Unsupervised Measurement of Translation Quality using Multi-Engine, Bi-Directional Translation

Menno van Zaanen and Simon Zwarts

Division of Information and Communication Sciences
Department of Computing
Macquarie University
2109 Sydney, NSW, Australia
{menno, szwarts}@ics.mq.edu.au

Abstract. Lay people discussing machine translation systems often perform a round trip translation, that is translating a text into a foreign language and back, to measure the quality of the system. The idea behind this is that a good system will produce a round trip translation that is exactly (or perhaps very close to) the original text. However, people working with machine translation systems intuitively know that round trip translation is not a good evaluation method. In this article we will show empirically that round trip translation cannot be used as a measure of the quality of a machine translation system. Even when using translations of multiple machine translation systems into account, to reduce the impact of errors of a single system, round trip translation cannot be used to measure machine translation quality.

1 Introduction

The on-going development of new Machine Translation (MT) systems requires a solid evaluation framework that allows for comparison of the improvements of these systems. Evaluation of MT systems is therefore an important aspect in the field of MT research.

Evaluating MT systems is non-trivial. When several human translators translate a text from one language to another, there is a high likelihood that the resulting translations are not exactly the same. There are multiple ways of translating one text. Among other reasons, this has to do with differences of the two languages. Often, no one-to-one mapping of sentences in the two languages exists. The meaning of words is almost never exactly the same and ambiguity in the sentences can be perceived differently by different readers (and translators).

Research in MT system evaluation concentrates on finding a fair way to measure the quality of MT systems. If a system is perceived to be better than another system by a user, an evaluation metric should also reflect this. This is not necessarily easy, as many different translations for one text exist and there is no one correct answer.

Automatic evaluation (where the computer compares the translation against one or more human translations) has advantages over human evaluation (where people directly analyse the outcome of MT systems). Not only is automatic evaluation much cheaper and less expertise intensive (human translations only need to be produced once and can be reused afterwards), it is also faster.

In this article, we investigate how far fully unsupervised, automatic MT evaluation can be taken. Unsupervised evaluation means that we evaluate using no human-made translation of any text. This requirement is much stricter than that of existing measures, that need one or more human translations of the text under investigation.

This article is structured as follows: in section 2, we will briefly discuss current MT evaluation methods. Next, we introduce a new, unsupervised evaluation methods (in section 3) and measure the usefulness of these approaches in section 4. Finally, we will mention possible future work in this direction and draw a conclusion.

2 Evaluating MT systems

There are many approaches to the evaluation of MT systems, ranging from manually analysing MT system output to automatically measuring quality against texts that were translated by humans previously. All current approaches to MT evaluation have their advantages and disadvantages.

With manual evaluation, a text in the original language (i.e. the source text written in the source language) is translated by an MT system into the target language. The output of this system, i.e. the target text, is handed to a human evaluator, who has knowledge of both the source and target languages. By manually examining the translation, an assessment of the quality of the MT system can be made.

The advantage of manual evaluation is that it gives a good idea of the quality of the MT system. The evaluator can precisely pinpoint errors made by the MT system and it also gives an idea of the readability and "naturalness" of the translation. However, manual evaluation is costly and time and expertise intensive. This can be a major drawback when the MT system needs to be tested regularly, for example, when measuring improvements or when evaluating performance on different genres of text.

To reduce the disadvantages of manual evaluation, automatic evaluation methods have been devised. These metrics compare text translated by an MT system against pre-made human translations of the same text. The advantage of this approach is that a text only needs to be manually translated once (by one or more human translators). The translations can be reused in future experiments. Examples of such metrics are BLEU [1], F-Measure [2], and Meteor [3][1].

The underlying idea behind these automatic evaluation approaches is that the metric correlates highly with human evaluation. Only when the correlation

[1] In this article, we used the implementations of the BLEU score and the F-Measure available from `http://www.ics.mq.edu.au/~szwarts/Downloads.php`.

between the automatic evaluation and manual evaluation is high, can we assume the final outcomes of automatic evaluation would be (almost) the same as the outcomes of manual evaluation.

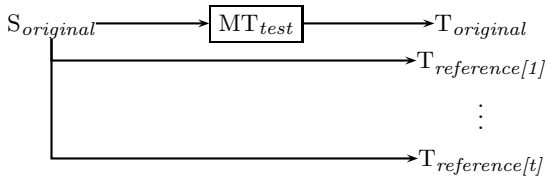

Fig. 1. Translation flow, current evaluation methods

Figure 1 illustrates the traditional automatic evaluation. $S_{original}$ stands for the text the source language. This text is translated using an MT system and is converted into $T_{original}$, i.e. the original text in the target language. The evaluation of the MT system is done by comparing $T_{original}$ to one or more human made reference texts (illustrated as $T_{reference[x]}$ in the figure).

