Tolerant BLEU: a Submission to the WMT14 Metrics Task

Jindřich Libovický and **Pavel Pecina**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal ad Applied Linguistics
{libovicky, pecina}@ufal.mff.cuni.cz

Abstract

This paper describes a machine translation metric submitted to the WMT14 Metrics Task. It is a simple modification of the standard BLEU metric using a monolingual alignment of reference and test sentences. The alignment is computed as a minimum weighted maximum bipartite matching of the translated and the reference sentence words with respect to the relative edit distance of the word prefixes and suffixes. The aligned words are included in the n-gram precision computation with a penalty proportional to the matching distance. The proposed tBLEU metric is designed to be more tolerant to errors in inflection, which usually does not effect the understandability of a sentence, and therefore be more suitable for measuring quality of translation into morphologically richer languages.

1 Introduction

Automatic evaluation of machine translation (MT) quality is an important part of the machine translation pipeline. The possibility to run an evaluation algorithm many times while training a system enables the system to be optimized with respect to such a metric (e.g., by Minimum Error Rate Training (Och, 2003)). By achieving a high correlation of the metric with human judgment, we expect the system performance to be optimized also with respect to the human perception of translation quality.

In this paper, we propose an MT metric called tBLEU (tolerant BLEU) that is based on the standard BLEU (Papineni et al., 2002) and designed to suit better when translation into morphologically richer languages. We aim to have a simple language independent metric that correlates with human judgment better than the standard BLEU.

Several metrics try to address this problem as well and usually succeed to gain a higher correlation with human judgment (e.g. METEOR (Denkowski and Lavie, 2011), TerrorCat (Fishel et al., 2012)). However, they usually use some language-dependent tools and resources (METEOR uses stemmer and parahprasing tables, TerrorCat uses lemmatization and needs training data for each language pair) which prevent them from being widely adopted.

In the next section, the previous work is briefly summarized. Section 3 describes the metric in detail. The experiments with the metric are described in Section 4 and their results are summarized in Section 5.

2 Previous Work

BLEU (Papineni et al., 2002) is an established and the most widely used automatic metric for evaluation of MT quality. It is computed as a harmonic mean of the n-gram precisions multiplied by the brevity penalty coefficient which ensures also high recall. Formally:

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right),$$

where BP is the brevity penaly defined as follows:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{otherwise} \end{cases},$$

c is the length of the test sentence (number of tokens), r is the length of the reference sentence, and p_n is the proportion of n-grams from the test sentence found in the reference translations.

The original experiments with the English to Chinese translation (Papineni et al., 2002) reported very high correlation of BLEU with human judgments. However, these scores were computed using multiple reference translations (to capture translation variability) but in practice, only one

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 409–413,
Baltimore, Maryland USA, June 26–27, 2014.

