

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Machine Translation and the Evaluation of Its Quality

Mirjam Sepesy Maučec and Gregor Donaj

Abstract

Machine translation has already become part of our everyday life. This chapter gives an overview of machine translation approaches. Statistical machine translation was a dominant approach over the past 20 years. It brought many cases of practical use. It is described in more detail in this chapter. Statistical machine translation is not equally successful for all language pairs. Highly inflectional languages are hard to process, especially as target languages. As statistical machine translation has almost reached the limits of its capacity, neural machine translation is becoming the technology of the future. This chapter also describes the evaluation of machine translation quality. It covers manual and automatic evaluations. Traditional and recently proposed metrics for automatic machine translation evaluation are described. Human translation still provides the best translation quality, but it is, in general, time-consuming and expensive. Integration of human and machine translation is a promising workflow for the future. Machine translation will not replace human translation, but it can serve as a tool to increase productivity in the translation process.

Keywords: machine translation, statistical machine translation, neural machine translation, evaluation, post-editing

1. Introduction

Machine translation (MT) investigates the approaches to translating text from one natural language to another. It is a subfield of computational linguistics that draws ideas from linguistics, computer science, information theory, artificial intelligence, and statistics. For a long time, it had a bad reputation because it was perceived as low quality. Especially in the last two decades, we have been witnessing great progress in MT quality, which made it interesting also for the use in the translation industry. Its quality is still lower than human translation, but that does not mean it does not have good practical uses. In the past, translation agencies and other professional translators were the only actors in the translation industry, but, in recent years, we have been faced with the rapidly growing range of machine translation solutions entering the market and being of practical use. There is increasing pressure on the translation industry in terms of price, volume, and turnaround time. The emergence of commercial applications for MT is a welcome change in translation processes. In professional or official circumstances, human translation is inevitable, as humans are essential to making sure a translation is grammatically correct and carries the same meaning as the original text.

Machine translation is appropriate in different circumstances, mainly for unofficial purposes or for providing content for a human translator to improve upon it. MT has proved to be able to speed up the whole translation process, but it cannot replace the human translator. The questions that researchers in the translation industry are trying to answer today are: How much can human translators benefit from using MT? How could MT be integrated efficiently into translation processes? If MT is integrated into the translation workflow, will the quality of translation remain at the same level? These questions will not be answered explicitly in this chapter, but an effort will be made to show that MT is worth being part of the translation process, as its quality can be evaluated reliably. MT opens new opportunities for translators through using MT output only as a suggestion and, if necessary, post-editing it to the desired quality. It could be much faster than translation from scratch. This process is further discussed in the penultimate section of the chapter.

The aim of this chapter is to overview the methods of machine translation and the methods of the evaluation of its quality. This chapter is organised as follows. In Section 2, different approaches for machine translation are described: rule-based MT in Section 2.1, example-based MT in Section 2.2, statistical MT in Section 2.3, hybrid MT in Section 2.4, and neural MT in Section 2.5. Not all languages are equally difficult for MT. Section 3 exposes common problems when dealing with morphologically rich languages. Sections 4–7 are devoted to MT evaluation. In Section 5, basic metrics for automatic MT evaluation are described and in Section 6, the more advanced ones. Automatic MT evaluation makes sense if it gives similar results as manual evaluation. Section 7 discusses how the correlation between automatic and manual MT evaluation is determined. MT is never perfect. Section 8 discusses post-editing MT to correct the mistakes and make MT of practical use. Section 9 concludes this chapter.

2. Machine translation

Computer scientists began trying to solve the problem of MT in the 1950s. The first published machine translation experiment was performed by the Georgetown University and IBM. It involved automatic translation of more than 60 Russian sentences into English. The system had only 6 grammar rules and 250 lexical items in its vocabulary. It was by no means a fully featured system. The sentences for translation were selected carefully, as the idea of the experiment was to attract governmental and public interest and funding by showing the possibilities of MT. Many problems of MT had come to light right after, and, consequently, for a long time, MT was present only as a research area in computational linguistics. Over-time, different approaches for MT were defined and gained maturity for practical use today. The history of the development of MT approaches is given in **Figure 1**. In [1], it has been shown that 22% of the MT users in the translation industry use rule-based MT systems, 50% use statistical MT systems, and 36% of them use hybrid MT systems.

2.1 Rule-based machine translation

The first approaches for MT were based on linguistic rules that were used to parse the source sentence and create the intermediate representation, from which the target language sentence was created. Such approaches are appropriate to translate between closely related languages. The rule-based machine translation methods include dictionary-based MT, transfer-based MT, and interlingual MT.

