Advances in Intelligent Systems Research, volume 172
Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)

MBRTE: Modified BLEU Algorithm for Recognition Textual Entailment

Abdiansah ABDIANSAH[1*] and Alvi Syahrini UTAMI[2]

[1,2]*AIRLab, Comp. Sci. Dept., Universitas Sriwijaya, Indonesia*
**Corresponding author: abdiansah@unsri.ac.id*

Abstract
The BLEU algorithm has become a standard evaluation tool for Machine Translation (e.g Goole Translate). The advantages of this algorithm are simple and fast. This algorithm is quite widely applied in fields other than MT, including in the Recognition Textual Entailment (RTE). The implementation of BLEU algorithm to the RTE requires some modification because of several problems. One of these problems is related to the sensitivity of BLEU scores. This research has modified BLEU to overcome these problems. Experimental data set contains 200 hypotheses with each hypothesis have some text. The experimental results obtained an accuracy of 83,82 % for modified BLEU. While the accuracy results using BLEU without modification is obtained at 77,50%. So that the accuracy increases of 6,32%.

Keywords: *MBRTE, BLEU, textual entailment*

Introduction

RTE (Recognition Textual Entailment) is part of Natural Language Understanding (NLU). RTE aims to solve variability problems that are often encountered in natural language-based systems, such as Question Answering System (QAS), Machine Translation (MT), Automatic Summarization (AS) and others [1]. RTE task is to recognize the semantic similarity between two texts. If the meaning of a text (Text or T) can be infered by another text (Hypothesis or H) then the two texts are considered have entailment. Generally, the text similarity techniques used in RTE consist of lexical similarity and semantic similarity. Both these techniques are multilevel, ranging from simple to complex one. One of the lexical similarity technique and used as a research baseline in RTE is BLEU algorithm [2].

BLEU algorithm [3] becomes "de facto" standard that is used as an automatic measurement tool in the Machine Translation (MT). There are two reasong why BLEU is made as standard metric for MT, namely: simple and fast [4]. Nevertheless, BLEU has several limitations, especially if applied to other fields such as Recognition Textual Entailment (RTE) field. In principle, BLEU is specifically designed to handle long sentences according to domain of MT. Even, a particular condition is made to overcome the problem if the length of H is shorter than the T. In fact, the length of H in RTE is generally always short from T. Therefore, BLEU must be modified so that it can be used in the domain of RTE. Modified BLEU can be done in two ways, namely: (1) Modification in the data processing phases that aims to improve data quality; and (2) Modification of BLEU phases that aims to improve BLEU's performance. In this study, we used the first way because the data in dataset used is retrieved from Web. So that the data quality is still very low (semi-structured and noisy) and need to improved.

Furthermore, this article is organized as follows. Section 2 contains related researches in this study. Section 3 explains brief of BLEU algorithm. Section 4 contains the research methodology used. Section 5 contains discussion and experimental results. The Last section contains conclusion.

Related Works

The implementation of BLEU algorithm in the field other than MT has been done quite a lot. One of them is RTE field. The first work of BLEU for RTE was was done by [6] in the First Pascal RTE Challange in 2005 [2]. They implemented BLEU algorithm without modification. There are two focuses activity in they works, namely: (1) Determine the conversion of reference text as T or H. This applies also for candidate text; and (2) Finding thresholding value to determine whether the text pairs is TRUE (entailment) or FALSE (not entailment). The average accuracy results in the first experiment (T as reference and H as candidate) were obtained at 53%. The threshold value used is 0.157. Whereas in the second experiment obtained the average accuracy of 54% (T ascandidate and H as reference). The threshold value used is 0.1. The accuracy of all experiments results obtained above

50%. So that BLEU algorithm can be used as a baseline for RTE. Also there are still many opportunities to improve the accuracy.

Next, [2] modified one of the BLEU algorithm phase. They used the scheme T as reference and H as candidate since in RTE definition H is only expected to contain a subset of T. There are two parts of modifications they applied the BLEU algorithm, namely: (1) Eliminating Bravety Penalty (BP) since in RTE length of H is always shorten than T; and (2) Subtitute the average n-gram score from geometric to linear. They use RTE-1 (Development Set) dataset as experimental data. The result of BLEU algorithm accuracy without modification was obtained at 53.8 %. While the results of modified BLEU was obtained at 57.8%.

