10.31341/jios.47.2.12
Open Access
UDC 004.412:004.85
Original Scientific Paper

Towards a Combination of Metrics for Machine Translation

Mawloud Mosbah *mos_nasa@hotmail.fr*
LRES Laboratory, Informatics Department
Faculty of Sciences
University 20 Août 1955 of Skikda, Algeria

Abstract

In this scholar, we compare three metrics for machine translation, from English to French and vice versa, and we give some combination formulas based on some schemes, algorithms, and machine learning tools. As an experimental dataset, we consider 10 English and French theses abstracts published in the web with four free in charge machine translation systems. Five combinations, with the same implicit weights, are considered namely: (BLEU+NIST), (BLEU+ (1-WER)), (NIST+(1-WER)), (BLEU+NIST+(1-WER)), and (FR(BLEU)+FR(NIST)+FR(WER)). These combinations are also considered differently through generating weights parameters on the basis of regression. The results of 12 formulas are computed and compared then in total. According to the obtained results, average regression combinations based on machine learning step are the best, especially with the three basic metrics, followed by average WER metric in the case of English to French. For French to English, (FR(BLEU)+FR(NIST)+FR(WER)) combination is the best followed respectively by the average regression combination with both first parameters (Reg(α,β)) and average BLEU basic metric. Another performance criterion is considered here, in the second position, namely: the number of times, over the 10 abstracts, where the formula is the best. Based on the obtained results, combination with regression based on the first and the last parameters (Reg(α,γ)) outperforms the others, in the case of English to French, with 3 times followed by Reg(β,γ), Reg(α,β,γ), NIST+(1-WER), and the basic metrics (BLEU, NIST, and WER) with 2 times for each of them. For French to English, the basic WER metric outperforms the others with three times followed by BLEU, (BLEU+ (1-WER)), (FR(BLEU)+FR(NIST)+FR(WER)), and Reg(α,γ) with 2 times for each of them. To note that there is a room of improvement for the combinations with1.0914 in the case of English to French and 1.01 in the case of French to English.

Keywords: Machine Translation, Machine Translation Metrics, Combination of Machine Translation Metrics

1. Introduction

Evaluation is an important operation for any scientific field. Indeed, assessment enables us to know in which degree the addressed model is effective, what is its room of optimization and improvement through analyzing and identifying its various weaknesses, and to compare the model in question with other ones proposed in the literature. For machine translation, evaluation is qualified as a difficult task for the purpose that there are many possible ways to translate a given source sentence. There are two approaches for machine translation evaluation: manual, subjective, and qualitative assessment, done by human experts, and automatic and objective and numerical evaluation implemented by fully automatic metrics [1]. Three aspects are tied to human evaluation of machine translation: fluency, indicating how natural the evaluation segment sounds to a native speaker, adequacy, judging how much of the information from the original translation is expressed in the output, and acceptability, judging how easy to understand the translation is. Unfortunately, human evaluation is subjective and time consuming. Automatic evaluation proceeds to compare the output segment of the translation to the reference one either through associating a score for each translation [2] or to rank the different translations of the same input [3] to each other. In [4], authors have categorized automatic metrics into deterministic metrics, tending to focus on specific aspects of the evaluation, and learned metrics such as BLANC [5], trying to gather and combine all the aspects into a single metric. According to [6], human evaluation metrics are classified based on the criterion of whether human judges which expresses a so-called subjective evaluation judgment, such as 'good' or 'better than', or not. The former methods are based on directly expressed judgment (DEJ) while the latter are called 'non-DEJ-based evaluation. For the DEJ-based evaluation, there are tasks such as fluency and adequacy annotation, ranking and direct assessment (DA) such as Blend [7] whereas for the non-DEJ-based evaluation, there are tasks like error classification and post-editing.

Unfortunately, as reported in [8], there is no automatic metric that practically outperforms the other metrics of the literature or to well reflect human judgment. Combining automatic metrics seems to be then a good idea with two issues to be tackled, namely: (1) what are the metrics to combine and how many numbers of them and (2) which weight to attribute for each one. The scheme of combining metrics have been previously considered in natural language processing applications such as in information retrieval in the image of f-measure [9] which combines both precision and recall.

To the best of our knowledge, there is only one work, in literature, that deals with combination of evaluation metrics for machine translation as we address here. Indeed, in [10], authors have applied a loss function, as an approach from statistical decision theory for weighted cost estimates, to combine three basic metrics namely: correct response, non-response, and incorrect response rates. However, there are few works that address combination differently such as in [11] where authors have combined automatic metrics for predicting human assessment using binary classifiers and in [12] where author has quoted some works that combines evaluation metrics with error classification and analysis. A regression is also taken into consideration here in different ways as considered in [4] where authors have proceeded to combine different criteria and aspects of machine translation.

The methodology adopted in this paper is as follows: as a purpose, we look for the effective formula that combines three basic evaluation metrics for machine translation systems namely: BLEU, NIST, and WER. Four machine translation systems are considered namely: Google Translate, Promt, Babylon and Bing over 10 theses abstracts in both English and French. Human expert evaluation, through manual assessment and judgment of the various returned translation outputs, is also taken into consideration to evaluate the performance of BLEU, NIST, and WER as well as the different considered combinations. Based on each machine translation metric and the different combination metrics, the outputs of the adopted four machine translation systems are ranked. These ranks are compared with those given by human experts using an information retrieval evaluation metric which is NDCG. The performance of the three primitive machine translation metrics and especially their different combinations are assessed then using the NDCG metric.

