WupLeBleu: The WordNet-Based Evaluation Metric for Machine Translation

Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
{debajyoty.pcs13,asif,pb}@iitp.ac.in

Abstract

In this paper, we propose a machine translation (MT) evaluation metric based on paraphrase matching fuzzy logic and the n-gram feature. Paraphrase matching generally calculates the relatedness between two strings by considering the depth, content, and structure in WordNet taxonomy. Various metrics based on stem match exist for MT evaluation. Since a sentence can be represented in different forms using synonyms and morphological structures, stem match is found inadequate to evaluate the MT output. Our proposed WupLeBleu evaluation metric can handle this challenge. Empirical evaluation on the benchmark datasets show that our proposed metric significantly improves the correlations with respect to the human judgment.

1 Introduction

The usage of automatic evaluation metric aims at evaluating the output quality of machine translation (MT) systems quickly. This is less expensive in comparison to the evaluation carried out by the trained experts. A few techniques were proposed for automatic evaluation. Especially, BLEU metric (Papineni et al., 2002) was widely used to automatically evaluate the quality of machine translation output. BLEU is an n-gram based method. However, weakness of BLEU was addressed in recent years (Ananthakrishnan et al., 2007). Many other automatic MT evaluation metrics like LeBleu (Virpioja and Grönroos, 2015), METEOR (Lavie and Denkowski, 2009), NIST (Doddington, 2002) etc. were proposed to overcome the issues of BLEU. To a huge degree, proposed WupLeBleu works on the principle of the fuzzy matching logic of the n-gram words along with the Wu-Palmer (WUP) similarity (Wu and Palmer, 1994) using WordNet. WUP similarity computes semantic relatedness of word senses using the edge counting method (Wu and Palmer, 1994).

We present an analysis of WupLeBleu with various language pairs: Chinese-English, Turkish-English, Czech-English, Russian-English, Finnish-English, and German-English.

1.1 Related Works

The most popular automatic MT evaluation metric is BLEU that computes n-gram matchings of the candidate (C) with reference (R) translation. It computes the overall precision of n-grams by using geometric average along with the brevity penalty. But, there are many issues with the automatic evaluation of BLEU metric, as it solely focuses on the n-gram matchings. Researchers proposed NIST (Lin and Hovy, 2003) to calculate the score based on the information gain from each n-gram. NIST evaluation assigns more score to the n-gram which is more informative.METEOR (Lavie and Denkowski, 2009) is based on explicit word-to-word matchings using the stem, and synonym modules. RIBES was proposed (Neubig et al., 2012) with the primary focus on the word order of a sentence and by considering the brevity penalty for calculating the final score with the help of Kendall's correlation. Very recently, researchers introduced LeBleu (Virpioja and Grönroos, 2015) that considers fuzzy based matching and computes the similarity score based on Levenshtein distance. LeBleu uses arithmetic averaging for calculating the overall precision of score. LeBleu cannot handle paraphrase or synonym. This is regarded as one of the major drawbacks. The proposed WupLeBleu is designed in such a way so that it can properly handle all of the challenges like synonym matching, fuzzy matching and morphological differences altogether.

2 Issues in Existing Machine Translation Evaluation Metrics

Despite the fact that the BLEU is widely used metric for MT evaluation, it experiences a few shortcomings which we particularly intend to address in our proposed metric.

1. BLEU, a precision based metric that matches word n-grams of MT-translation output with multiple reference translations simultaneously. Lack of attention to recall within BLEU is a great shortcoming. The "Brevity Penalty" in the BLEU metric does not satisfactorily compensate for the absence of recall.

2. The n-gram matching focuses exact word matches and all the matched words weigh equally in BLEU. The geometric average of n-gram scores produces a result of zero if the individual n-gram scores are zero.

3. The correlation between BLEU score and human evaluation is very poor (Ananthakrishnan et al., 2007).

For example, let us consider the candidate and reference translations as stated below:
C: *H*e who fears as a result of conquered is a sound of defeat.
R: *H*e who fears being conquered is sure of defeat.
Here, R and C refer to reference and candidate translation of phrase-based statistical machine translation (PBSMT) system, respectively. The computed BLEU score will be zero for C, because of the absence of the four-gram matchings in C1 when checked against the reference translations.
C: *T*he 7^{th} era are as yet battling for their rights.
R: *T*he seventh generation is still fighting for their rights.
For example, both BLEU and LeBleu fail as the n-gram matchings are absent. METEOR, which considers only the precision of uni-gram matchings calculates the score based on explicit word-to-word matching. The default METEOR parameters prefer longer translations than the other metrics. Since precision and recall are computed for uni-gram matching, the high α values contribute more weight to uni-gram recall than precision. This puts METEOR in disadvantage position when being evaluated by the other metrics. The primary objective of our proposed WupLeBleu metric is to overcome the problems as mentioned above. Consider the following example. Here, both candidate and reference translations convey the same meaning, but with different vocabularies.