The automatic evaluation approach seems to overcome most of the problems of the human evaluation approach. However, this approach also has problems that need to be addressed. Most automatic evaluation metrics perform best when comparing the MT system translation against multiple human reference translations at the same time. There is a trade-off between the amount of data and the number of reference texts [4]. In practice, however, it is hard to obtain multiple translations of the same text. To make things worse, the performance of MT systems depends highly on the tested domain, i.e. systems can perform quite well on one genre but completely fail to perform adequately on another. This means that one cannot simply pick a random text for which (multiple) translations are available. The tested domain also needs to be taken into account.

In this article, we address the dependency of the automatic evaluation on reference translations. If we can remove the requirement of multiple reference translations (of multiple texts in different genres), evaluation of MT systems will be much more flexible. In such a setting, translation quality of new genres or even particular texts can easily be measured automatically.

3 Unsupervised MT system evaluation

Currently, there are several web-based MT systems freely available on the Internet. Not only are these systems used by people to translate their webpages, these systems also allow for interesting experiments, such as multi-engine MT systems. These systems try to average the translation quality of different translations engines (like Democrat [5]). The idea is that, on average, combining the output of multiple MT systems is better than the output of only one MT system.

There are several ways of evaluating MT systems in an unsupervised way. Here we will discuss three different approaches. The first translates a text from the source to the target language and tries to evaluate the system based on that translation. The other two approaches rely on round-trip translations, where the

source text is translated in the target language and then translated back to the source language. The evaluation takes place on the source language. All of these approaches will be treated in more detail.

3.1 One way translation

In the one-way quality assessment approach, we evaluate based on the outputs of different MT systems in the target language. The idea is that we can use the different outputs as reference translations with respect to the system that we are trying to measure. This means that we do not require any pre-translated text at all. In this approach, we use $c + 1$ different MT systems, which leads to one text under investigation and c reference texts.

Figure 2 illustrates the one-way translation approach. The source text that will be used to evaluate is translated into the target language using the both MT system to be evaluated and several "reference" MT systems. The output of the reference MT systems is used as reference translations, similar to the process described in Figure 1. This means that for this new approach the traditional automatic evaluation tools can be used.

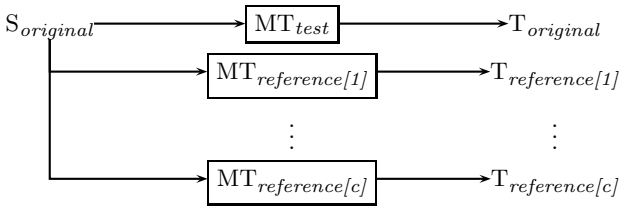

Fig. 2. Translation flow, one way translation

3.2 Round trip translation

Evaluation using "round trip translation" (RTT) is very popular among lay people. The idea here is that a text in the source language is translated into the target language and that text is then translated back into the source language again. The evaluation is performed by comparing the original text with the back translated text, all in the source language. Figure 3 illustrates the idea. Here, $S_{original}$ is compared against S_{RTT}.

Fig. 3. Translation flow, round trip translation

The advantage of RTT compared to other unsupervised evaluation methods is that we can use the original text as a reference text, which we know is a correct "translation". This is different from the other approaches, where the evaluation is performed on texts translated by other (unreliable[2]) MT systems.

[2] The systems are unreliable in that their quality has not been measured on the text that is being evaluated.

The assumption behind RTT is that the closer the S_{RTT} is to $S_{original}$, the better the MT system is. The underlying idea is that if the MT system translates the original text well and the back translation is also good, the two texts should be equivalent. Measuring the quality of the back translation against the original text will then give a measure of the forward translation.

There are, however, reasons to believe why RTT does not work. MT systems often make the same mistakes in the forward translation as in the back translation. For example, if the MT system completely failed in putting the words in the right order in the target language and again failed to reorder it during the back translation, the words in the round trip translation are still in the right order, even though the translation in the target language is incorrect.

RTT also gives incorrect results when the two translation steps are asymmetrical. For example, the forward translation may be perfect, but only the back translation is incorrect. This leads to a low RTT evaluation score, even though the forward translation is correct.

