

# Application of the Bleu algorithm for recognising textual entailments

**Diana Pérez and Enrique Alfonseca**

Department of Computer Science

Universidad Autónoma de Madrid

Madrid, 28049, Spain

{diana.perez, enrique.alfonseca}@uam.es

## Abstract

The BLEU algorithm has been applied to many different fields. In this paper, we explore a new possible use: the automatic recognition of textual entailments. BLEU works at the lexical level by comparing a candidate text with several reference texts in order to calculate how close the candidate text is to the references. In this case, the candidate would be the text part of the entailment and the hypothesis would be the unique reference. The algorithm achieves an accuracy around 50% that proves that it can be used as a baseline for the task of recognising entailments.

## 1 Introduction

In the framework of the Pascal Challenges, we are now in the position of tackling a new application: the automatic recognition of textual entailments. It is, without doubt, a complex task that, as it is first approached in this event, needs both a preliminary study to find out which are the best techniques that can be applied, and the development of new techniques specifically designed for it. Another issue is to study if a combination of shallow techniques is able to face this problem, or whether it will be necessary to go into deeper techniques. If so, it will be interesting to know what the advantages of deep analyses are, and how the results differ from just using shallow techniques.

In the current situation, textual entailment is defined as the relation between two expressions, a text

(T), and something entailed by T, called an entailment hypothesis (H). Our approach consists in using the BLEU algorithm (Papineni et al., 2001), that works at the lexical level, to compare the entailing text (T) and the hypothesis (H). Next, the entailment will be judged as true or false according to BLEU's output.

Once the algorithm is applied, we have seen that the results confirm the use of BLEU as baseline for the automatic recognition of textual entailments. Furthermore, they show how a shallow technique can reach around a 50% of accuracy.

The article is organised as follows: Section 2 explains how BLEU works in general, next Section 3 details the application of this algorithm for recognising entailments and gives the results achieved using the development and test sets. Finally, Section 4 ends with a discussion about the contribution that BLEU can make to this task and as future work, how far it can be improved to increase its accuracy.

## 2 The BLEU Algorithm

The BLEU (BiLingual Evaluation Understudy) algorithm was created by (Papineni et al., 2001) as a procedure to rank systems according to how well they translate texts from one language to another. Basically, the algorithm looks for n-gram coincidences between a candidate text (the automatically produced translation) and a set of reference texts (the human-made translations).

The pseudocode of BLEU is as follows:

- For several values of N (typically from 1 to 4), calculate the percentage of n-grams from the

candidate translation which appears in any of the human translations. The frequency of each n-gram is limited to the maximum frequency with which it appears in any reference.

- Combine the marks obtained for each value of N, as a weighted linear average.
- Apply a brevity factor to penalise short candidate texts (which may have n-grams in common with the references, but may be incomplete). If the candidate is shorter than the references, this factor is calculated as the ratio between the length of the candidate text and the length of the reference which has the most similar length.

It can be seen from this pseudocode that BLEU is not only a keyword matching method between pairs of texts. It takes into account several other factors that make it more robust:

- It calculates the length of the text in comparison with the lengths of reference texts. This is because the candidate text should be similar to the reference texts (if the translation has been well done). Therefore, the fact that the candidate text is shorter than the reference texts is considered an indicative of a poor quality translation and thus, BLEU penalises it with a Brevity Penalty factor that lowers the score.
- The measure of similarity can be considered as a precision value that calculates how many of the n-grams from the candidate appear in the reference texts. This value has been modified, as the number of occurrences of an n-gram in the candidate text is clipped at the maximum number of occurrences it has in the reference texts. Therefore, an n-gram that is repeated very often in the candidate text will not increment the score if it only appears a few times in the references.
- The final score is the result of the weighted sum of the logarithms of the different values of the precision, for n varying from 1 to 4. It is not interesting to try higher values of n since coincidences longer than four-grams are very unusual.