The rest of the paper is organized as follows: section 2 presents the considered four free in charge machine translation systems. In section 3, we show, in details, the different three considered evaluation metrics for machine translation. Section 4 depicts the different considered combinations with and without regression. Datasets and results with their discussions are given in section 5. In Section 6, we establish a conclusion and we draw some perspectives may be implemented later in our future works.

2. The Considered Free in Charge Machine Translation Systems

In this work, we consider the following four machine translation systems available free in the web:

- *Google Translate* [13]: is a multilingual neural machine translation [14] service developed by Google to translate text, documents and websites from one language to another. As of July 2021, Google Translate supports 109 languages at various levels. Moreover, as of April 2016, Google Translate claimed over 500 million total users, with more than 100 billion words daily translated.
- *Promt* [15]: is a Russian company focused upon the development of machine translation systems. At the moment, Promt translator exists for more than 25 languages.
- *Babylon* [16]: is a computer dictionary and translation program developed by the company Babylon Software.
- *Bing* [17]: is a service which allows users to translate text and web pages into different languages using the Microsoft statistical machine translation system.

In addition to the free in charge aspect, the four considered machine translation systems adopt different approaches which enables us to obtain theoretically different results.

3. Machine Translation Metrics

We consider here three statistical metrics to evaluate machine translation namely: *BLEU*, *NIST*, and *WER*.

3.1. BLEU

BLEU (*Bi-Lingual Evaluation Understudy*) is an automatic metric baptized by *IBM* employing several references [18]. In its simple manner, *BLEU* measures how many sequence of words in the block for text under evaluation (the candidate block text) match the sequence of words of some reference blocks of text. It also contains a penalty for translations whose length differs significantly from that of the reference translation. *BLEU* metric is based firstly on computing n-grams (or chunks; that are sequence of words) for both the block of text under evaluation and the reference block of text. Secondly, the clipper chunk Counts for the candidate block text is added and divided by the number of candidate chunks in the reference block of text to compute its modified precision score p_n as follows:

$$p_n = \frac{\sum_{C\in(Candidates)}\sum_{ngrams\in C} Count_{clip}(ngram)}{\sum_{C'\in(Candidates)}\sum_{ngram'\in C'} Count\left(ngram'\right)'} \quad (1).$$

Where: $Count_{clip}(ngram)$ is the maximum number of n-grams co-occurring in a candidate translation and a reference translation, and $Count(ngram)$ is the number of n-grams in the candidate translation.

Let c be the length of the candidate translation and r be the effective reference block of text length, the brevity penalty *BP*, for preventing very short translations that try to maximize their precision scores, is computed as follows:

$$BP = \begin{cases} 1 & if\ \ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad (2)$$

Then

$$BLEU = BP * e^{\left(\sum_{n=1}^{N} w_n * log(P_n)\right)} \quad (3)$$

Where: w_n represents the weights given to the number of words constituting the chunks or n-grams. According to [18], the ranking behaviour is more immediately apparent in the logarithm domain as follows:

$$log(BLEU) = min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^{N} w_n log(P_n) \quad (4)$$

Commonly, the following factors are set as follows: N=4 (ie. we use uni-gram, bi-grams, three-grams, and four-grams) and $w_n = {}^{1}\!/_{N}$.

3.2. NIST

NIST (*US National Institute of Standards and Technology*) score weights more heavily on those n-grams occurring less frequently according to their information value [19], [20]. That is to say when a correct n-gram is found, the rarer that n-gram is, the more weight it will be given. The formula of *NIST* score is given as follows:

$$score_{NIST} = \sum_{n=1}^{N} \left\{ {\sum_{all\, W_1..w_N cooccur} Info(W_1..W_n)} \Big/ {\sum_{all\, W_1..W_n}(1)} \right\} * e^{\left\{\beta log_2\left[min\left(\frac{L_{sys}}{L_{ref}},1\right)\right]\right\}} (5)$$

Where:

$$Info(W_1..W_n) = log_2\left(\frac{the\ \#\ of\ occurrences\ of\ W_1..W_{n-1}}{the\ \#\ of\ occurrences\ of\ W_1..W_n}\right) \quad (6)$$

L_{ref} is the average length of reference.
L_{sys} is the length of target translation.
β is used as a parameter conditioning penalty translation.

To note that both *BLEU* and *NIST* metrics are not cumulative since they are based on the candidate and reference segments at whole rather than to the addition of its sentences. That is to say it is not needed to identify the different sentences.

3.3. WER

WER (Word Error Rate) represents the percentage of words, which are to be inserted, deleted or replaced in the translation for obtaining the sentence of the reference [21]. It can be computed automatically by using the editing distance between the candidate and the reference sentences.

In the aim to encounter the dependency on the sentences of the reference, several references may be considered for each sentence. Indeed, *mWER* (multi reference *WER*) is a version of WER metric where for each sentence the editing distance will be computed with regard to the various references and the smallest one is chosen [22]. Nevertheless, adopting *mWER*, considering many references, presents the drawback of requiring a great human effort to generate references although this effort is worthwhile to be used later for hundreds of evaluations [23]. *aWER* is another version of *WER* which calculates the percentage of words to be inserted, detected or replaced in order to obtain a correct translation. Involving automatically synonyms seems to be an essential pre-processing needed in the case of *aWER*.

It is worthy to note that there are other metrics not considered here such as *METEOR*, *ROUGE*, *TER*, *SER*, and *OOV*. For more information about them, authors may ask [8] and [12].