BLEU ALGORITHM

The BLEU (Bilingual Evaluation Understudy) algorithm aims to evaluate the output of Machine Translation (MT). The basic idea of the algorithm is to compare the result of MT (as candidat) with one or more human translations (as references). The comparison is calculated using average n-gram for each candidat sentence and references. The output is a BLEU score between 0 – 1. If the score is getting closer to 1 then the text pairs are getting similar. Following is stages of BLEU algorithm:
For each i up to N, calculate a score Si that is the ratio of the count of i-gram co-appearing in both candidate and references ($C_{cand,refs}$) and the count of i-gram appearing in the candidate (C_{cand}).

$$Si = Ccand,refs / Ccand \quad \dots (1)$$

Average the values of Si. This is accomplished with a weighted geometric mean. The weight wi is typically kept constant for all i (wi=1/N for all i).

$$SN = e^{(SIGN (wi*log(Si)))} \dots (2)$$

Calculate the brevity penalty. If the length of the candidate (c) is greater than the length of the reference (r), then there is no penalty (b = 1). Otherwise, the penalty is logarithmically derived from the ratio of the two lengths:

$$b = e^{(1-(r/c))} \text{ if } t<r; 1 \text{ if } t>r \dots (3)$$

Finally, calculate the overall score (BLEU Score) as the mean of all scores multiplied by the brevity penalty.

BLEUScore = SN * b … (4)
then there is no penalty (b = 1). Otherwise, the penalty is logarithmically derived from the ratio of the two lengths:

$$b = e^{(1-(r/c))} \text{ if } t<r; 1 \text{ if } t>r \dots (3)$$

Finally, calculate the overall score (BLEU Score) as the mean of all scores multiplied by the brevity penalty.

BLEUScore = SN * b … (4)

RESEARCH BACKGROUND

This study uses DS-200-R dataset derived from [5]. The dataset contains 200 hypotheses with each hypothesis have several references. Furthermore, the experiment was conducted in two ways, namely: (1) Using BLEU without modification; and (2) Using Modified BLEU (MBRTE). If the BLEU score is above 0.5 (threshold) then the value is TRUE (entailment), otherwise it will be considered FALSE (not entailment). The results of study were measured by accuracy using Eq. (5).

accuracy = Total of hypotesis (BLEU score >= 0.5) / Total of hypotesis … (5)

RESULTS AND DISCUSSION

Based on analysis of experiment results of BLEU algorithm trial, the information was obtained that the score produced by BLEU was very sensitive to the difference in length of sentence. The difference and the addition of one word between H and T can cause a considerable difference in scores. Therefore, to see the changes of BLEU score due to the difference in length of sentence, an experiment was conducted by calculating several variants of T patterns against H. Table 1 shows the results of experiment. The length of T will be longer than H because the addition of another word (denoted as X). In the table also we seen that T2 has two patterns that produce the same score. The pattern of T3 has “X” which is more than pattern of T2 pattern, so that the BLEU score decreases. Furthermore, to see the various other patterns, then we made a list of possibilities patterns for all T. The results can be seen in Table 2. If there is a sentence “A B C D” then five types of patterns will be generated. While the number of adding X is symbolized by Xn.

Table 1. BLEU score comparison based on sentence patterns.

Hypothesis/Text	Sentence Patterns	BLEU Score
H	{ABCD}	1
T_1	{ABCD}	1
T_2	{A B C D X} or {X A B C D}	0,77
T_3	{A B C D X X} or {X X A B C D}	0,60
T_4	{A B C X D} or {A X B C D}	0,59
T_5	{A B C X X D} or {A X X B C D}	0,46

Table 2. List of possibility patterns from adding "X" words.

Pattern Types	Sentence Patterns
Pattern-1	{A B C D Xn} or {Xn A B C D X}
Pattern-2	{A B C Xn D} or {A Xn B C D}
Pattern-3	{A B Xn C D}
Pattern-4	{A B Xn C Xn D} or {A Xn B Xn C D}
Pattern-5	{A Xn B Xn C Xn D}

Table 3 contains the results of BLEU scores from the patterns in the Table 2. The value of textit n in the table is the number of adding "X". If n = 0, then there is no addition of "X" so the score is

While in Figure 1 can be seen the comparison graph of BLEU score for all patterns. In the graph can be seen that the more X increases, the more BLEU score decreases. In addition, the decline is quite drastic. For example in Pattern-4 and Pattern-5, adding of "X" can reduce almost half of score, that is from 1 to 0.4

Table 3. BLEU Score for all patterns in Table 2 with adding n-word of "X".