C: हर स्थान शांति छा गया।
ETL: *h*ar sthaan shaanti chha gaya.
ET: *E*very place has peace.
R: हर जगह सन्नाटा छा गया।
ETL: *h*ar jagah sannaata chha gaya.
ET: *S*ilence everywhere.
Here, ETL, ET are the English transliteration and English translation, respectively. But, the computed BLEU score would be zero, as exact n-gram matchings are absent. Also, LeBleu partially solves this problem by using the fuzzy matching technique. But, WupLeBleu metric has the power of solving a fuzzy n-gram matching technique along with the WUP similarity.

3 Methodology

WupLeBleu calculates the score based on the precision of n-gram matching with fuzzy logic along with the WUP similarity score. The WUP similarity score uses WordNet to improve the correlation of automatic evaluation metric with human evaluation. This score provides the detailed idea of candidate words with respect to reference words in terms of synonyms and lemmas. The WUP similarity method (Wu and Palmer, 1994) generally calculates the relatedness between the two words by considering the depth, content and structure of two strings in WordNet taxonomies. The similarity measure is computed based on the ratio of the information content of the least common subsumer of the candidate and the reference string. LCH (Leacock and Chodorow, 1998), the WUP similarity (Wu and Palmer, 1994) and the path length are three similarity measures that are considered based on the path length (Pedersen et al., 2004) between C and R sentences. LCH method calculates the minimum path between the source and the target string, and then scales the minimum path by the maximum path length found in the hierarchy in which they occur. The WUP similarity score is calculated as the sum of the depth of LCS (Least Common Subsumer) between the words from C and R sentences. The path score is equal to the inverse of the shortest path between two strings (Pedersen et al., 2004). The final WUP similarity score is calculated based on the above three measures.

$$WUP\,similarity = 2 * \frac{depth(lcs)}{(depth(s1) + depth(s2))}$$

If the WUP similarity score is more than the predefined threshold parameter δ, then both the words are considered to be nearly similar and their matching n-gram precision is taken into account while calculating the overall n-gram precision, else ignored. Fuzzy matching works on the fact, that the n-gram matching is said to be a fuzzy match if the similarity score is more than the threshold parameter ∂. The fuzzy based similarity score is calculated as one minus letter edit distance. The letter edit distance (levenshtein distance) is a measure of the similarity between two strings (Heeringa, 2004). The distance ($lev_{a,b}(i,j)$) is calculated by the required number of insertions, deletions, or substitutions, to transform a source into target string.

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & if\ min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & otherwise \end{cases}$$

The brevity penalty (BP) considers the total number of characters rather than the words present in reference and candidate translations. Overall precision is calculated by combining the individual n-gram precisions through arithmetic averaging.

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

Here, the variables r and c refer to the total number of characters in the reference and candidate translations respectively.

4 Experiments

We conduct the experiments on WMT 14 dataset (Machacek and Bojar, 2014) and HinEnCorp (Bojar et al., 2014) for five different language pairs. At first, the WUP similarity scores are calculated among the words of the aligned sentence. If this WUP similarity score is more than the tuned threshold parameter δ, then two words are considered as a matching. After fine tuning we found δ value as 0.80. If there is a matching of more than one single word pair, then the word pair with greater WUP similarity value will be chosen as matching. All synonyms, morphological structure, and other representation of the words are covered by using this step. It then calculates the similarity score which uses fuzzy logic. Basically, it calculates the Levenshtein distance between the two strings. TH final score is then computed by calculating the arithmetic average of the individual n-gram matchings multiplied by brevity penalty (BP).

5 Performance in WMT 14 Dataset

We also evaluate our proposed algorithm using WMT 2014 dataset [1]. The highest correlation with human judgment is found for Hindi to English (hi-en) and French to English (fr-en) language pairs. After calculating the correlation with human judgment on average, we found the score as 0.951. This shows that our proposed model stands out on top, considering the average score. In most of the cases the proposed metric achieves better correlation than the standard metrics, shown in Table 1.
Main challenge in WUP similarity approach is to tune δ. If this value is too small then synonymous words may not be considered as similar words. For large value of δ may cover distance words as similar one. For example, this

[1] http://www.statmt.org/wmt14/

Table 1: Comparison: Correlation with different metrics in WMT 14 Dataset

Metric	Pearson Correlation					
	de-en	ru-en	cs-en	fr-en	hi-en	Average
WupLeBleu	**0.931**	**0.882**	**0.985**	**0.973**	**0.984**	**0.951**
LeBleu	0.892	0.896	0.912	0.971	0.969	0.928
LAYRED	0.893	0.843	0.940	0.973	0.976	0.925
BLEU	0.831	0.774	0.908	0.952	0.956	0.884
NIST	0.810	0.785	0.983	0.955	0.783	0.863
METEOR	0.926	0.792	0.980	0.975	0.457	0.826
TER	0.774	0.796	0.977	0.952	0.618	0.823

metric may identify ”foot-ball” and ”basketball” as similar words which is not true.
We have done significance tests, and observe that results are significant with 95% confidence level (with p=0.1 which is $<$ 0.05).