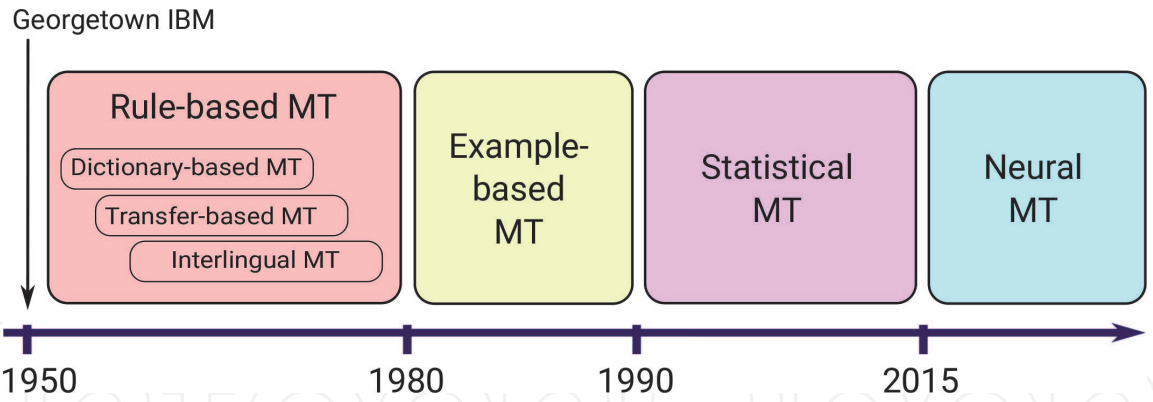


Figure 1.
Timeline of MT evolution.

Dictionary-based MT uses entries in a language dictionary to find a words equivalent in the target language. Using a dictionary as the sole information source for translation means that the words will be translated as they are translated in a dictionary. As this is, in many cases, not correct, grammatical rules are applied afterwards.

Transfer-based MT belongs to the next generation of machine translation. The source sentence is transformed into an intermediate, less language-specific structure. This structure is then transferred into a similar structure of the target language, and, finally, the sentence is generated in the target language. The transfer uses morphological, syntactic, and/or semantic information about the source and target languages.

In interlingual MT, the source sentence is transformed into an intermediate, artificial language. It is a neutral representation that is independent of any language. The target sentence is then generated out of the interlingua.

To be useful in practice, rule-based MT systems consist of large collections of rules, developed manually over time by translation experts, mapping structures from the source language to the target language. They are costly and time-consuming to implement and maintain. As rules are added and updated, there is the potential of generating ambiguity and translation degradation. Rule-based MT requires linguistic experts to apply language rules to the system.

2.2 Example-based machine translation

Example-based MT is based on the idea of analogy. It is grounded upon a search for analogous examples of sentence pairs in the source and target languages. Example-based MT belongs to corpus-based approaches because examples are extracted from large collections of bilingual corpora. Given the source sentence, sentences with similar sub-sentential components are extracted from the source side of the bilingual corpus, and their translations to the target language are then used to construct the complete translation of the sentence.

2.3 Statistical machine translation

Statistical MT is based on statistical methods [2]. It also belongs to corpus-based approaches, as statistical methods are applied on large bilingual corpora. Building a statistical MT system does not require linguistic knowledge. Statistical MT utilises statistical models generated from the analysis of texts, being either monolingual or bilingual. It is called training data. If more training data are available, better and larger MT systems can be built. Statistical MT systems are computationally

expensive to build and store. Statistical MT can be adapted easily to a specific domain if enough bilingual and/or monolingual data from that domain are available.

Statistical MT is defined using the noisy-channel model from the information theory:

$$\mathbf{e} = \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e}). \quad (1)$$

where \mathbf{f} is the source sentence and \mathbf{e} is the target sentence. The source sentence consists of words f_j and the target sentence of words e_i . Words f_j belong to the source vocabulary F and the words e_i to the target vocabulary E . In the phrase-based model, the source sentence f is broken down into I phrases \bar{f}_i , and each source phrase \bar{f}_i is translated into a target phrase \bar{e}_i .