No	Pattern-1	Pattern-2	Pattern-3	Pattern-4	Pattern-5
0	1	1	1	1	1
1	0,77	0,59	0,70	0,46	0,47
2	0,60	0,46	0,54	0,28	0,22
3	0,47	0,35	0,42	0,17	0,10
4	0,36	0,28	0,33	0,10	0,05

The analysis results of RTE dataset [5] found that quite a number of T were same patterned on patterns in Table 2. These discoveries became our motivation for reducing the length of T with the assumption that, "The smaller of difference in length of T and H, the higher of BLEU score". The sentence reduction in this study uses word removal techniques. The basic idea is to remove several words in T so that the length of T becomes shorter or close to H. The sentence reduction algorithm can be seen in Algorithm-1.

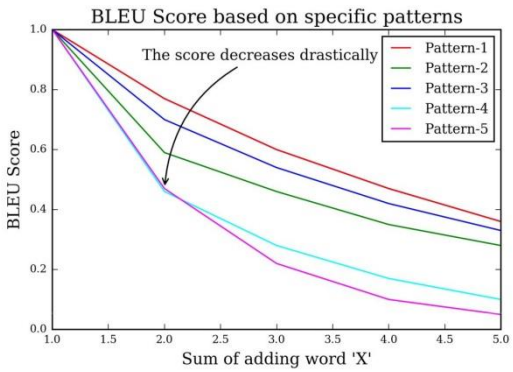

Figure 1. BLEU Score comparison between all patterns

Table 4. The experimental results using dataset DS-200-R.

Dataset	Accuracy (%)
Original BLEU	77,5
Modified BLEU (MBRTE)	83,82

Algorithm-1

```
H ← hypothesis
T ← references
T new ← new references
idx f irst ← 0
idx last ← 0
idx ← 0
for the first word of T to the last word of T do
    if word of T in H then
        if idx f irst < idx last then
            idx f irst ← idx last
        end if
    if idx f irst = idx last then
        idx last ← idx
    end if
    temp = T [idx f irst : idx last + 1]
    if length(temp) = 1, 2, or 3 then
        T new.append(temp[length(temp) – 1]
    else
        for i to length(temp) – 1 do
            if i mod 2 != 0 then
                T new.append(temp[i + 1])
            end if
        end for
        if length(temp) mod 2 = 0 then
            T baru.append(temp[length(temp) – 1])
        end if
    end if
    end if
    idx ← idx + 1
end for
return T new
```

CONCLUSION

This article explainsis about modified BLEU algorithm for RTE or called MBRTE. BLEU algorithm was a standard evaluation tool in Machine Translation. The algorithm operates at level of lexical similarity. The advantages of BLEU are simple and fast. Nevertheless, there are some roblems if it is applied to the other fields such as RTE. One of the problems examined in this study is the changes of BLEU score which is quite drastic when there are additional words in the sentence. The proposed solution is to reduce the reference text (Text/T) so that the length of reference is shorter or close to the candidate text (hypotesis/H). The experimental results show that the accuracy increases about 6.32% when using MBRTE, which is from 77.5% to 83.82%. These results are quite promising compared to the previous study [6] [2] even though the dataset used is different.

REFERENCES

[1] Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M. 2013. Recognizing textual entailment: Models and applications. Synthesis Lectures on Human Language Technologies, 6(4), 1-220. Morgan & Claypool Publisher.

[2] Terrence Szymanski, https://www.researchgate.net/publication/266017019

[3] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.

[4] Song, X., Cohn, T., & Specia, L. (2013). BLEU deconstructed: Designing a better MT evaluation metric. International Journal of Computational Linguistics and Applications, 4(2), 29-44.

[5] Abdiansah A., Azhari A., & Anny K. S. 2018. WERTES: Web as External Resources for Textual Entailment Systems. International Journal of Intelligent Engineering and Systems. Vol.11, no.3, pp.91.

[6] Prez, D., & Alfonseca, E. (2005). Application of the Bleu algorithm for recognising textual entailments. In Proceedings of the First Challenge Workshop Recognising Textual Entailment (pp. 9-12