To evaluate the performance of the ranks, for the four considered machine translation systems, given by *BLEU*, *NIST*, *WER*, and their different combinations, we use the *NDCG*.

3.4. Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain (*NDCG*) [24] is computed according to the following stages:

Given a number of recommendations or ranks where every recommendation has its associated relevance score. Cumulative Gain (CG), as shown in equation (7), is the sum of all the relevance scores in a recommendation set.

$$CG = \sum_{i=1}^{n} relevance_i \quad (7)$$

In order to well distinguish the evaluation of two different recommendations, Cumulative Gain is unfortunately not enough. For overcoming this issue, the computation involves to discount the relevance score by dividing it with the log of the corresponding position.

$$DCG = \sum_{i=1}^{n} \frac{relevance_i}{log_2(i+1)} \quad (8)$$

Alternatively, Discounted Cumulative Gain can be computed using the following expression:

$$DCG = \sum_{I+\&}^{N} \frac{2^{relevance_i -1}}{log_2(i+1)} \quad (9)$$

The recommendation should be measured relatively regarding a reference recommendation called ideal order, whose *DCG* is *iDCG*. Designating its proper upper and lower bounds simplifies comparison with other recommendations of other characteristics and parameters such as the number of considered elements. Normalized Discounted Cumulative Gain is given then in the following equation:

$$NDCG = \frac{DCG}{iDCG} \quad (10)$$

As we consider here four machine translation systems, the best value of *NDCG* (which is done manually) is then *10.563*.

4. The different Considered Combinations

In this section, we present the different combinations, for the basic metrics, considered here.

4.1. BLEU+NIST

As shown in section 3, both *BLEU* and *NIST* metrics have a trade on with machine translation performance. Adding *BLEU* to the *NIST* value seems then to be an intuitive and a logical combination way to think about for introducing a novel metric, that keeps the trade on relationship, and that we hope to be more effective. In addition, there is no need to weight the *BLEU* and *NIST* values because as shown in experimental results, given previously in [8], their values are each other closed and belong commonly to the same scale which is the [0, 1] range. We associate then the same importance for both considered basic metrics which is implicitly 1. The first considered combination formula is simply given then as follows:

$$formula1 = BLEU + NIST \quad (11)$$

4.2. BLEU+(1-WER)

In the same optic of the formula1, *BLEU* is combined now with *WER* metric. The difference is that *WER* has a trade off with machine translation performance that has a trade on with the BLEU metric. For this purpose, we need then to consider *(1-WER)* values that have a trade on with machine translation performance. Since WER values belong all to the [0, 1] range, as shown previously in [7], *(1-WER)* values belong too to the same range that of [0, 1]. As we attribute the same importance for both values: *BLEU* and *(1-WER)*, an implicit weight set as 1 is then enough. The second proposed combination formula is given then as follows:

$$formula2 = BLEU + (1 - WER) \quad (12)$$

4.3. NIST+(1-WER)

With the same thinking way, the third formula combining both *NIST* and *(1-WER)* is given as follows:

$$formula3 = NIST + (1 - WER) \quad (13)$$

4.4. BLEU+NIST+(1-WER)

The fourth formula combining *BLEU*, *NIST*, and *(1-WER)* is given as follows:

$$formula4 = BLEU + NIST + (1 - WER) \quad (14)$$

4.5. FR(BLEU)+FR(NIST)+FR(1-WER)

Another combination way is considering an algebraic expression based on a first rank function for the three basic metrics: *BLEU*, *NIST*, and *(1-WER)*. The first rank function for a machine translation system *(MTS)* is given as follows:

$$FR(MTS) = \begin{cases} 1 \; if \; MTS \; is \; ranked \; first \; based \; on \; NDCG \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad else \end{cases} \quad (15)$$

The fifth formula is given then as follows:

$$formula5 = FR(BLEU) + FR(NIST) + FR(1 - WER) \quad (16)$$

To note that *FR()* means the metric first ranked.

4.6. Combination based on Regression

A regression, as a statistical notion, refers to the prediction of the value of a dependent variable from the values of one or more independent (explanatory) variables [25]. It is able to show whether changes observed in the dependent variable are associated

with changes in one or more of the explanatory variables. To note that we consider here a multiple regression model because we have more than independent variable (*BLEU*, *NIST*, and *WER*) affecting a dependent variable which is the ranking of machine translation systems measured using *NDCG*.
Four combinations may be considered here, namely:

$$\left\{\begin{array}{c} \alpha BLEU + \beta NIST \\ \alpha BLEU + \gamma(1 - WER) \\ \beta NIST + \gamma(1 - WER) \\ \alpha BLEU + \beta NIST + \gamma(1 - WER) \end{array}\right. \quad (17)$$

Learning consists then to find the adequate or at least the optimal parameters: α, β, and γ that may give more performance in the test step. Table1 presents an illustrative learning example.