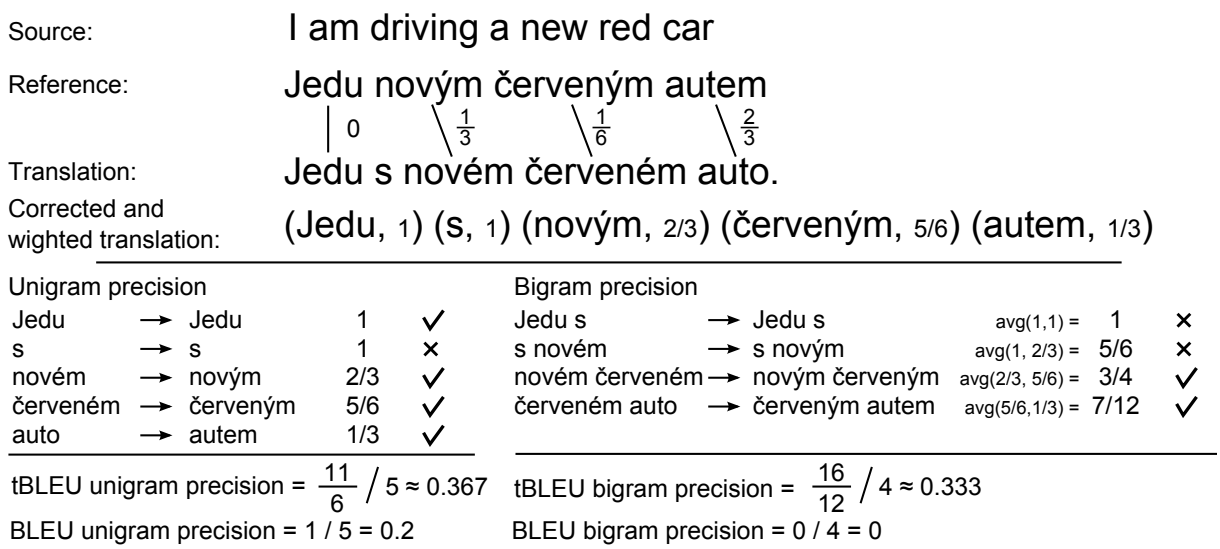

Figure 1: An example of the unigram and bigram precision computation for translation from English to Czech with the test sentence having minor inflection errors and an additional preposition. The first two lines contain the source sentence in English and a correct reference translation in Czech. On the third line, there is an incorrectly translated sentence with errors in inflection. Between the second and the third line, the matching with respect to the affix distance is shown. The fourth line contains the corrected test sentence with the words weights. The bottom part of the figure shows computation of the unigram and bigram precisions. The first column contains the original translation n-grams, the second one the corrected n-grams, the third one the n-gram weights and the last one indicates whether a matching n-gram is contained in the reference sentence.

reference translation is usually available and therefore the BLEU scores are often underestimated.

The main disadvantage of BLEU is the fact that it treats words as atomic units and does not allow any partial matches. Therefore, words which are inflectional variants of each other are treated as completely different words although their meaning is similar (e.g. *work*, *works*, *worked*, *working*). Further, the n-gram precision for $n > 1$ penalizes difference in word order between the reference and the test sentences even though in languages with free word order both sentences can be correct (Bojar et al., 2010; Condon et al., 2009).

There are also other widely recognized MT evaluation metrics: The NIST score (Doddington, 2002) is also an n-gram based metric, but in addition it reflects how informative particular n-grams are. A metric that achieves a very high correlation with human judgment is METEOR (Denkowski and Lavie, 2011). It creates a monolingual alignment using language dependent tools as stemmers and synonyms dictionaries and computes weighted harmonic mean of precision and recall based on the matching.

Some metrics are based on measuring the edit distance between the reference and test sentences. The Position-Independent Error Rate (PER) (Leusch et al., 2003) is computed as a length-normalized edit distance of sentences treated as bags of words. The Translation Edit Rate (TER) (Snover et al., 2006) is a number of edit operation needed to change the test sentence to the most similar reference sentence. In this case, the allowed editing operations are insertions, deletions and substitutions and also shifting words within a sentence.

A different approach is used in TerrorCat (Fishel et al., 2012). It uses frequencies of automatically obtained translation error categories as base for machine-learned pairwise comparison of translation hypotheses.

In the Workshop of Machine Translation (WMT) Metrics Task, several new MT metrics compete annually (Macháček and Bojar, 2013). In the comptetition, METEOR and TerrorCat scored better that the other mentioned metrics.

3 Metric Description

tBLEU is computed in in two steps. Similarly to the METEOR score, we first make a monolingual alignment between the reference and the test sentences and then apply an algorithm similar to the standard BLEU but with modified n-gram precisions.

The monolingual alignment is computed as a minimum weighted maximum bipartite matching between words in a reference sentence and a translation sentence[1] using the Munkres assignment algorithm (Munkres, 1957).