6 Evaluation with other Datasets

We evaluate the WupLeBleu for English to Hindi (en-hi) translation. Due to the unavailability of en-hi language in WMT dataset we study the proposed evaluation score by using miscellaneous domain data sets from the HinEnCorpora. We choose three systems: Moses's default configuration for SMT system[2], Google[3] and Bing[4] translator) for the correctness checking of our proposed metrics. We take 271877 and 1001 sentence pairs for training and tuning of SMT, respectively. For evaluation we use 1002 sentence pairs. After detailed analysis (with 1002 sentence pair), we achieve better Pearson correlation for the proposed WupLeBleu. The Pearson correlations are BLEU: 0.9103, METEOR: 0.9137, Lebleu: 0.9278 and WupLeBleu: 0.9434.

We also manually evaluate the F-beta scores (Figure 1) for different automatic evaluation metrics and compare their ratio (Figure 2) to estimate how close these are to human evaluation. It is clearly understood from Figure 2) that LeBleu and proposed WupLeBleu's evaluation preferences are closer to manual judgment. WUP similarity makes WupLeBleu better. We have added details of manual evaluation in the additional sheet.

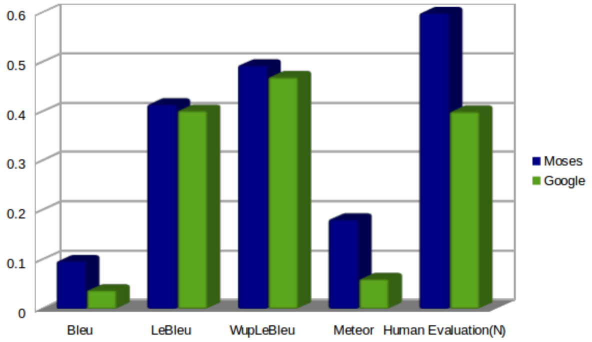

Figure 1: Ranking of correctness (hi-en)

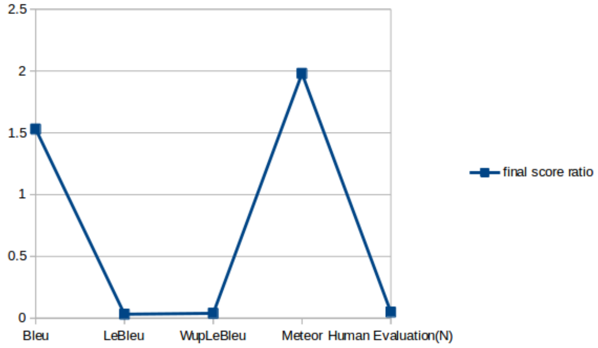

Figure 2: Score ratio between dataset-1 and dataset-2

7 Conclusion

In this paper we have proposed an automatic MT evaluation metric, WupLeBleu. Based on a large study and several experiments, we can conclude that fuzzy logic based n-gram matching with the WUP similarity method can perform more accurate MT evaluation than the existing metrics. We believe that our proposed approach that uses WUP similarity and fuzzy logic has a higher similarity to human evaluation. In future we will also evaluate the WupLeBleu metric on the other language pairs.

[2] http://www.statmt.org/moses/manual/manual.pdf
[3] https://translate.google.com/
[4] https://www.bing.com/translator

References

R Ananthakrishnan, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2007. Some issues in automatic evaluation of english-hindi mt: more blues for bleu. *ICON*.

Ondrej Bojar, Vojtech Diatka, Pavel Rychlỳ, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. 2014. Hindencorp-hindi-english and hindi-only corpus for machine translation. In *LREC*, pages 3550–3555.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Wilbert Jan Heeringa. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University Library Groningen][Host].

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram cooccurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.

Matous Machacek and Ondrej Bojar. 2014. Results of the wmt14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Sami Virpioja and Stig-Arne Grönroos. 2015. Lebleu: N-gram-based translation evaluation score for morphologically complex languages. In *WMT@ EMNLP*, pages 411–416.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.