CMT: Predictive Machine Translation Quality Evaluation Metric

Michal Tyszkowski, Dorota Szaszko

Centrum Lokalizacji CM
Parkowa 19, 51-616, Wroclaw, Poland
E-mail: michal.tyszkowski@cmlocalization.eu, dorota.szaszko@cmlocalization.eu

Abstract

Machine Translation quality is evaluated using metrics that utilize human translations as reference materials. This means that the existing methods do not allow to predict the quality of Machine Translation if no human translations of the same material exist. To use Machine Translation in the translation industry, it is essential to have a metric that allows to evaluate the quality of a newly created Machine Translation in order to decide whether or not it can increase productivity. As a translation company that uses Statistical Machine Translation to generate translation memories for many projects, we decided to develop a metric that can predict its usability on a project basis. This metric, called CMT, is a combination of human assessment and statistics comparable to existing metrics. Our investigations showed that the mean difference between CMT and BLEU is 9.10, and between CMT and METEOR it is 9.69, so it correlates with the existing metrics in more than 90%. CMT is very easy to use and allows to evaluate each translation memory, regardless of its size, in 5 minutes without any reference material.

Keywords: Machine Translation, MT evaluation, CMT metric, predictive MT evaluation metric

1. Introduction

Machine Translation at its early stage of development was mainly used for investigation purposes. The aim of these investigations was to assess the similarity of translations produced by computers with human translations. To make the assessments comparable, a number of metrics were developed, amongst which BLEU (Papineni, Roukos, Ward, & Zhu, 2002), NIST (Doddington, 2002) and METEOR (Lavie, Sagae & Jayaraman 2004) are most commonly used. All these metrics show better or worse the similarity between machine and human translation, but they are retroactive. This means that they are calculated against reference translations after both machine and human translation are completed, which is correct for the purpose they were created.

Machine Translation, however, is not used only for investigation purposes any more. It is now used also as a tool for rough understanding of a text written in an unknown language, and also as a productivity tool for technical translations. Retroactive evaluation metrics is useless for the latter purpose, because the translation industry needs to know whether or not Machine Translation has an acceptable quality before it is used.

Translation buyers often request using Machine Translation for their projects to get lower prices. MT proposals are useful and really increase productivity only if their quality is high, meaning that their post-editing requires less time than translating from scratch. Our experience shows that this condition is fulfilled if the average BLEU score of MT-generated database is not less than 50. The existing evaluation metrics allow to calculate this score after the human translation or post-editing is completed, however, we need to estimate it in advance to be able to decide if a customer's database is really usable and speeds up the process or it is better not to use it and translate from scratch. This is the reason for developing our own tool that could estimate Machine Translation quality before it is used and before the reference human translations exist.

This paper describes briefly the main concept of the predictive MT evaluation methodology and presents the results of a comparison to the metrics calculated after post-editing that shows the high efficiency of our tool.

2. Methods

The investigation required 3 components:

2.1 Machine Translation Engine

An engine to generate MT proposals for as many translation projects as possible and use it for research purposes.

NiuTrans (Xiao, Zhu, Zhang & Li, 2012) an Open Source solution, was used as a Statistical Machine Translation engine. Based on the translation projects completed by our company, 2 corpora were prepared – the first one for IT translations, containing about 30 million of words, and the second one for medical translations, with about 15 million words. For both corpora, a phrase-based translation model was trained. Selected translation projects in the field of IT and medicine from last year were subjected to an MT-evaluation procedure that consisted in producing MT proposals for all segments that did have no matches in our Translation Memories, evaluating these proposals with the use of our proprietary tool, and, depending on the score they achieved, incorporation of MT proposals during the translation phase.

2.2 Estimation Program

A computer program that estimates the quality of a newly generated MT database and calculates a score that can be then compared to a BLEU score or other metrics.