The effectiveness of RTT has been investigated earlier [6] and the results presented there are not promising. It seems that RTT evaluation has little correlation to evaluation based on human translations. Here, we extend this research. We aim to reduce the impact of errors of the back translation step by using more MT systems. This is done in the multi-engine RTT evaluation as described in the next section.

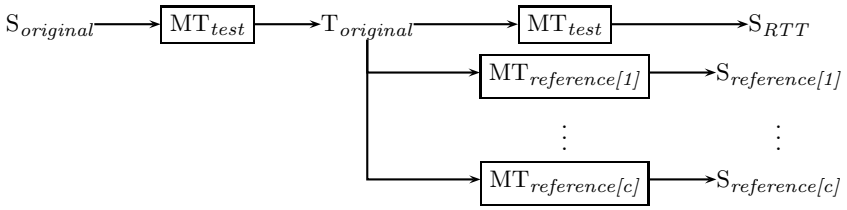

Fig. 4. Translation flow, single round trip translation

3.3 Multi-engine round trip translation

To reduce the impact of the errors made by the MT system in RTT, we also investigated approaches using multiple MT systems. We identified two of these approaches, which we call *single RTT* and *multi RTT* (in contrast to the "standard" RTT, which we will call *default RTT* here).

Figure 4 illustrates single RTT evaluation. The translated text is back translated using several MT systems. S_{RTT} is now evaluated with the output of the reference systems ($S_{reference[x]}$) as reference translations.

Multi RTT (see Figure 5) is quite similar to single RTT. The idea is to generate even more reference texts by performing forward translations as well as back translations using multiple systems. The added amount of data allows us to average the translations to reduce the impact of the translation mistakes of the individual systems (in both the forward and back translation steps). Using $c + 1$ translation engines (including the one that is being tested) we generate $(c + 1)^2 - 1$ reference texts and compare these against one candidate text.

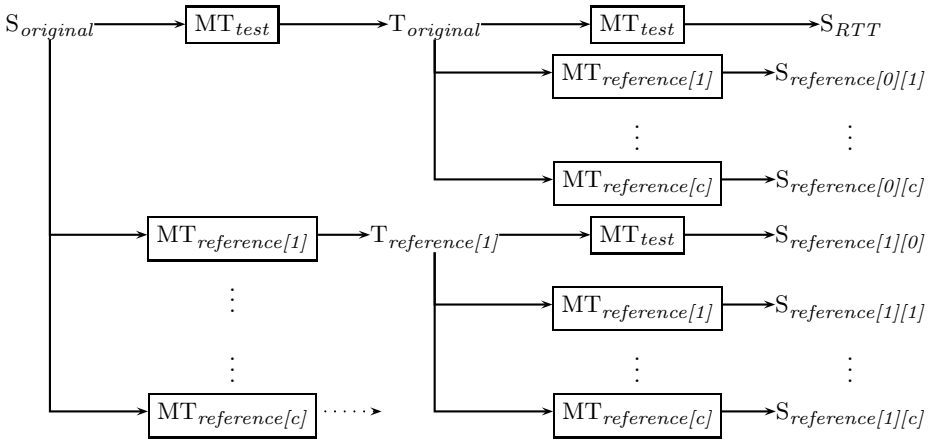

Fig. 5. Translation flow, multi round trip translation

4 Experiments

We are interested in how effective the unsupervised evaluation methods (one-way translation, default RTT, single RTT, and multi RTT) are at measuring the quality of a single MT system without having to rely on human translated texts. To measure this, we have conducted several experiments that measure the performance of the unsupervised methods and we have compared these against a control experiment on the same texts. The control experiment is simply the approach depicted in Figure 1, i.e. comparing the output of the MT system against human translated reference texts.

A method is successful if the result of the unsupervised evaluation method correlates highly with the results of the control experiment. We assume that the results of the control experiment have a high correlation with human translations.

4.1 Setting

There are several parameters that need to be set before the actual experiments can take place. First of all, MT systems behave differently on different types of texts. One translation engine may achieve high quality translations on technical text, while another may perform better on more informal texts. To reduce this impact, we have conducted the experiments on texts from different genres.

The choice of languages may also influence translation quality. We have used different language pairs to restrict the influence of language choice. Texts are translated from English to German or from French to English.