Manual Ranking		BLEU	NIST	(1-WER)
3		Google Translate		
		0.540	0.923	0.303
4		Promt		
		0.478	0.867	0.289
1		Babylon		
		0.534	0.895	0.282
2		Bing		
		0.504	0.902	0.281

Table 1. An Illustrative Learning Example

According to the example, the different non equalities to find parameters α and β considering only *BLEU* and *NIST* are:

$$\left\{\begin{array}{c} 0.534\alpha + 0.895\beta > 0.504\alpha + 902\beta \\ 0.534\alpha + 0.895\beta > 0.54\alpha + 923\beta \\ 0.534\alpha + 0.895\beta > 0.478\alpha + 0.867\beta \\ 0.504\alpha + 902\beta > 0.54\alpha + 923\beta \\ 0.504\alpha + 902\beta > 0.478\alpha + 0.867\beta \\ 0.54\alpha + 923\beta > 0.478\alpha + 0.867\beta \end{array}\right. \quad (18)$$

The different non equalities to find parameters α and γ considering only *BLEU* and *(1-WER)* are:

$$\left\{\begin{array}{c} 0.534\alpha + 0.282\gamma > 0.504\alpha + 0.281\gamma \\ 0.534\alpha + 0.282\gamma > 0.54\alpha + 0.303\gamma \\ 0.534\alpha + 0.282\gamma > 0.478\alpha + 0.289\gamma \\ 0.504\alpha + 0.281\gamma > 0.54\alpha + 0.303\gamma \\ 0.504\alpha + 0.281\gamma > 0.478\alpha + 0.289\gamma \\ 0.54\alpha + 0.303\gamma > 0.478\alpha + 0.289\gamma \end{array}\right. \quad (19)$$

The different non equalities to find parameters β and γ considering only *NIST* and *(1-WER)* are:

$$\begin{cases} 0.895\beta + 0.282\gamma > 0.902\beta + 0.281\gamma \\ 0.895\beta + 0.282\gamma > 0.923\beta + 0.303\gamma \\ 0.895\beta + 0.282\gamma > 0.867\beta + 0.289\gamma \\ 0.902\beta + 0.281\gamma > 0.923\beta + 0.303\gamma \\ 0.902\beta + 0.281\gamma > 0.867\beta + 0.289\gamma \\ 0.923\beta + 0.303\gamma > 0.867\beta + 0.289\gamma \end{cases} \quad (20)$$

The different non equalities to find parameters α, β, and γ considering *BLEU*, *NIST*, and *(1-WER)* are:

$$\begin{cases} 0.534\alpha + 0.895\beta + 0.282\gamma > 0.504\alpha + 0.902\beta + 0.281\gamma \\ 0.534\alpha + 0.895\beta + 0.282\gamma > 0.54\alpha + 0.923\beta + 0.303\gamma \\ 0.534\alpha + 0.895\beta + 0.282\gamma > 0.478\alpha + 0.867\beta + 0.289\gamma \\ 0.504\alpha + 0.902\beta + 0.281\gamma > 0.54\alpha + 0.923\beta + 0.303\gamma \\ 0.504\alpha + 0.902\beta + 0.281\gamma > 0.478\alpha + 0.867\beta + 0.289\gamma \\ 0.54\alpha + 0.923\beta + 0.303\gamma > 0.478\alpha + 0.867\beta + 0.289\gamma \end{cases} \quad (21)$$

As we can see, there are six non-equalities for each combination which gives 60 non-equalities in total (over the 10 considered abstracts). These non-equalities are solved using an experimental simple automatic algorithm with 0.01 as a walking step for the different considered parameters (α, β, and γ). The best configuration of the parameters is which satisfies the great number of non-equalities let alone all the 60 non-equalities.

5. Experimental Results

5.1. Dataset and Tools

We consider English and French abstracts for 10 theses published in the web. Four free in charge web machine translation systems, based on various schemes, are used, namely: *Google Translate*, *Promt*, *Babylon*, and *Bing*. Three basic metrics are adopted namely: *BLEU*, *NIST*, and *WER*. For combination based on regression, we use 10 abstracts for English and French. In total, the machine learning dataset is composed of 100 texts over both languages (20 references + 80 results). For combination based on regression, we consider 60 non equalities for English to French (6 non equalities over 10 learning examples) and the same number for French to English. To solve these non equalities and find parameter weights α, β, and γ, we consider a walking step of 0.01. We consider then two respective sub-sections for presenting results namely: from English to French and from French to English.

Theses Abstractions	Characteristics			
	# of Abstract in Words		# of Abstract in Sentences	
	In English	In French	In English	In French
Thesis #1	264	294	14	13
Thesis #2	295	293	10	09
Thesis #3	289	372	07	07
Thesis #4	317	439	15	15
Thesis #5	250	275	10	08
Thesis #6	578	767	25	27
Thesis #7	170	188	09	07
Thesis #8	275	313	12	12
Thesis #9	287	312	11	11
Thesis #10	301	379	11	10
Total	3026	3632	124	119

Table 2. The characteristics of the considered dataset

5.2. From English to French

Theses Abstractions	Manual Performance	The performance of the primitive metrics		
		BLEU	NIST	WER
#1	10.563	10.463	10.031	9.31
#2		10.12	9.31	10.463
#3		9.654	9.365	9.986
#4		9.365	7.911	10.175
#5		10.563	8.299	9.365
#6		8.544	9.986	8.488
#7		8.011	10.563	9.31
#8		10.375	8.821	9.121
#9		8.011	8.544	9.365
#10		8.011	8.821	8.444

Table 3. The performance of the considered primitive metrics

Theses Abstractions	The performance of the different considered combination formulas				
	BLEU+ NIST	BLEU+ (1-WER)	NIST+ (1-WER)	BLEU+NIST+ (1-WER)	FR(MTS)
#1	10.031	10.175	9.31	10.031	10.563
#2	10.563	9.931	10.375	10.563	9.654
#3	9.654	9.654	9.931	9.654	9.931
#4	8.299	9.986	10.086	7.911	9.365
#5	10.086	10.463	8.299	8.299	10.086
#6	8.544	8.011	8.388	8.388	9.11
#7	10.175	9.11	10.375	9.31	8.388
#8	6.524	9.31	8.821	9.121	9.11
#9	8.544	8.388	8.299	9.11	7.911
#10	8.444	8.544	8.821	8.444	8.544