Standard phrase-based SMT models consist of three components:

1. A translation model of phrases (denoted as $\phi(\bar{f}|\bar{e})$). In practice, both translation directions, with the proper weight setting, are used: $\phi(\bar{f}|\bar{e})$ and $\phi(\bar{e}|\bar{f})$.
2. A reordering model (denoted as d). It is based on distance. The reordering distance is computed as $start_i - end_{i-1} - 1$, where $start_i$ is the position of the first word in phrase i , end_{i-1} is the position of the last word of phrase $i - 1$, and d is the probability distribution of reordering.
3. A language model (denoted as $p_{LM}(\mathbf{e})$). It makes the output a fluent sequence of words in the target language and is most commonly an n -gram language model.

Log-linear models of phrase-based SMT are most commonly used:

$$p(\mathbf{e}, a | f) = \exp \left[\lambda_\phi \sum_{i=1}^I \log \phi(\bar{f}_i | \bar{e}_i) + \lambda_d \sum_{i=1}^I \log d(start_i - end_{i-1} - 1) + \lambda_{LM} \sum_{i=1}^N \log p_{LM}(e_i | e_1 \dots e_{i-1}) \right]. \quad (2)$$

where a is an alignment between source and target sentences and N is the length of the target sentence.

Statistical MT faces many obstacles. Data sparsity of highly inflected languages limits the effectiveness of statistical MT. Advanced statistical MT systems try to overcome the limitations by introducing data preprocessing and data post-processing. In **Figure 2** data preprocessing is used, where morphosyntactic tags (MSD) and lemmas are assigned to words and used in translation and language models. Reordering model captures short-term dependencies. The operation sequence model (OSM) is able to capture long-distance dependencies [3]. It models translation by a linear sequence of operations. The operation generates translation, performs reordering, jumps forward and backward, etc. Having morphosyntactic tags and lemmas available, OSM could be constructed based on them, as depicted in **Figure 2**.

2.4 Hybrid machine translation

While statistical methods still dominate research work in MT, most commercial MT systems were, from the beginning, only rule-based. Recently, boundaries

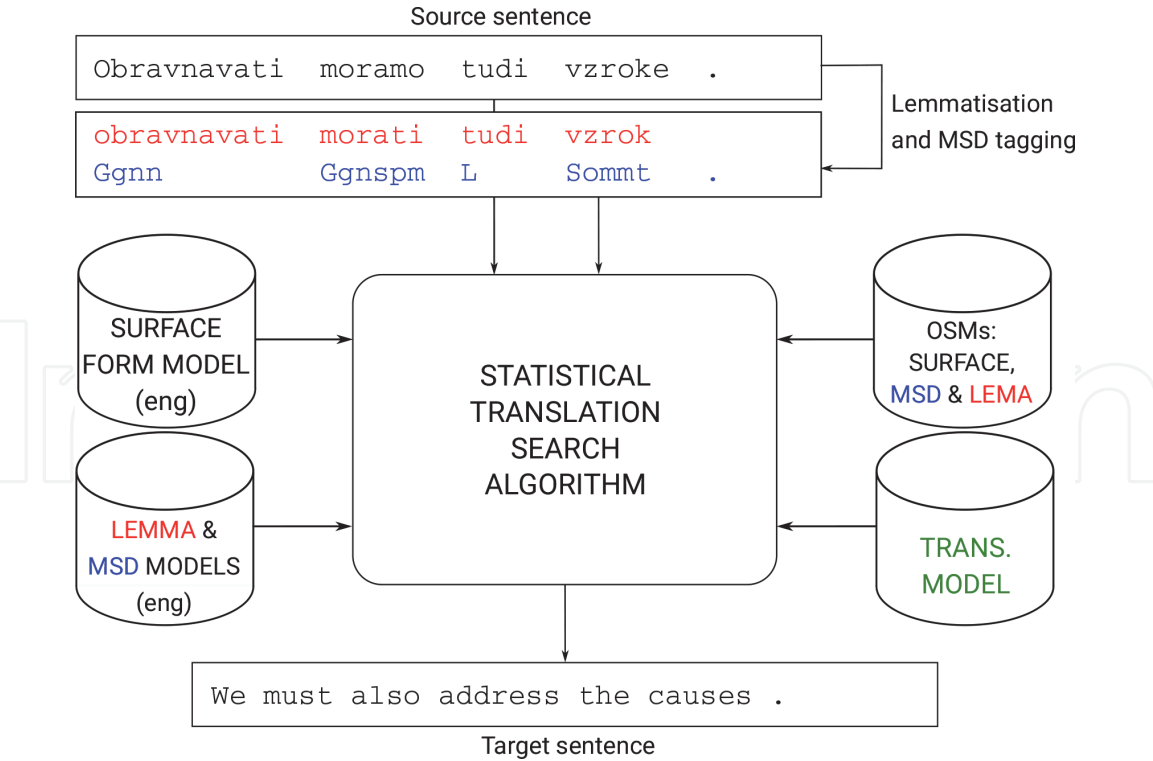


Figure 2.
Statistical machine translation system using a language model based on surface forms, a language model based on MSD tags, a language model based on lemmas, and three OSMs.

