BLEU in characters: towards automatic MT evaluation in languages without word delimiters

Etienne Denoual
etienne.denoual@atr.jp

Yves Lepage
yves.lepage@atr.jp

ATR – Spoken language communication research labs
Keihanna gakken tosi, 619-0288 Kyoto, Japan

Abstract

Automatic evaluation metrics for Machine Translation (MT) systems, such as BLEU or NIST, are now well established. Yet, they are scarcely used for the assessment of language pairs like English-Chinese or English-Japanese, because of the word segmentation problem. This study establishes the equivalence between the standard use of BLEU in word n-grams and its application at the character level. The use of BLEU at the character level eliminates the word segmentation problem: it makes it possible to directly compare commercial systems outputting unsegmented texts with, for instance, statistical MT systems which usually segment their outputs.

1 Introduction

Automatic evaluation metrics for Machine Translation (MT) systems, such as BLEU (PAPINENI et al., 2001) or NIST (DODDINGTON, 2002), are now well established. They serve as quality assessment methods or comparison tools and are a fast way of measuring improvement. Although it is claimed that such objective MT evaluation methods are language-independent, they are usually only applied to English, as they basically rely on word counts. In fact, the organisers of campaigns like NIST (PRZYBOCKI, 2004)[1], TIDES[2] or IWSLT (AKIBA et al., 2004)[3], prefer to evaluate outputs of machine translation systems which are already segmented into words before applying such objective evaluation methods. The consequence of this state of affairs is that evaluation campaigns of English to Japanese or English to Chinese machine translation systems for instance, are not, to our knowledge, widely seen or reported.

2 Overview

2.1 The word segmentation problem

As statistical machine translation systems basically rely on the notion of words through their lexicon models (BROWN et al., 1993), they are usually capable of outputting sentences already segmented into words when they translate into languages like Chinese or Japanese. But this is not necessarily the case with commercial systems. For instance, Systran[4] does not output segmented texts when it translates into Chinese or Japanese.

As such, comparing systems that translate into languages where words are not an immediate given in unprocessed texts, is still hindered by the human evaluation bottleneck. To compare the performance of different systems, segmentation has to be performed beforehand.

[1] http://www.nist.gov/speech/tests/mt-/doc/mt04_evalplan.v2.1.pdf
[2] http://www.nist.gov/speech/tests/mt-/mt_tides01_knight.pdf
[3] http://www.slt.atr.jp/IWSLT2004-/archives/000619.html
[4] http://www.systranbox.com/systran-/box.

One can always apply standard word segmentation tools (for instance, The Peking University Segmenter for Chinese (DUAN et al., 2003) or ChaSen for Japanese (MATSUMOTO et al., 1999)), and then apply objective MT evaluation methods. However, the scores obtained would be biased by the error rates of the segmentation tools on MT outputs[5]. Indeed, MT outputs still differ from standard texts, and their segmentation may lead to a different performance. Consequently, it is difficult to directly and fairly compare scores obtained for a system outputting non-segmented sentences with scores obtained for a system delivering sentences already segmented into words.

2.2 BLEU in characters

Notwithstanding the previous issue, it is undeniable that methods like BLEU or NIST have been adopted by the MT community as they measure complementary characteristics of translations: namely fluency and adequacy (AKIBA et al., 2004, p. 7). Although far from being perfect, they definitely are automatic, fast, and cheap. For all these reasons, one cannot easily ask the MT community to give up their practical know-how related to such measures. It is preferable to state an equivalence with well established measures than to merely look for some correlation with human scores, which would indeed amount to propose yet another new evaluation method.

Characters are always an immediate given in any electronic text of any language, which is not necessarily the case for words. Based on this observation, this study shows the effect of shifting from the level of words to the level of characters, *i.e.*, of performing all computations in characters instead of words. According to what was said above, the purpose is not to look for any correlation with human scores, but to establish an equivalence between BLEU scores obtained in two ways: on characters and on words.

Intuitively a high correlation should exist. The contrary would be surprising. However, the equivalence has yet to be determined, along with the corresponding numbers of characters and words for which the best correlation is obtained.