As already stressed in the introduction, no predictive MT evaluation metrics exists, so there is also no program that could calculate it. Our task was to create both: the metric and the program. We assumed that there is no possibility to create a fully automated metric that can evaluate translation quality with no reference. In fact, the existing metrics do not tell anything about translation quality, because it can only be assessed by a human. The metrics only measure the distance between a machine-translated and

human-translated sentence. If the reference translation is wrong, the BLEU value can be high, but the real quality is low. That is why we decided to base our tool on a human judgement, but built into software and rendered as a numeric value.
In our solution, a source text and machine-translated target are displayed in two columns. An evaluator looks at the source and target sentence and decides whether the target is:
- Correct (no errors) – the segment does not get any score,
- Almost correct (1–2 minor errors) – the segment scores 1 point,
- Acceptable (more errors, but still easily understandable) – the segment scores 2 points,
- Unacceptable – the segment scores 3 points.

The decision is taken by clicking one of 4 buttons. From our experience, an evaluator needs about 5 seconds per sentence to evaluate it. The sentences are chosen randomly, based on a selected number of words. For our investigation, 500 words per project were chosen.
The tool calculates the final score using the following formula:

$$100 - \frac{100 \sum s}{smax\, n}$$

where:
s is a segment score with a value from 1 to 3
smax is a max segment score

n is a number of evaluated segments

2.3 Score Evaluation Program
A computer program that calculates MT evaluation metrics using our MT database and translated targets as a reference. To prove the usability of our score, it was compared with the results of MT evaluation performed using two metrics: BLUE and METEOR. For this investigation, we used machine-translated sentences matched with the same sentences translated by humans. 60 data sets saved as plain text files with sentences separated by line feed characters were processed by the iBLEU 2.6.2 (Madnani, 2011) program for BLEU metric calculation and METEOR 1.5 (Denkowski & Lavie, 2014) for METEOR calculation.
For each data set, a difference between our predictive metric and BLEU and METEOR metrics was calculated, as well as a mean value and standard deviation. The results are described below.

3. Results

The results of our research on comparison between CMT and known metrics BLEU and METEOR are shown in the Table 1:

		BLEU		METEOR			
Project	CMT	iBLEU	Delta BLEU	METEOR	Delta METEOR	Usability diff.	Number of words
IT01	32.70	17.90	14.80	24.96	7.74	No	1132
IT02	52.08	42.84	9.24	50.03	2.05	**Yes**	1671
IT03	56.25	35.65	20.60	39.92	16.33	**Yes**	9832
IT04	22.52	27.51	4.99	33.46	10.94	No	1491
IT05	18.33	23.60	5.27	29.37	11.04	No	912
IT06	33.33	22.20	11.13	30.52	2.81	No	939
IT07	19.61	17.12	2.49	25.75	6.14	No	228
IT08	20.16	27.07	6.91	32.72	12.56	No	1071
IT09	25.00	21.82	3.18	29.91	4.91	No	673
IT10	23.76	27.91	4.15	36.14	12.38	No	543
IT11	36.84	30.17	6.67	39.66	2.82	No	715
IT12	76.42	54.57	21.85	65.68	10.74	No	2343
IT13	62.37	41.62	20.75	45.78	16.59	**Yes**	8161
IT14	23.42	37.88	14.46	46.21	22.79	No	467
IT15	23.53	25.30	1.77	29.74	6.21	No	5407
IT16	52.52	29.41	23.11	39.91	12.61	**Yes**	567
IT17	73.33	68.38	4.95	71.78	1.55	No	7654
IT18	67.42	67.97	0.55	69.35	1.93	No	3423
IT19	71.27	60.61	10.66	64.40	6.87	No	6000
IT20	60.00	59.37	0.63	64.47	4.47	No	19,314
IT21	61.36	57.20	4.16	62.41	1.05	No	10,943
IT22	71.32	60.63	10.69	63.47	7.85	No	38,020
IT23	57.62	66.6	8.98	70.44	12.82	No	6099
IT24	47.62	43.2	4.42	49.14	1.52	No	1934
IT25	52.71	54.09	1.38	55.23	2.52	No	3051
IT26	61.81	59.05	2.76	62.55	0.74	No	446
IT27	48.41	47.03	1.38	53.88	5.47	No	5717
IT28	42.64	48.14	5.50	53.47	10.83	No	11,364
IT29	50.17	59.47	9.30	65.27	15.10	No	2861
IT30	50.00	58.56	8.56	60.12	10.12	No	3423
IT31	61.81	70.00	8.19	68.70	6.89	No	543
MED01	59.03	45.84	13.19	50.03	9.00	**Yes**	6708
MED02	41.18	38.62	2.56	45.35	4.17	No	3813