We used the same 500 sentences that were used for the evaluation of the Democrat [5] system:

euEG, euFE English to German and French to English translations of the Europarl corpus [7]. These, quite formal texts, are extracts from sessions of the European Parliament over a couple of years;

maFE parts of a tourism website of Marseilles available in French and English translations are available. This text is less formal than the Europarl corpus;
miFE parts from the novel *20.000 Leagues Under the Sea* by Jules Verne. The original French version and an English translation are used.

The different MT systems that are used in this article are listed in the appendix. It is not important to know which MT system produced which output, so we assigned numbers to all of them.

4.2 Empirical results

To find out whether unsupervised automatic MT evaluation is effective, we correlate the output of the different unsupervised methods with the standard evaluation method as described in Figure 1. Finding a high correlation between the unsupervised evaluation and the standard evaluation indicates that the unsupervised method provides similar results and is as usable.

We compute the BLEU scores of the different MT systems by comparing them against the human translated reference texts, which serves as the control data. We use the BLEU metric with n-gram sizes up to 4. The results can be found in Table 1.

Table 1. BLEU scores for the different MT-systems and corpora

MT system	euEG	euFE	maFE	miFE
system 1	0.110304	0.186572	0.186828	0.140024
system 2	0.106163	0.178278	0.230331	0.180803
system 3	0.105912	0.181388	0.205101	0.139894
system 4	0.100855	0.141064	0.147441	0.145592
system 5	0.095173	0.188239	0.179050	0.160152

Table 2. Formulas used to compute the measures

A.1 $\textsc{Bleu}(\mathrm{T}_{original}, \mathrm{T}_{reference[1]}, \ldots, \mathrm{T}_{reference[c]})$

A.2 $\frac{1}{c}\sum_{i=1}^{c} \textsc{Bleu}(\mathrm{T}_{original}, \mathrm{T}_{reference[i]})$

B $\textsc{Bleu}(\mathrm{S}_{RTT}, \mathrm{S}_{original})$

C.1 $\frac{1}{c}\sum_{i=1}^{c} \textsc{Bleu}(\mathrm{S}_{reference[i]}, \mathrm{S}_{original})$

C.2 $\textsc{Bleu}(\mathrm{S}_{RTT}, \mathrm{S}_{reference[1]}, \ldots, \mathrm{S}_{reference[c]})$

C.3 $\frac{1}{c}\sum_{i=1}^{c} \textsc{Bleu}(\mathrm{S}_{RTT}, \mathrm{S}_{reference[i]})$

D $\frac{1}{(c+1)^2-1}\sum_{i=0,j=0}^{c,c} \textsc{Bleu}(\mathrm{S}_{RTT}, \mathrm{S}_{reference[i][j]})$, where $i \neq 0 \vee j \neq 0$

E $\mathrm{stddev}_{i=0,j=0}^{c,c}\textsc{Bleu}(\mathrm{S}_{RTT}, \mathrm{S}_{reference[i][j]})$, where $i \neq 0 \vee j \neq 0$

As is seen in Table 1 there is no one MT system that consistently performs best (on different language pairs and genres). For example, system 5 ranks highest on French to English (euFE) but lowest on English to German (euEG). The unsupervised evaluation metrics should reflect these differences as well.

The first experiment of unsupervised evaluation measures the effectiveness of one-way translation as described in Figure 2. We calculate the BLEU score of the translation of the original text ($T_{original}$) with respect to the reference translations generated by the five MT systems ($T_{reference[x]}$ with $x = 1 \ldots 4$). This is computed using formula A.1 of Table 2[3].

In experiment A.2 we compute the BLEU score of the translation with respect to each of the single reference translations, which results in four BLEU scores. BLEU expects all reference texts to be correct, which is certainly not the case here. Hence we propose averaging individual BLEU scores here.

Experiment B evaluates the default round trip translation as described in Figure 3. Since only one MT system is used in the evaluation (the same system is used to translate to the target language and back), this is essentially a reproduction of Somers's experiments [6]. Although he does not publish exact figures, we suspect that he used n-grams with n=1 only, whereas we use n=4.

Table 3. Correlations of all experiments (formulas given in Table 2)

Exp.	**Corr.**	**Exp.**	**Corr.**	**Exp.**	**Corr.**	**Exp.**	**Corr.**	**Exp.**	**Corr.**
A.1	0.295	B	0.193	C.1	-0.063	D	0.341	E	0.394
A.2	0.441			C.2	0.532				
				C.3	0.401				
one-way		*default RTT*		*single RTT*		*multi RTT*			

In experiment C we measure the effectiveness of the single RTT approach (Figure 4). We first compute the average BLEU score of the reference back translations with respect to the original text (C.1) and also compute the BLEU score of the round trip translation of the test system and use the round trip translations of the reference systems as reference texts (C.2). C.3 (like A.2) computes the average BLEU score using each of the reference texts separately.