[5]Such error rates are around 5% to 10% for standard texts. An evaluation of the segmentation tool is in fact required. on MT outputs alone.

3 Experimental setup

The most popular off-the-shelf objective methods currently seem to be BLEU and NIST. As NIST was a modification of the original definition of BLEU, the work reported here concentrates on BLEU. Also, according to (BRILL and SORICUT, 2004), BLEU is a good representative of a class of automatic evaluation methods with the focus on precision[6].

3.1 Computation of a BLEU score

For a given maximal order N, a baseline BLEU$_{wN}$ score is the product of two factors: a brevity penalty and the geometric average of modified n-gram precisions computed for all n-grams up to N.

$$\text{BLEU}_{wN}\text{ score} = BP \times \sqrt[N]{\prod_{n=1}^{N} p_n}$$

The brevity penalty is the exponential of the relative variation in length against the closest reference:

$$BP = \begin{cases} 1 & \text{if } |\mathcal{C}| > |\mathcal{R}_{\text{closest}}| \\ e^{1-r/c} & \text{if } |\mathcal{C}| \leq |\mathcal{R}_{\text{closest}}| \end{cases}$$

where $\mathcal{C}$ is the candidate and $\mathcal{R}_{\text{closest}}$ is the closest reference to the candidate according to its length. $|\mathcal{S}|$ is the length of a sentence $\mathcal{S}$ in words. Using a consistent notation, we note as $|\mathcal{S}|_{\mathcal{W}}$ the number of occurrences of the (sub)string $\mathcal{W}$ in the sentence $\mathcal{S}$, so that $|\mathcal{S}|_{w_1 \ldots w_n}$ is the number of occurrences of the word n-gram $w_1 \ldots w_n$ in the sentence $\mathcal{S}$.

With the previous notations, a modified n-gram precision for the order n is the ratio of two sums[7]:

$$p_n = \frac{\sum_{w_1 \ldots w_n \in \mathcal{C}} \min\left(|\mathcal{C}|_{w_1 \ldots w_n}, \max_{\mathcal{R}} \left(|\mathcal{R}|_{w_1 \ldots w_n} \right) \right)}{\sum_{w_1 \ldots w_n \in \mathcal{C}} |\mathcal{C}|_{w_1 \ldots w_n}}$$

- the numerator gives the number of n-grams of the candidate appearing in the references,

[6]ROUGE (LIN and HOVY, 2003) would be a representative of measures with the focus on recall.

[7]We limit ourselves to the cases where one candidate or one reference is one sentence.

limited to the maximal number of occurrences of the n-gram considered in a single reference[8];

- the denominator gives the total number of n-grams in the candidate.

We leave the basic definition of BLEU untouched. The previous formulae can be applied to character n-grams instead of word n-grams. In the sequel of this paper, for a given order N, the measure obtained using words will be called BLEU$_{wN}$, whereas the measure in characters for a given order M will be noted BLEU$_{cM}$.

3.2 The test data

We perform our study on English because a language for which the segmentation is obvious and undisputable is required. On Japanese or Chinese, this would not be the case, as different segmenters differ in their results on the same texts[9].

The experiments presented in this paper rely on a data set consisting of 510 Japanese sentences translated into English by 4 different machine translation systems, adding up to $2,040$ candidate translations. For each sentence, a set of 13 references had been produced by hand in advance.

Different BLEU scores in words and characters were computed for each of the $2,040$ English candidate sentences, with their corresponding 13 reference sentences.

4 Results: equivalence BLEU$_{wN}$ / BLEU$_{cM}$

To investigate the equivalence of BLEU$_{wN}$ and BLEU$_{cM}$, we use three methods: we look for the best correlation, the best agreement in judgements between the two measures, and the best behaviour, according to an intrinsic property of BLEU.

4.1 Best correlation

For some given order N, our goal is to determine the value of M for which the BLEU$_{cM}$ scores (in characters) are best correlated with the scores obtained with BLEU$_{wN}$. To this end, we compute for all possible Ns and Ms all Pearson's correlations between scores obtained with BLEU$_{wN}$ and BLEU$_{cM}$. We then select for each N, that M which gives a maximum in correlation. The results[10] are shown in Table 1. For $N = 4$ words, the best M is 17 characters.