Table 4. The performance of the different considered combination formulas without regression

Theses Abstractions	The performance of the different metric combinations based on regression			
	$\alpha BLEU + \beta NIST$	$\alpha BLEU + \gamma(1 - WER)$	$\beta NIST + \gamma(1 - WER°$	$\alpha BLEU + \beta NIST + \gamma(1 - WER)$
#1	10.031	9.31	9.31	9.31
#2	10.563	10.375	10.463	10.463
#3	9.654	10.563	10.463	10.563
#4	8.299	10.175	10.175	10.175
#5	10.086	9.365	9.365	9.11
#6	8.544	8.488	8.388	9.11
#7	10.175	9.31	9.31	9.31
#8	6.524	9.121	9.121	9.121
#9	8.544	9.365	9.365	9.11
#10	8.444	8.444	8.444	8.444
Number of verified non equalities	35	34	32	37
Different parameters	$\alpha = 0.01, \beta = 0.01$	$\alpha = 0.04, \gamma = 0.99$	$\beta = 0.05, \gamma = 0.96$	$\alpha = 0.06, \beta = 0.03, \gamma = 0.99$

Table 5. The performance of the different metric combinations based on regressio.

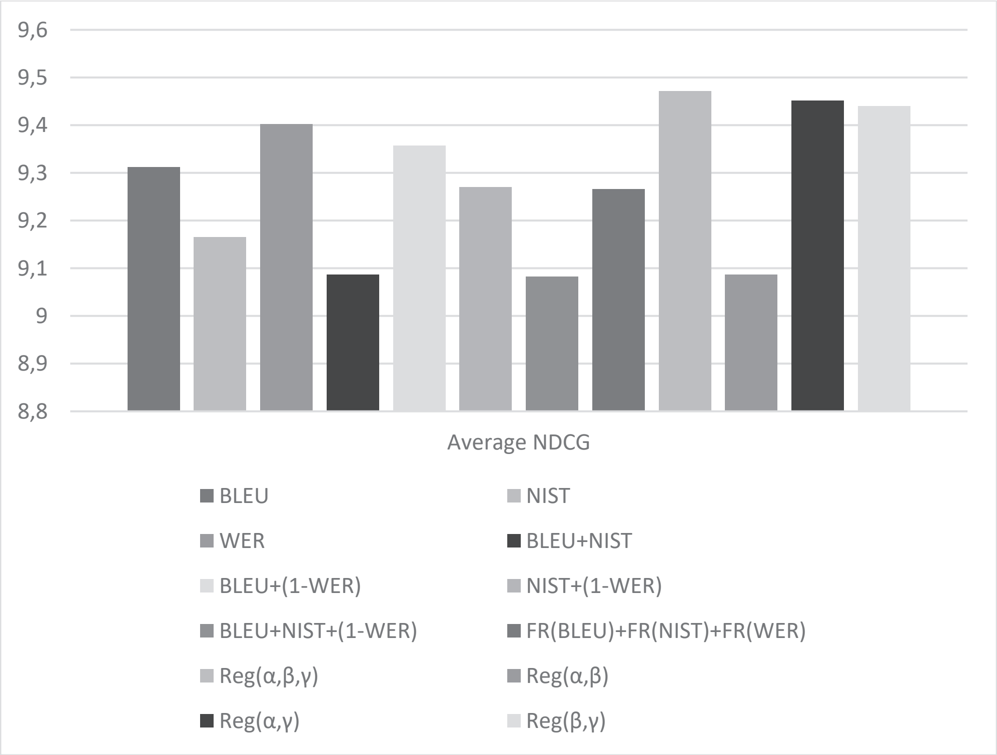

Figure 1. The considered primitive machine translation metrics and their various combinations in the case of 'From English to French'