We define a weight of an alignment link as the *affix distance* of the test sentence word w_i^t and the reference sentence word w_j^r: Let S be the longest common substring of w_i^t and w_i^r. We can rewrite the strings as a concatenation of a prefix, the common substring and a suffix:

$$w^t = w^t_{i,p} S w^t_{i,s}$$
$$w^r = w^r_{j,p} S w^r_{j,s}$$

Further, we define the affix distance as:

$$\mathrm{AD}(w^r, w^t) = \max\left\{1, \frac{L(w^r_{j,p}, w^t_{i,p}) + L(w^r_{s,j}, w^t_{s,i})}{|S|}\right\}$$

if $|S| > 0$ and $\mathrm{AD}(w^r, w^t) = 1$ otherwise. L is the Levensthein distance between two strings.

For example the affix distance of two Czech words *vzpomenou* and *zapomenout* (different forms of verbs remember and forget) is computed in the following way: The longest common substring is *pomenou* which has a length of 7. The prefixes are *vz* and *za* and their edit distance is 2. The suffixes are an empty string and *t* which with the edit distance 1. The total edit distance of prefixes and suffixes is 3. By dividing the total edit distance by the length of the longest common substring, we get the affix distance $\frac{3}{7} \approx 0.43$.

We denote the resulting set of matching pairs of words as $M = \{(w_i^r, w_i^t)\}_{i=1}^m$ and for each test sentence $S^t = (w_1^t, ..., w_m^t)$ we create a corrected sentence $\hat{S}^t = (\hat{w}_1^t, ..., \hat{w}_m^t)$ such that

$$\hat{w}_i^t = \begin{cases} w^r & \text{if } \exists w^t : (w^r, w^t) \in M \;\&\; \mathrm{AD}(w^r, w^t) \leq \epsilon \\ w_i^t & \text{otherwise.} \end{cases}$$

This means that the words from the test sentence which were matched with the affix distance smaller than ϵ are "corrected" by substituting them by the matching words from the reference sentence. The threshold ϵ is a free parameter of the metric. When the threshold is set to zero, no corrections are made and therefore the metric is equivalent to the standard BLEU.

[1]The matching is always one-to-one which means that some words remain unmatched if the sentences have different number of words.

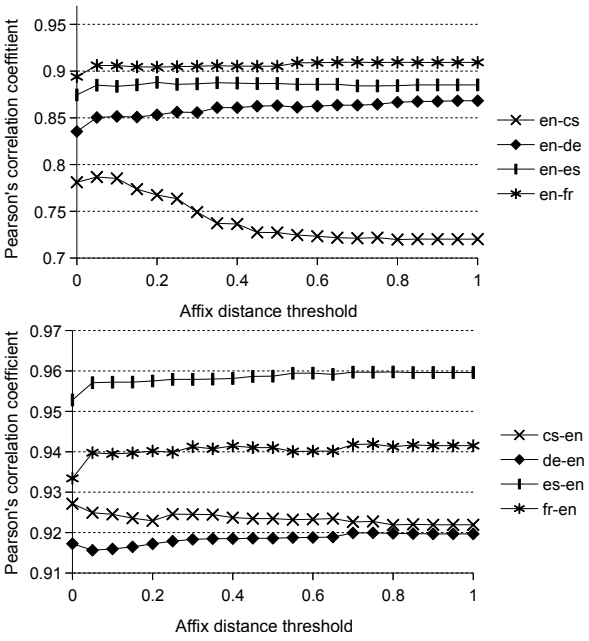

Figure 2: Dependence of the Pearson's correlation of tBLEU with the WMT13 human judgments on the affix distance threshold for translations from English and to English.

The words in the corrected sentence are assigned the weights as follows:

$$v(\hat{w}_i^t) = \begin{cases} 1 - \mathrm{AD}(\hat{w}_i^t, w_i^t) & \text{if } \hat{w}_i^t \neq w_i^t \\ 1 & \text{otherwise.} \end{cases}$$

In other words, the weights penalize the corrected words proportionally to the affix distance from the original words.