4.2 Best agreement in judgement

Similar to the previous method, we compute for all possible Ms and Ns all Kappa coefficients between BLEU$_{wN}$ and BLEU$_{cM}$ and then select, for each given N, that M which gives a maximum. The justification for such a procedure is as follows.

All BLEU scores fall between 0 and 1, therefore it is always possible to recast them on a scale of grades. We arbitrarily chose 10 grades, ranging from 0 to 9, to cover the interval $[\,0\,,\,1\,]$ with ten smaller intervals of equal size. A grade of 0 corresponds to the interval $[\,0\,,\,0.1\,[$, and so on, up to grade 9 which corresponds to $[\,0.9\,,\,1\,]$. A sentence with a BLEU score of, say 0.435, will be assigned a grade of 4.

By recasting BLEU scores as described above, they become judgements into discrete grades, so that computing two different BLEU scores first in words and then in characters for the same sentence, is tantamount to asking two different judges to judge the same sentence. A well-established technique to assess the agreement between two judges being the computation of the Kappa coefficient, we use this technique to measure the agreement between any BLEU$_{wN}$ and any BLEU$_{cM}$.

The maximum in the Kappa coefficients is reached for the values[11] given in Table 1. For $N = 4$ words, the best M is 18 characters.

[8] This operation is referred to as clipping in the original paper (PAPINENI et al., 2001).

[9] Although we already applied the method in characters on unsegmented Japanese or Chinese MT outputs, this is not the object of the present study, which, again, is to show the equivalence between BLEU in words and characters.

[10] The average ratio M/N obtained is 4.14, which is not that distant from the average word length in our data set: 3.84 for the candidate sentences.

Also, for $N = 4$, we computed all values of Ms for each sentence length. See Table 2.

[11] Except for $N = 3$, where the value obtained (14) is quite different from that obtained with Pearson's correlation (10), the values obtained with Kappa coefficients atmost differ by 1.

4.3 Best analogical behaviour

BLEU depends heavily on the geometric average of modified n-gram precision scores. Therefore, because one cannot hope to find a given n-gram in a sentence if neither of the two included $(n-1)$-grams is found in the same sentence, the following property holds for BLEU:

> For any given N, for any given candidate, for any given set of references,
>
> $$\text{BLEU}_{wN} \leq \text{BLEU}_{w(N-1)}$$

The left graph of Figure 2 shows the correspondence of BLEU_{w4} and BLEU_{w3} scores for the data set. Indeed all points are found on the diagonal or below.

Using the property above, we are interested in finding experimentally the value M such that $\text{BLEU}_{cM} \leq \text{BLEU}_{w(N-1)}$ is true for almost all values. Such a value M can then be considered to be the equivalent in characters for the value N in words.

Here we look incrementally for the M allowing BLEU_{cM} to best mimic BLEU_{wN}, that is leaving at least 90% of the points on or under the diagonal. For $N = 4$, as the graph in the middle of Figure 2 illustrates, such a situation is first encountered for $M = 18$. The graph on the right side shows the corresponding layout of the scores for the data set. This indeed tends to confirm that the M for which BLEU_{cM} displays a similar behaviour to BLEU_{w4} is around 18.

5 The standard case of system evaluation

5.1 $\text{BLEU}_{w4} \simeq \text{BLEU}_{c18}$

According to the previous results, it is possible to find some M for some given N for which there is a high correlation, a good agreement in judgement and an analogy of behaviour between measures in characters and in words. For the most widely used value of N, 4, the corresponding values in characters were 17 according to correlation, 18 according to agreement in judgement, and 18 according to analogical behaviour. We thus decide to take 18 as the number of characters corresponding to 4 words (see Figure 1 for plots of scores in words against scores in characters).

5.2 Ranking systems

We recomputed the overall BLEU scores of the four MT systems whose data we used, with the usual BLEU_{w4} and its corresponding method in characters, BLEU_{c18}. Table 3 shows the average values obtained on the four systems.

When going from words to characters, the values decrease by an average of 0.047. This is explained as follows: a sentence of less than N units, has necessarily a BLEU score of 0 for N-grams in this unit. Table 4 shows that, in our data, there are more sentences of less than 18 characters (350) than sentences of less than 4 words (302). Thus, there are more 0 scores with characters, and this explains the decrease in system scores when going from words to characters.