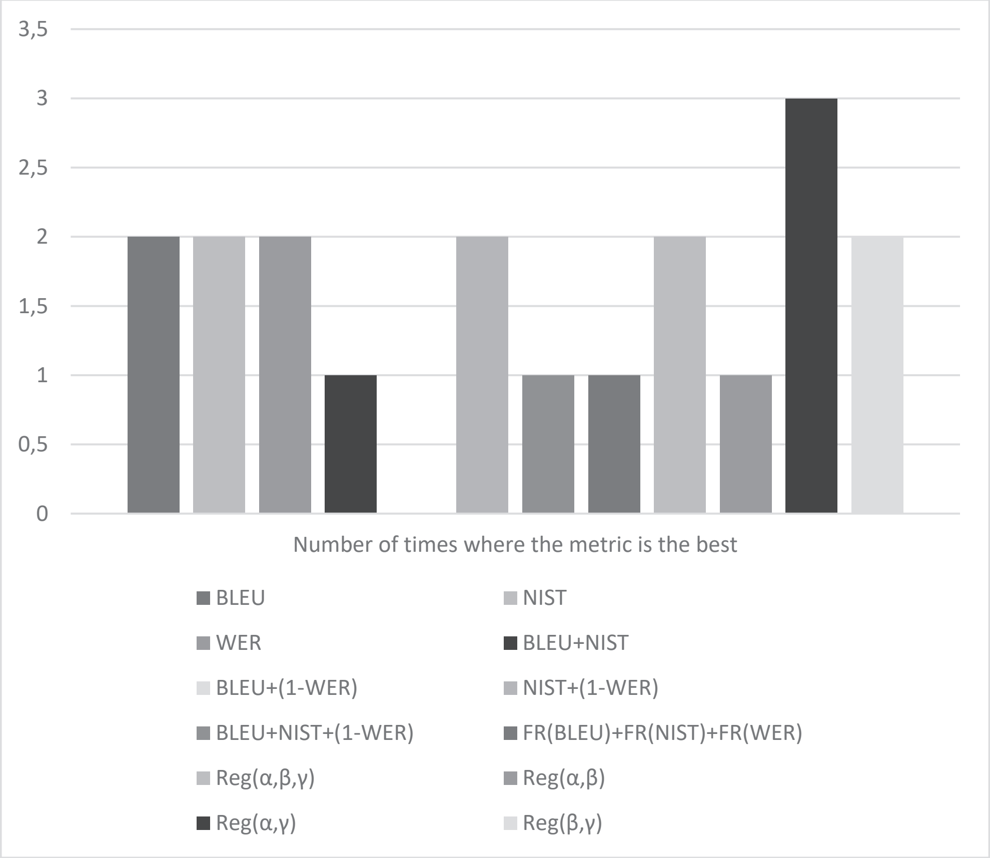

Figure 2. The number of times where the metric, from those considered and their combinations, is the best, in the case of 'From English to French'

5.3. From French to English

Theses Abstractions	Manual Performance	The performance of the primitive metrics		
		BLEU	NIST	WER
#1	10.563	8.544	8.544	10.031
#2		9.31	9.121	7.911
#3		9.654	8.544	8.488
#4		9.121	9.11	10.463
#5		9.365	8.488	10.563
#6		8.488	9.354	8.011
#7		9.986	10.031	8.388
#8		10.463	9.931	8.821
#9		9.654	10.563	8.299
#10		8.488	9.165	9.664

Table 6. The performance of the considered primitive metrics

Theses Absts	The performance of the different considered combination formulas				
	BLEU+ NIST	BLEU+ (1-WER)	NIST+ (1-WER)	BLEU+NIST+ (1-WER)	BLEU or NIST or WER
#1	8.544	8.544	8.544	8.544	9.31
#2	9.31	10.175	9.31	9.31	8.544
#3	9.31	9.31	9.121	9.31	8.544
#4	9.31	9.121	9.31	9.121	9.509
#5	8.821	10.563	10.031	9.654	10.544
#6	10.175	8.011	7.739	8.388	10.463
#7	10.031	8.388	10.563	10.086	10.463
#8	8.099	8.388	9.354	8.099	9.31
#9	9.654	9.354	9.365	9.654	10.031
#10	8.299	8.444	8.821	8.444	8.821

Table 7. The performance of the different considered combination formulas without regression

Theses Absts	The different metric combinations based on regression			
	$\alpha BLEU + \beta NIST$	$\alpha BLEU + \gamma WER$	$\beta NIST + \gamma WER$	$\alpha BLEU + \beta NIST + \gamma WER$
#1	8.821	10.175	9.986	10.175
#2	9.31	8.299	8.299	8.299
#3	9.31	8.544	8.544	8.544
#4	10.375	8.388	8.011	8.011
#5	10.175	10.563	10.031	10.031
#6	8.444	8.388	8.388	8.011
#7	9.354	8.444	10.175	10.175
#8	10.463	9.121	9.121	9.121
#9	10.031	9.165	9.165	9.165
#10	7.911	9.165	9.165	9.354
Number of verified non equalities	30	35	35	35
Parameters	$\alpha = 0.01,$ $\beta = 0.03$	$\alpha = 0.04,$ $\gamma = 0.97$	$\beta = 0.01,$ $\gamma = 0.97$	$\alpha = 0.01,$ $\beta = 0.03, \gamma = 0.89$

Table 8. The performance of the different considered combination formulas without regression.

5.4. Discussions

- Two performance evaluation criteria have been considered, here, in both cases: from English to French and vice versa. Firstly, the average accuracy computed by the automatic NDCG metric which measures the closeness of the various ranks, for the four adopted translators, given by the different formulas regarding the reference rank generated from the evaluation done by human experts. Secondly, the number of times where the formula is the best over the ten considered abstracts.
- Unfortunately, the solutions given, in the case of regression, do not satisfy all the considered 60 non-equalities. The best solution satisfies only 37 in both senses

'from English to French' and vice versa. Considering step value more little than 0.01 may improve the performance.