BLEU's output is always a number between 0 and 1. This value indicates how similar the candidate and reference texts are. In fact, the closer the value is to 1, the more similar they are. (Papineni et al., 2001) report a correlation above 96% when comparing BLEU's scores with the human-made scores. This algorithm has also been applied to evaluate text summarisation systems (Lin and Hovy, 2003) and to help in the assessment of open-ended questions (Alfonseca and Pérez, 2004).

### 3 Application of BLEU for recognising textual entailments

For recognising entailments using BLEU, the first decision is to choose whether the candidate text should be considered as the text part of the entailment (T) or as the hypothesis (and, as a consequence whether the reference text should be considered as the H or the T part). In order to make this choice, we did a first experiment in which we considered the T part as the reference and the H as the candidate. This setting has the advantage that the T part is usually longer than the H part and thus the reference would contain more information than the candidate. It could help the BLEU's comparison process since the quality of the references is crucial and in this case, the number of them has been dramatically reduced to only one (when in the rest of the applications of BLEU the number of references is always higher).

Then, the algorithm was applied according to its pseudocode (see Section 2). The output of BLEU was taken as the confidence score and it was also used to give a TRUE or FALSE value to each entailment pair. We performed an optimisation procedure for the development set that chose the best threshold according to the percentage of success of correctly recognised entailments. The value obtained was 0.157. Thus, if the BLEU's output is higher than 0.157 the entailment is marked as TRUE, otherwise as FALSE.

The results achieved are gathered in Table 1. Besides, in order to confirm that this setting was truly better, we repeated the experiment this time choosing the T part of the entailment as the candidate and the H part as the reference. The results are shown in Table 2. In this case, the best threshold has been 0.1.

Task	NTE	A	NTR	NFR	NTW	NFW	Task	NTE	A	NTR	NFR	NTW	NFW
CD	98	77%	39	36	12	11	CD	98	72%	40	31	17	10
IE	70	44%	16	15	20	19	IE	70	50%	23	12	23	12
MT	54	52%	18	10	17	9	MT	54	52%	21	7	20	6
QA	90	41%	9	28	17	36	QA	90	50%	22	23	22	23
RC	103	51%	30	23	28	22	RC	103	50%	33	19	32	19
PP	82	57%	22	25	18	17	PP	82	60%	25	24	19	14
IR	70	44%	10	21	14	25	IR	70	41%	8	21	14	27
Total	567	53%	144	158	126	139	Total	567	54%	172	137	147	111

Table 1: Results for the development sets considering the T part of the entailment as the reference text (threshold = 0.157). Columns indicate: task id; number of entailments (NTE); accuracy (A); number of entailments correctly judged as true (NTR); number of entailments correctly judged as false (NFR); number of entailments incorrectly judged as true (NTW); and, number of entailments incorrectly judged as false (NFW).

This is the value that has been fixed as threshold for the test set.

It is important to highlight that the average correlation achieved was 54%. Moreover, it reached a 72% accuracy for the Comparable Document (CD) task as it could be expected since Bleu’s strength relies on making comparisons among texts in which the lexical level is the most important. For example, the snippet of the development test with identifier 583, whose T part is “*While civilians ran for cover or fled to the countryside, Russian forces were seen edging their artillery guns closer to Grozny, and Chechen fighters were offering little resistance*” and H part is “*Grozny is the capital of Chechnya*”, is an ideal example case for BLEU. This is because only the word Grozny is present both in the T and H texts. BLEU will mark it as false since there is no n-gram co-occurrence between both texts.

On the other hand, BLEU cannot deal examples in which the crucial point to correctly recognise the entailment is at the syntactical or semantics level. For example, those cases in which the T and H parts are the same except for just one word that reverses the whole meaning of the text. For example, the snippet 148 of the development set, whose T part is “*The Philippine Stock Exchange Composite Index rose 0.1 percent to 1573.65*” and the H part is

Table 2: Results for the development sets considering the T part of the entailment as the candidate text (threshold = 0.1). Columns indicate: task id; number of entailments (NTE); accuracy (A); number of entailments correctly judged as true (NTR); number of entailments correctly judged as false (NFR); number of entailments incorrectly judged as true (NTW); and, number of entailments incorrectly judged as false (NFW).