On the whole, Table 3 shows that happily enough, shifting from words to characters in the application of the standard BLEU measure leaves the ranking unchanged[12].

6 Conclusion

We studied the equivalence of applying the BLEU formula in character M-grams instead of word N-grams. Our study showed a high correlation, a good agreement in judgement, and an analogy of behaviour for definite corresponding values of M and N. For the most widely used value of N, 4, we determined a corresponding value in characters of 18.

Consequently, this study paves the way to the application of BLEU (in characters) in objective evaluation campaigns of automatic translation into languages without word delimiters, like Chinese or Japanese, as it avoids any problem with segmentation.

Acknowledgements

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

[12] (ZHANG et al., 2004) reported confidence intervals of around 2% (*i.e.*, in this case, ± 0.01) for BLEU, so that system 2 and 3 are undistinguishable by BLEU_{w4}.

References

Yasuhiro AKIBA, Marcello FEDERICO, Noriko KANDO, Hiromi NAKAIWA, Michael PAUL, and Jun'ichi TSUJII. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Eric BRILL and Radu SORICUT. 2004. A unified framework for automatic evaluation using n-gram co-occurence statistics. In *Proceedings of ACL 2004*, pages 613–620, Barcelone.

Peter E. BROWN, Vincent J. DELLA PIETRA, Stephen A. DELLA PIETRA, and Robert L. MERCER. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, Special Issue on Using Large Corpora: II*, 19(2):263–311.

George DODDINGTON. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of Human Language Technology*, pages 128–132, San Diego.

Huiming DUAN, Xiaojing BAI, Baobao CHANG, and Shiwen YU. 2003. Chinese word segmentation at Peking University. In Qing Ma and Fei Xia, editors, *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 152–155.

Chin-Yew LIN and Eduard HOVY. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL 2003*, pages 71–78, Edmonton.

Y. MATSUMOTO, A. KITAUCHI, T. YAMASHITA, Y. HIRANO, H. MATSUDA, and M. HASAHARA. 1999. Japanese morphological analysis system ChaSen version 2.0. Technical report NAIST-IS-TR99009, Nara Institute of Technology.

Kishore PAPINENI, Salim ROUKOS, Todd WARD, and Wei-Jing ZHU. 2001. Bleu: a method for automatic evaluation of machine translation. Research report RC22176, IBM.

Mark PRZYBOCKI. 2004. The 2004 NIST machine translation evaluation plan (MT-04).

Ying ZHANG, Stefan VOGEL, and Alex WAIBEL. 2004. Interpreting BLEU/NIST scores: how much improvement do we need to have a better system? In *Proceedings of LREC 2004*, volume V, pages 2051–2054, Lisbonne.