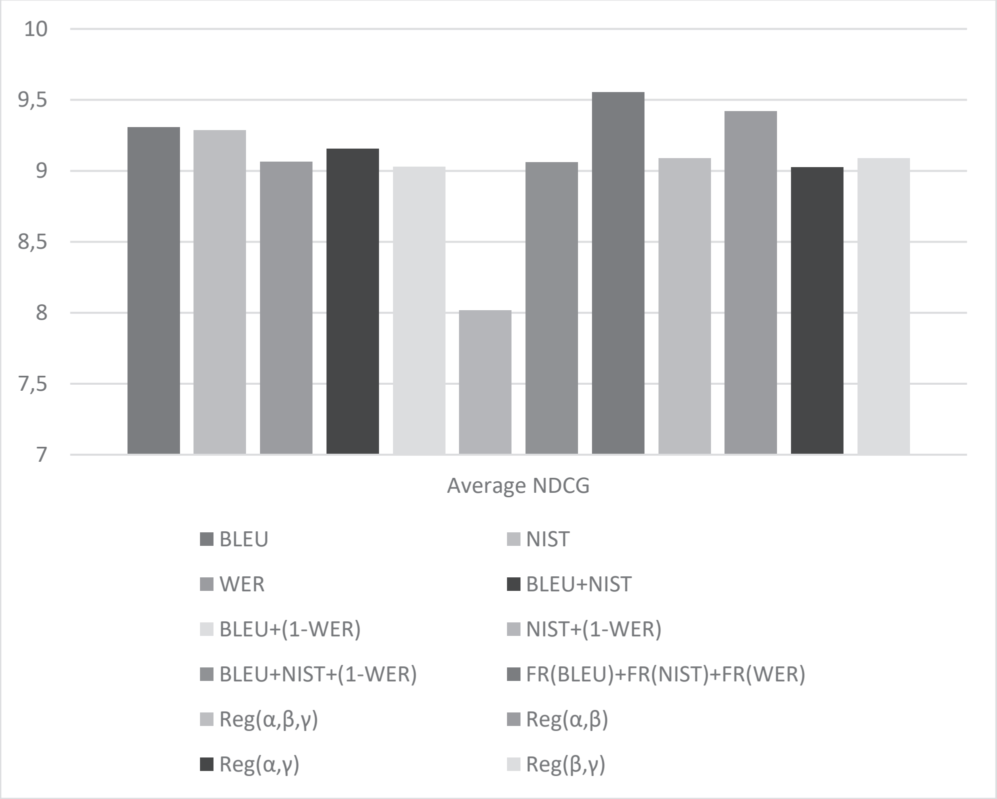

Figure 3. The considered primitive machine translation metrics and their various combinations in the case of 'From French to English'

5.4.1. From English to French

- In terms of *NDCG*, Regression, with its respective three combinations *(Reg(α,β,γ)*, *Reg(α,γ)*, *and Reg(β,γ))*, outperforms the others followed by the basic *WER* metric. Unfortunately, regression with the first and the second parameter weights *(Reg(α,β)* downgrades the performances as well as *(BLEU+NIST)* and *(BLEU+NIST+(1-WER))*.
- In terms of number of times where the formula is the best, *Reg(α,γ)* outperforms the others with 3 times followed by (*NIST+(1-WER))*, *Reg(α,β,γ)*, *Reg(β,γ)*, and the three basic metrics *(BLEU, NIST, and WER)* with 2 times for each one. *(BLEU+(1-WER))* is the only formula where there is no time to be the best.
- Although the scope of this work is to test the possible different combinations of the basic metrics for machine translation systems, it is worthy to note that the rank in terms of performance for the three considered primitive metrics is as follows: *WER*, *BLEU*, and *NIST*.

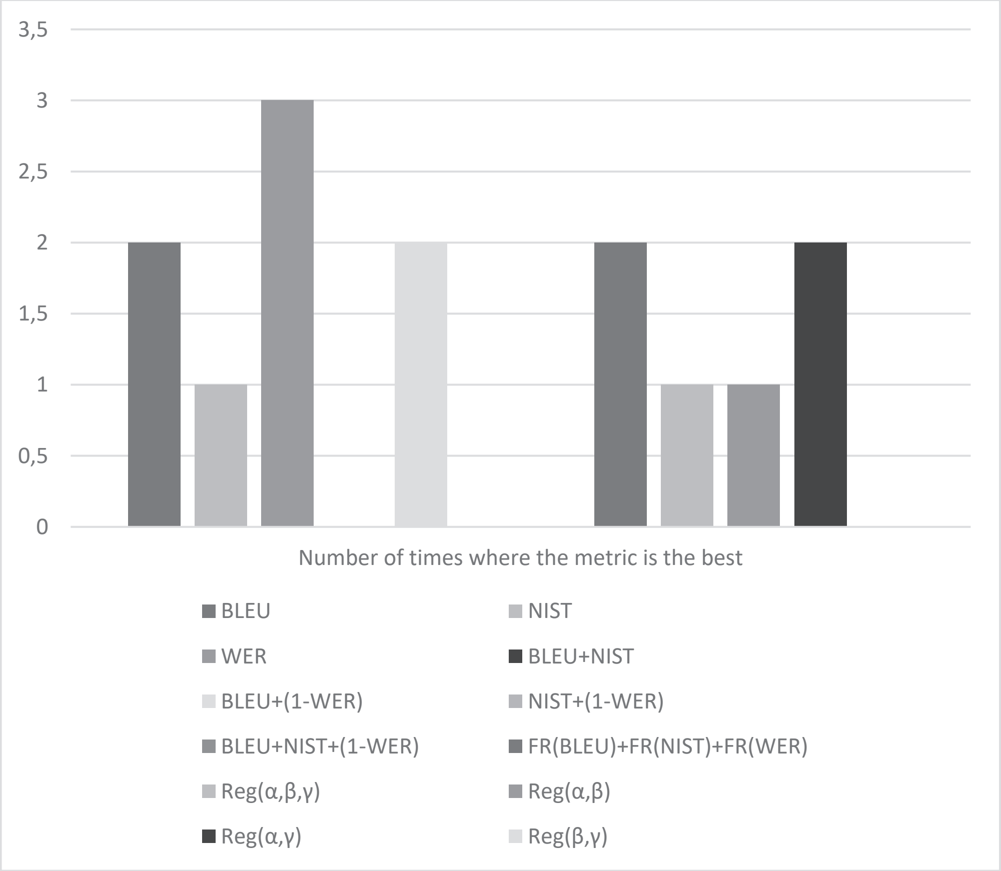

Figure 4. The number of times where the metric, from those considered and their combinations, is the best, in the case of 'From French to English'