between the two approaches have narrowed, and hybrid approaches emerged, which try to gain benefit from both of them. We distinguish two groups of hybrid MT, those guided by rule-based MT and those guided by statistical approaches. Hybrid systems, guided by rule-based MT, use statistical MT to identify the set of appropriate translation candidates and/or to combine partial translations into the final sentence in the target language. Hybrid systems, guided by statistical MT, use rules at pre-/post-processing stages.

2.5 Neural machine translation

Neural MT emerged as a successor of statistical MT. It has made rapid progress in recent years, and it is paving its way into the translation industry as well. Neural MT is a deep learning-based approach to MT that uses a large neural network based on vector representations of words. If compared with statistical MT, there is no separate language model, translation model, or reordering model, but just a single sequence model, which predicts one word at a time. The prediction is conditioned on the source sentence and the already produced sequence in the target language. The prediction power of neural MT is more promising than that of statistical MT, as neural networks share statistical evidence between similar words. In **Figure 3** one of the proposed topology for neural machine translation is given with the same example sentence as in **Figure 2**. The input words are passed through the layers of the encoder (blue circles) to its last layer, the context vector, updating it for every input word. The context layer is then passed through the decoder layers (red circles) to output words, and it is again updated for each output word.

The encoder-decoder recurrent neural network architecture with attention is currently the state of the art for machine translation.

Although effective, the neural MT systems still suffer some issues, such as scaling to larger vocabularies of words and the slow speed of training the models.

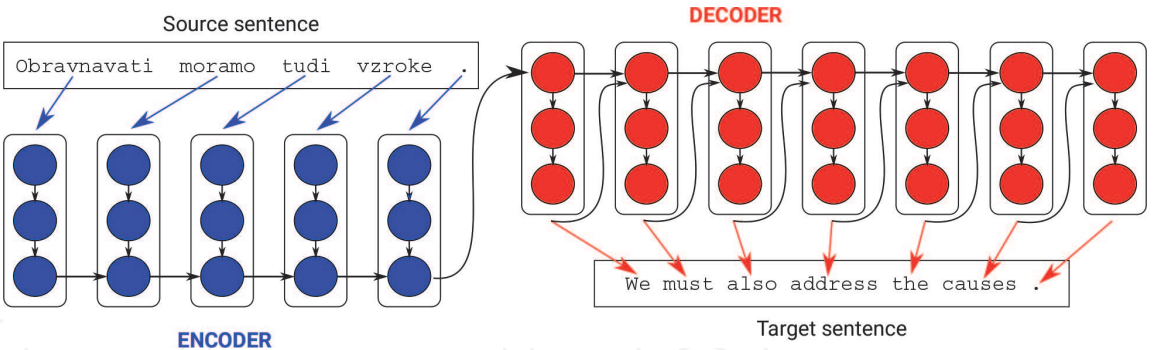


Figure 3. Neural machine translation system using the encoder-decoder topology.

In addition, large corpus is needed to train neural MT systems with performance comparable to statistical machine translation. Researchers continue to work on solving open problems.

3. Problems in machine translation

The fast progress of MT has boosted translation quality significantly, but, unfortunately, machine translation approaches are not equally successful for all language pairs. Morphologically rich languages are problematic in MT, especially if the translation is from a morphologically less complex to a morphologically more complex language. Morphological distinctions not present in the source language need to be generated in the target language. Much work on morphology-aware approaches relies heavily on language-specific tools, which are not always available. Many morphologically rich languages fall in the category of low-resource languages.

One group of morphologically rich languages is a group of highly inflected languages. They are difficult not only for MT but also for other language technology applications [4, 5]. The main problem in highly inflected languages is that the large number of inflected word forms lead to data sparsity (see example in **Table 1**), which results in unreliable estimates in statistical MT [6]. Most words in a given corpus occur at most a handful of times. Therefore, the translation rule coverage is partial, and the estimation of translation probabilities is poor. Some approaches try to reduce the problem of data sparsity by using modelling units other than words; for example, stems and endings, lemmas and morphosyntactic tags, etc. Relaxed word order in inflectional languages poses another problem (see example in **Table 2**). Usually, very little information about the target word order is obtainable from the source sentence. Pre-ordering approaches learn to preprocess the source

	Singular	Dual	Plural
Nominative	študent	študenta	študenti
Genitive	študenta	študentov	študentov
Dative	študentu	študentoma	študentom
Accusative	študenta	študenta	študente
Locative	študentu	študentih	študentih
Instrumental	študentom	študentoma	študenti

In the example, the word has nine different endings.