“*The Philippine Stock Exchange Composite Index dropped.*” This is a very difficult case for BLEU. It will be misleading since BLEU would consider that both T and H are saying something very similar, while in fact, the only words that are different in both texts, “*rose*” and “*dropped*”, are antonyms, making the entailment FALSE.

It can also be seen how the results contradict the insight that the best setting would be to have the T part as the reference text. In fact, the results are not so much different for both configurations. A possible reason for this could be that all cases when BLEU was misled into believing that the entailment was true (because the T and H parts have many n-grams in common except the one that is the crucial to solve the entailment) are still problematic. It should be noticed that BLEU, irrespectively of the consideration of the texts as T or H, cannot deal with these cases.

The results for the test set confirm the same conclusions drawn for the development tests. In fact, for the first run in which BLEU was used for all the tasks, it achieved a 52% confidence-weighted score and a 50% accuracy. See Table 3 for details.

As can be seen, not only the overall performance continues being similar to accuracy obtained with the development test. Also the best task for the test set keeps being the CD. To highlight this fact, we

TASK	CWS	A
CD	0.7823	0.7000
IE	0.5334	0.5000
MT	0.2851	0.3750
QA	0.3296	0.4231
RC	0.4444	0.4571
PP	0.6023	0.4600
IR	0.4804	0.4889
TOTAL	0.5168	0.4950

Table 3: Results for the test set (threshold = 0.1). Columns indicate: task id; confidence-weighted score or average precision (CWS); and, the accuracy (A).

implemented a preliminary step of the algorithm in which there was a filter for the CD snippets, and only they were processed by BLEU. In this way, we created a second run with the CD set that achieved a CWS of 78% and a 70% accuracy. This high result indicates that, although, in general, BLEU should only be considered as a baseline for recognising textual entailments, in the case of CD, it can be used as a stand-alone system.

#### 4 Conclusion and future work

Some conclusions can be drawn from the experiments previously described:

- BLEU can be used as a baseline for the task of recognising entailments, considering the candidate text as T and the reference text as the H part of the entailment, since it has achieved an accuracy above 50%.
- BLEU’s results depend greatly on the task considered. For example, for the Comparable Documents (CD) task it reaches its maximum value (77%) and for Information Retrieval (IR) the lowest (41%).
- BLEU has a slight tendency to consider a hypothesis as TRUE. In 319 out of 567 pairs, BLEU said the entailment was true. Out of these, it was right in 172 cases, and it was wrong in 147 cases. On the other hand, there were only 111 false negatives.

It is also interesting to observe that, although the origin of BLEU is to evaluate MT systems, the results for the MT task are not specially higher. The reason for that could be that BLEU is not being used here to compare a human-made translation to a computer-made translation, but two different sentences which contain an entailment expression, but which are not alternative translations of the same text in a different language.

The main limit of BLEU is that it does not use any semantic information and, thus, sentences with many words in common but with a different meaning will not be correctly judged. For instance, if T is “*The German officer killed the English student*” and H is “*The English students killed the German Officer*”, BLEU will consider the entailment hypothesis as TRUE, while it is FALSE.

It would be interesting, as future work, to complement the use of BLEU with some kind of syntactic processing and some treatment of synonyms and antonyms. For example, if BLEU were combined with a parser translating all sentences from passive to active and allowed the comparison by syntactic categories such as subject, direct object, indirect object, etc., it would be able to recognise more entailments.

#### References

- E. Alfonseca and D. Pérez. 2004. Automatic assessment of short questions with a BLEU-inspired algorithm and shallow nlp. In *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 25–35. Springer Verlag.
- C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003)*.
- K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Research report, IBM.