5.4.2. From French to English

- (FR(BLEU)+FR(NIST)+FR(WER)) outperforms the others followed respectively by Reg(α,β), BLEU and NIST. Unfortunately, all the other combinations downgrade the performances regarding to the primitive metrics especially BLEU and NIST.
- In terms of number of times where the formula is the best, WER is the best with (three times followed by *BLEU*, *(BLEU+(1-WER))*, *FR(BLEU)+FR(NIST)+FR(WER))*, and *Reg(α,γ)*. Unfortunately, *(BLEU+NIST)*, *(NIST+(1-WER))*, and *(BLEU+NIST+(1-WER))* are the formulas where there is no time to be the best.
- Although the scope of this work is to test the possible different combinations of the basic metrics for machine translation systems, it is worthy to note that the rank in terms of performance for the three considered primitive metrics is as follows: *BLEU*, *NIST*, and *WER*.

6. Conclusion

In this paper, we have tested the effectiveness of the combination for the basic machine translation metrics. Two kinds of combination have been considered: a simple combination with the same implicit parameter weight for each primitive adopted metric, and a combination with regression which designates the parameter weights for each considered basic metric. According to the results obtained, combination may improve performance compared with the basic metrics but it is not always the case. Indeed, some combinations may downgrade basic performance but there are always some ones which upgrade it. There is no specific combination which guarantees performance improvement in both senses ‘from English to French’ and vice versa. In total, combinations based on regression, which relies on a machine learning collection, are very promising in both senses. Unfortunately, we have considered here only three basic machine translation metrics with only one evaluation metric namely NDCG. In the future works, we hope to adopt more primitive machine translation metrics as well as more evaluation metrics and not only *NDCG*.

References

[1] Chauhan, S., & Daniel, P. (2022). A Comprehensive Survey on Various Fully Automatic Machine Translation Evaluation Metrics. Neural Processing Letters, 1-55.

[2] Przybocki, M., Peterson, K, & Bronsart, S. (2008). Metrics for Machine Translation Challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results.

[3] Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., & Zaidan, O. (2010). Findings of the 2010 joint workshop on smt and metrics for machine translation. In Proceedings of the Joint Fifth Workshop on SMT and MetricsMATR, pages 17-53, Uppsala, Sweden. Association for Computational Linguistics.

[4] Albrecht, J. S., & Hwa, R. (2008). Regression for machine translation at the sentence level. Machine Translation, 22(1), 1-27.

[5] Lita, L. V., Rogati, M., & Lavie, A. (2005, October). Blanc: Learning evaluation metrics for mt. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 740-747).

[6] Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering, 26(2), 137-161.

[7] Ma, Q., Graham, Y., Wang, S., & Liu, Q. (2017, September). Blend: a novel combined MT metric based on direct assessment--CASICT-DCU

submission to WMT17 metrics task. In Proceedings of the second conference on machine translation (pp. 598-603).

[8] Mosbah, M. Web Meta-Machine Translation System based on Voting Algorithm. International Journal of Web Science (In Press).

[9] Makhoul, J., Kubala, F., Schwartz, R., & Weischedel, R. (1999, February). Performance measures for information extraction. In Proceedings of DARPA broadcast news workshop (Vol. 249, p. 252).

[10] Tate, C., & Voss, C. (2006). Combining evaluation metrics via loss functions. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 242-250).

[11] Paul, M., Finch, A., & Sumita, E. (2012). Predicting human assessment of machine translation quality by combining automatic evaluation metrics using binary classifiers. International Journal of Computer Applications, 59(10), 9581-406.

[12] Rivera-Trigieros, I. (2022). Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, 56(2), 593-619.

[13] https://www.translate.google.com visited on September 07th 2021.

[14] Dabre, R., Chu, C., & Kunchuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5), 1-38.

[15] https://www.online-translator.com/traduction visited on September 07th 2021.

[16] https://www.babylon-software.com/?lang=fr visited on September 07th 2021.

[17] https://www.bing.com/translator visited on September 07th 2021.

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

[19] Doddington, G. (2002, March). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research (pp. 138-145).

[20] Qin, Y., Q., & Wang, J. (2009, September). Automatic evaluation of translation quality using expanded N-gram co-occurrences. In 2009 International Conference on Natural Language Processing and Knowledge Engineering (pp. 1-5).

[21] Vidal, E. (1997, April). Finite-state speech-to-speech translation. In 1997 IEEE International Conference on Acoustic, Speech and Signal Processing (Vol. 1, pp. 111-114). IEEE.

[22] Niesen, S., Och, F. J., Leusch, G., & Ney, H. (2000, May). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In LREC.

[23] Tomas, I., Mas, J. A., Casacuberta, F. (2003, April). A quantitative method for machine translation evaluation. In Proceedings of the E ACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable? (pp. 27-34).

[24] Järvelin, K., Kekäläinen, J. Cumulative Gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS). 2002; 20(4): 422-446.

[25] Mousavi, S. M. N., & Nagy, J. (2021). Evaluation of plant characteristics related to grain yield of FAO410 and FAO340 hybrids using regression models. Cereal Research.