Experiment D is similar to experiment C, only in this case multi RTT is evaluated (Figure 5). This means that more data points are available, because the forward translation step is also performed by several MT systems (in contrast to single RTT, where only one MT system is used in the forward translation). Due to space restrictions, we only show the results of the average BLEU score here. The other results are similar to those in experiment C.

Finally, we realised that the average of the evaluation metrics might not be a good indication for MT quality. We are more interested in how similarly the MT systems perform. Consequently, instead of comparing the actual BLEU scores, we need to measure the standard deviation between the BLEU scores of the systems. Experiment E computes this.

Looking at the results presented in Table 3, we see that none of the experiments show a correlation that is high enough to use for a final MT quality assessment. The highest correlation is found in experiment C.2, where the round

[3] Note that the names of the texts in the formulas refer to the names in the figures corresponding to the setup of the experiment.

trip translation of the tested system is compared against all other round trip translations based on the forward translation of the tested system. Even the correlation of experiment D, which uses more data points (and as such may have given us more precise results) is lower.

We performed several more experiments. We wondered whether the size of the n-gram used in BLEU has any influence on the correlation, so we also ran the experiment D for n-gram sizes 6, 8, 9, and 12. It is interesting to see that increasing n also increases the correlation. The highest correlations were found with n=9. However, the shortest sentence is 9 tokens, which meant that increasing the size of n introduces smoothing errors, which is also reflected on the lower correlation of n=12.

Choosing BLEU as the evaluation metric may be somewhat arbitrary. We also calculated the results of experiment D using the F-measure[4]. The correlation between the BLEU score and the F-measure was 0.960. As such, the correlations of all experiments are quite similar, which means that the choice between these metric is unimportant here. And indeed the F-measure in all our results followed the trend with BLEU quite precisely.

5 Conclusion

In this article, we investigated unsupervised evaluation methods in the context of machine translation. Evaluation of MT systems is hard, for several reasons. Human evaluation is costly and automatic evaluation requires several reference texts in different genres and language pairs. Unsupervised evaluation is an automatic evaluation that removes the requirement of human translated reference texts by evaluating on texts that are translated from one language to another and back.

Unfortunately, the experiments show that unsupervised evaluation (including round trip translation which is often used by lay people) is unsuitable for the selection of MT systems. The highest correlation we found was 0.532. We think that this is not enough to base a clear verdict on. (Note that BLEU scores computed between MT system output and human made reference texts have a reported correlation between 0.96 and 0.99 against human evaluation of MT systems [1].) Essentially, RTT should never be used to illustrate MT system quality.

The research described here covers most possible combinations of one way and round trip translation using one or more MT systems. At the moment, we believe that unsupervised evaluation is not suitable to measure the quality of MT systems. However, it might be desirable to repeat these experiments in the future, when more systems are available.

The evaluation methods were compared against reference translations that correspond to "good" translations. However, perhaps the metrics should measure a form of readability. Even though the output of some MT systems may be worse

[4] BLEU has n-gram size as a parameter. F-measure also has a non-trivial parameter, the exponent. We used $e = 2$.

than others, the readability and understandability of the translations may be better. None of the current metrics addresses this issue and more research needs to be done in this area.

Bibliography

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Jul 2002. URL `http://www1.cs.columbia.edu/nlp/sgd/bleu.pdf`.

[2] Joseph P. Turian, Luke Shen, and I. Dan Melamed. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*, 2003. URL `http://nlp.cs.nyu.edu/publication/papers/turian-summit03eval.pdf`.

[3] Santanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for MT Evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, June 2005.

[4] Ying Zhang and Stephan Vogel. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, October 2004.

[5] Menno van Zaanen and Harold Somers. DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation. In *Proceedings of MT Summit X*, 2005.

[6] Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop*, 2005.

[7] Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation, 2002. URL `http://people.csail.mit.edu/~koehn/publications/europarl.ps`.

Appendix

URLs for the online MT systems in random order:
`http://world.altavista.com/babelfish`
`http://www.freetranslation.com`
`http://www.translate.ru`
`http://www.worldlingo.com`
`http://www.reverso.net`