		BLEU		METEOR			
Project	CMT	iBLEU	Delta BLEU	METEOR	Delta METEOR	Usability diff.	Number of words
MED03	29.17	30.50	1.33	45.47	16.30	No	1217
MED04	17.65	26.84	9.19	32.65	15.00	No	9757
MED05	20.00	47.87	27.87	42.94	22.94	No	2076
MED06	35.24	31.37	3.87	36.96	1.72	No	1410
MED07	11.90	31.20	19.30	41.69	29.79	No	494
MED08	22.76	22.25	0.51	31.52	8.76	No	11,702
MED09	43.17	45.78	2.61	49.98	6.81	No	3119
MED10	39.61	43.88	4.27	48.20	8.59	No	3020
MED11	37.25	24.06	13.19	31.80	5.45	No	1054
MED12	40.14	36.52	3.62	42.94	2.80	No	1278
MED13	43.21	34.71	8.50	38.39	4.82	No	1238
MED14	39.13	46.51	7.38	50.24	11.11	No	1809
MED15	35.83	49.21	13.38	55.03	19.20	No	1107
MED16	33.33	45.23	11.90	52.07	18.74	No	2995
MED17	25.68	35.57	9.89	32.38	6.70	No	958
MED18	21.90	31.47	9.57	34.23	12.33	No	1958
MED19	54.50	39.04	15.46	42.10	12.40	**Yes**	3713
MED20	21.11	22.37	1.26	26.47	5.36	No	2152
MED21	25.49	29.71	4.22	31.48	5.99	No	373
MED22	6.06	24.74	18.68	32.74	26.68	No	2744
MED23	57.00	28.02	28.98	34.19	22.81	**Yes**	7111
MED24	24.76	25.96	1.20	28.00	3.24	No	836
MED25	29.00	21.55	7.45	26.30	2.70	No	6746
MED26	25.93	30.63	4.70	35.51	9.58	No	51,802
MED27	9.76	17.92	8.16	20.33	10.57	No	2375
MED28	19.19	32.08	12.89	25.81	6.62	No	12,375
MED29	57.75	30.28	27.47	34.95	22.80	**Yes**	12,875
Mean value			9.10		9.69	Total words:	315,759
Standard deviation			7.30		6.90		

Table 1. Comparison between CMT, BLEU, and METEOR metrics

31 IT and 29 medical translation projects were used for creating a machine-generated translation memory which was evaluated using the CMT metric. After human translation of these projects, the BLEU and METEOR metrics were calculated using human translation as a reference. Then, the results of the CMT, BLEU, and METEOR metrics were calculated by subtracting the values.

The highest difference between CMT and BLEU was 29.98, and the lowest difference was 0.51. The mean value of this difference was 9.10 and the standard deviation was 7.30.

The highest difference between CMT and METEOR was 29.79, and the lowest difference was 0.74. The mean value of this difference was 9.69 and the standard deviation was 6.90.

Apart from the difference between metrics, we also checked in how many cases the decision about the usability of Machine Translation for post-editing taken on the basis of the CMT metrics appeared to be wrong. As already mentioned, our practice shows that it is reasonable to use Machine Translation as a productivity tool only if the BLEU score is not less than 50. This threshold was obtained in an empiric way. Translators have always choice either post-edit an MT proposal or translate the sentence from scratch. While analyzing sentences that translators post-edited we noticed that their BLEU score was never below 50. This means that sentences with lower quality were not used for post-editing but translated from scratch. Because the CMT score corresponds to BLEU, the machine generated translations were used for post-editing only if the CMT score was 50 or more. After calculating the BLEU score, it appeared that our decision was wrong only in 8 cases and it was right in 52 cases.