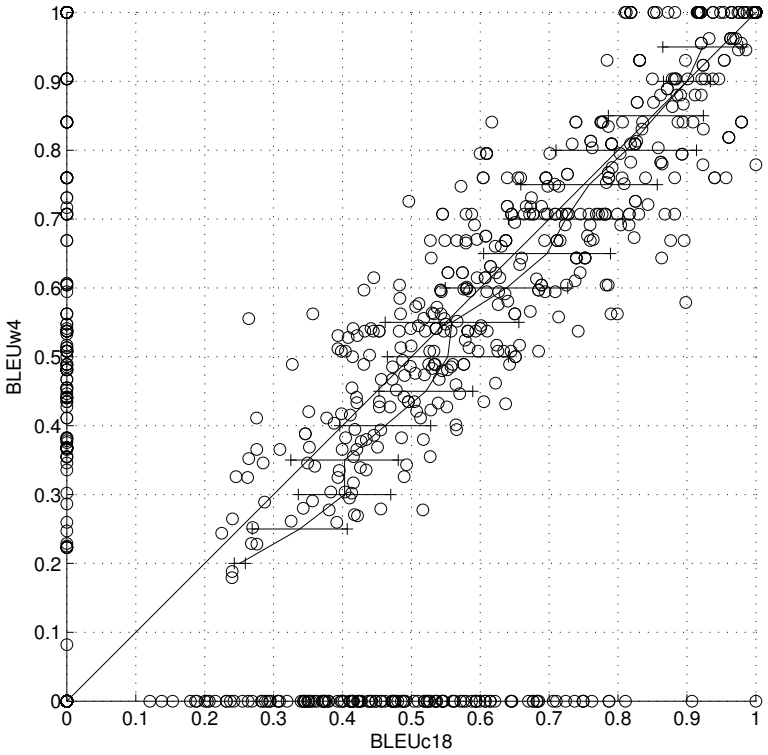

Figure 1: BLEU_{w4} in ordinates against BLEU_{c18} in abscissae.

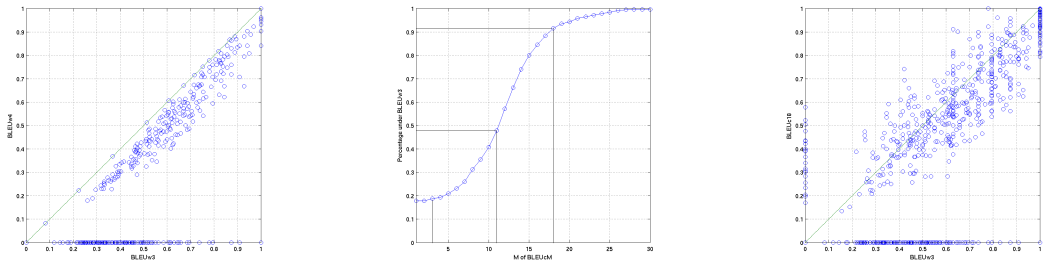

Figure 2: On the left, experimental scores for BLEU$_{w4}$ versus BLEU$_{w3}$: all points are on the diagonal or below. On the right, BLEU$_{c18}$ scores versus BLEU$_{w3}$: 90% of the points are on the diagonal or below. In the middle, proportion of BLEU$_{cM}$ scores under BLEU$_{w3}$ for M varying from 1 to 30.

Table 1: Equivalent Ns and Ms for BLEU$_{wN}$ and BLEU$_{cM}$ obtained by different methods.

	BLEU$_{w1}$	BLEU$_{w2}$	BLEU$_{w3}$	BLEU$_{w4}$
Pearson's correlation (best M)	0.89 (5)	0.90 (8)	0.85 (10)	0.83 (17)
Kappa value (best M)	0.17 (5)	0.29 (9)	0.34 (14)	0.35 (18)
best M for analogical behaviour wrt to $(N-1)$ (threshold = 90%)		(9)	(14)	(18)

Table 2: Correlation of BLEU$_{w4}$ scores with BLEU$_{c18}$ scores by sentence length.

sentence length	4	5	6	7	8	9	10	>10
points	12.9%	18.2%	13.6%	13.4%	7.5%	6.5%	5.0%	8.1%
average BLEU$_{w4}$ score	0.188	0.300	0.252	0.364	0.345	0.318	0.321	0.015
std. dev.	$\pm$0.389	$\pm$0.416	$\pm$0.376	$\pm$0.382	$\pm$0.363	$\pm$0.3150	$\pm$0.346	$\pm$0.291
local best M	16	17	16	19	17	17	16	12
Pearson's correlation	0.827	0.795	0.797	0.824	0.899	0.894	0.952	0.919
global best M				18				
Pearson's correlation	0.788	0.794	0.779	0.805	0.883	0.871	0.929	0.861

Table 3: Overall BLEU scores for 4 different systems in BLEU$_{w4}$ and BLEU$_{c18}$.

	system 1		system 2		system 3		system 4
overall BLEU$_{w4}$ score	0.349	$>$	0.305	$\sim$	0.312	$>$	0.232
overall BLEU$_{c18}$ score	0.292	$>$	0.279	$>$	0.267	$>$	0.183
difference in scores	-0.057		-0.036		-0.045		-0.049

Table 4: Distribution of the 510 sentences by lengths in words and characters.

length	<4 words	≥ 4 words	total
<18 characters	266	84	**350**
≥ 18 characters	37	123	160
total	**302**	208	510