6 and 8) have produced very low results (0.12 and 0.41 correlations), given that the algorithm is not capable of discerning whether the student is citing something as an advantage or a disadvantage, so the marks are equally high.
- The yes/no questions have the same problem, because the explanation might be the same regardless on the boolean answer. Thus, the correlation for 7 was 0.275. In the case of question 9, the correlation is very high because there are just a few answers and most of them are correct.
- Finally, the definitions and descriptions, which are usually very narrow questions, have produced better results, with correlations between 0.36 and 0.76.

Comparison with other algorithms As can be seen, BLEU clearly outperforms other keyword-based algorithms. Although it is not directly comparable to VSM, given that the evaluation procedure is different, the results hint that it has produced a better correlation.

Last remarks Therefore, we conclude that the current version of the system could be effectively used in an e-learning environment as a help to teachers who want to double check the scores they are giving, as well as for students who want to solve more practicals than the ones they receive in the classroom. Nevertheless, we do not recommend the use of this version as a stand-alone application.

The following are some ideas for future work:

- Automatise the production of the reference texts.
- Perform the evaluation against yet more existing systems.
- Explore how to extend the procedure with other linguistic processing modules, such as parsing or treatment of synonyms.

6. References

J. Burstein, C. Leacock, and R. Swartz. 2001. Automated evaluation of essay and short answers. In *M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.

D. Callear, J. Jerrams-Smith, and V. Soh. 2001. Caa of short non-mcq answers. In *Proceedings of the 5th International CAA conference*, Loughborough, UK.

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.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation.

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.