We also investigated whether or not the good results could be accidental. To verify this, the distribution of values was examined. The results are shown in Figure 1. The graph shows that the majority of values are in the ranges 0–2.8, 2.9–5.6, and 8.5–11.2, so the shape of this graph is far from the Gaussian curve.

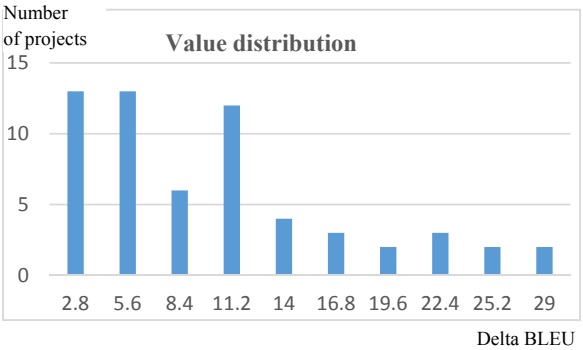

Figure 1. The distribution of the values of the difference between the CMT and BLEU scores

The last aspect that was checked was the dependency between the number of machine-translated words and the difference between the CMT and BLEU score. The results are illustrated in Figure 2. The graph shows that there is no significant dependency between these values, as the curves

have completely different shapes.

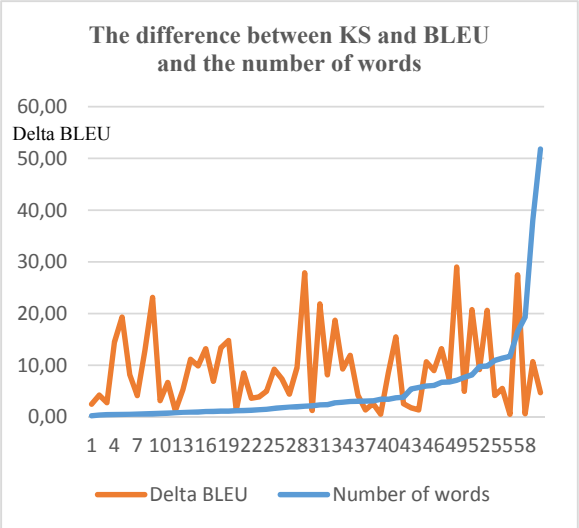

Figure 2. The dependency between number of words in MT and the difference between the CMT and BLEU scores

4. Conclusions

CMT is the first predictive MT evaluation metrics which means that it is designed to evaluate the usability of Machine Translation without any reference material.

CMT is calculated with a dedicated software utilizing an algorithm that uses a human judgement and statistics. The value of CMT ranges from 0 to 100, which makes it comparable to known MT evaluation metrics such as BLEU and METEOR.

The software used for CMT calculation allows to evaluate any Machine Translation in about 5 minutes, regardless of its size.

The comparative research conducted using 60 translation projects with a total wordcount of 315,759 words showed that the mean difference between CMT compared to BLEU and METEOR was below 10 (9.10 for BLEU and 9.69 for METEOR), which means that the correlation between CMT and the metrics calculated using human translations as reference is above 90%.

The correlation between CMT, BLEU, and METEOR does not depend on the number of words evaluated and the most values placed in range 0–8.4, which means that the distribution is not normal, but shifted towards the smallest values.

Unlike BLEU, METEOR, and other retroactive metrics, CMT does not rely on the quality of reference materials, so it is much more comparable to the human judgement.

CMT is a score that can be used in the translation industry to support the decision whether or not it is reasonable to use Machine Translation as a productivity tool for a given translation project.

5. References

Denkowski M. and Lavie A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language, *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*

Doddington, G. (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Human Language Technology Conference (HLT)*, San Diego, CA pp. 128–132

Lavie, A., Sagae, K. and Jayaraman, S. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation *in Proceedings of AMTA 2004*, Washington DC. September 2004

Madnani, N. (2011). iBLEU: Interactively Debugging & Scoring Statistical Machine Translation Systems *in Proceedings of the fifth IEEE International Conference on Semantic Computing*, Sep 19-21, 2011

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation *in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311–318

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. (2012). NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation. *In Proc. of ACL, demonstration session*