While computing the n-gram precision, two matching n-grams $(\hat{w}_1^t, \ldots \hat{w}_n^t)$ and $(w_1^r, \ldots w_n^r)$ contribute to the n-gram precision with a score of

$$s(w_1^t, \ldots, w_n^t) = \sum_{i=1}^{n} v(\hat{w}_i^t) \;/\; n$$

instead of one as it is in the standard BLEU. The rest of the BLEU score computation remains unchanged. While using multiple reference translation, the matching is done for each of the reference sentence, and while computing the n-gram precision, the reference sentences with the highest weight is chosen. The computation of the n-gram precision is illustrated in Figure 1.

direction	BLEU	METEOR	tBLEU
en-cs	.781	.860	.787
en-de	.835	.868	.850
en-es	.875	.878	.884
en-fr	.887	.906	.906
from English	.844	.878	.857

Table 1: System level Pearson's correlation with the human judgment for systems translating from English computed on the WMT13 dataset.

direction	BLEU	METEOR	tBLEU
cs-en	.925	.985	.927
de-en	.916	.962	.917
es-en	.957	.968	.953
fr-en	.940	.983	.933
to English	.923	.974	.935

Table 2: System level Pearson's correlation with the human judgment for systems translating to English computed on the WMT13 dataset.

4 Evaluation

We evaluated the proposed metric on the dataset used for the WMT13 Metrics Task (Macháček and Bojar, 2013). The dataset consists of 135 systems' outputs in 10 directions (5 into English 5 out of English). Each system's output and the reference translation contain 3000 sentences. According to the WMT14 guidelines, we report the the Pearson's correlation coefficient instead of the Spearman's coefficient that was used in the last years.

Twenty values of the affix distance threshold were tested in order to estimate what is the most suitable threshold setting. We report only the system level correlation because the metric is designed to compare only the whole system outputs.

5 Results

The tBLEU metric generally improves the correlation with human judgment over the standard BLEU metric for directions from English to languages with richer inflection.

Examining the various threshold values showed that dependence between the affix distance threshold and the correlation with the human judgment varies for different language pairs (Figure 2). For translation from English to morphologically richer languages than English – Czech, German, Spanish and French – using the tBLEU metric increased the correlation over the standard BLEU. For Czech the correlation quickly decreases for threshold values bigger than 0.1, whereas for the other languages it still grows. We hypothesize this because the big morphological changes in Czech can entirely change the meaning.

For translation to English, the correlation slightly increases with the increasing threshold value for translation from French and Spanish, but decreases for Czech and German.

There are different optimal affix distance thresholds for different language pairs. However, the threshold of 0.05 was used for our WMT14 submission because it had the best average correlation on the WMT13 data set. Tables 1 and 2 show the results of the tBLEU for the particular language pairs for threshold 0.05. While compared to the BLEU score, the correlation is slightly higher for translation from English and approximately the same for translation to English.

The results on the WMT14 dataset did not show any improvement over the BLEU metric. The reason of the results will be further examined.

6 Conclusion and Future Work

We presented tBLEU, a language-independent MT metric based on the standard BLEU metric. It introduced the affix distance – relative edit distances of prefixes and suffixes of two string after removing their longest common substring. Finding a matching between translation and reference sentences with respect to this matching allows a penalized substitution of words which has been most likely wrongly inflected and therefore less penalizes errors in inflection.

This metric achieves a higher correlation with the human judgment than the standard BLEU score for translation to morphological richer languages without the necessity to employ any language specific tools.

In future work, we would like to improve word alignment between test and reference translations by introducing word position and potentially other features, and implement tBLEU in MERT to examine its impact on system tuning.

7 Acknowledgements

This research has been funded by the Czech Science Foundation (grant n. P103/12/G084) and the EU FP7 project Khresmoi (contract no. 257528).

References

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling sparse data issue in machine translation evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91. Association for Computational Linguistics.

Sherri Condon, Gregory A Sanders, Dan Parvaz, Alan Rubenstein, Christy Doran, John Aberdeen, and Beatrice Oshika. 2009. Normalization for automated metrics: English and arabic speech translation. *Proceedings of MT Summit XII. Association for Machine Translation in the Americas, Ottawa, ON, Canada*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 85–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. Terrorcat: a translation error categorization-based mt quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, Hermann Ney, et al. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pages 240–247. Citeseer.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

James Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.