Table 1. Inflected word forms of the word “student” (masculine) in Slovene.

1.	Angleščino študiram dve leti.
2.	Dve leti študiram angleščino.
3.	Študiram angleščino, dve leti.

Third example is used in colloquial speech.

Table 2.
Word permutations of the English sentence “I have been studying English for two years” in Slovene.

sentence during training in such a way that the words on the source side appear closer to their final positions on the target side. A frequent problem of inflectional languages is also an inaccurate translation of pronouns. There are also many cases in inflectional languages where the subject is dropped completely. Problematic also is differences in the expression of negation. Slavic languages fall into the category of highly inflected languages, and they cause many problems in machine translation [7, 8].

Another group of morphologically rich languages is a group of agglutinative languages, which are even more difficult to use in machine translation. In an agglutinative language, words may consist of more than one, and possibly many, morphemes. Each morpheme in a sequence indicates a particular grammatical meaning. Morphemes are used commonly as basic units in MT for those groups of languages. All these phenomena cause errors in translations produced by MT systems and make the use of MT questionable. It is necessary to evaluate MT quality before use in practice.

4. Machine translation evaluation

As MT emerges as an important mode of translation, its quality is becoming more and more important. Judging translation quality is called machine translation evaluation. It is defined commonly by technical terms. It means, with the exception of human (i.e. manual) evaluation, that it is defined as an algorithm that can be coded into a programme and run by a computer that calculates the evaluation score, which tells the user how good a translation is. Translation evaluation methods count word- and/or sentence-based errors that can be detected automatically, while general text-level aspects are not taken into account. This weakness of automatic MT evaluation is one of the main criticisms in the translation community. Despite that, in the last decade, we have been witnessing great progress in automatic MT evaluation.

MT quality can be measured in many different ways, depending on the goal of the evaluation and the means available. Traditionally, there are two paradigms of machine translation evaluation: Glass-box evaluation and black-box evaluation. Glass-box evaluation measures the quality of a system based on internal system properties. Black-box evaluation examines only the system output, without connecting it to the internal mechanisms of the translation system. The focus in this section will be on black-box evaluation. It is concerned only with the objective behaviour of the system upon a predetermined evaluation set. An evaluation set is a set of sentences in the source language and their translations into the target language, obtained by the translation system. These sentence pairs are then exposed to the evaluation. An evaluation set needs to be selected carefully to cover all data features important for future use of the translation system. The same translation quality can then be expected on the other data that is of the same type as the evaluation set; if not, translations of quite different quality could be obtained.

The reason is in the fact that MT systems are trained on translation examples. If these examples are of a different type to the text that is afterwards translated by the system, the system has only weak knowledge about its translation and, consequently, produces poor translations. Different types of data mean variations in structure, genre, and style. Evaluation, on the other hand, can focus on testing the systems's robustness. In this case, the evaluation set is composed of subsets of different data types. One should be aware that obtaining a robust MT system means at least training it with translation examples of different data types.

There is also a difference between judging and measuring the quality of MT output as a final product and judging and measuring the usability of MT output for subsequent corrections by humans, called post-editing (PE). As regards the latter, it is interesting to know how much editing effort is needed to make the MT output match a reference translation or to become an acceptable translation. What is acceptable translation is left to be decided by the translation expert. In the MT community, there are no criteria for it.

For a long time, methods for evaluating human and MT quality have been disconnected. The comparison between them was impossible. In recent years, a framework called multidimensional quality metrics (MQM) [9] was developed for evaluating the quality of both human and machine translations. It includes over 100 issue types that cover all of the major translation quality evaluation metrics. For the specific translation quality judgement task, relevant issues may be chosen from MQM. The focus of this section is only on the evaluation of MT quality, whereas human translation is taken as the gold standard.

4.1 Manual evaluation

The most common option for judging and measuring machine translation quality is human evaluation. The quality of MT output is judged by experts in translation and linguistics from two different perspectives. The first perspective is the degree of adherence to the target text and target language norms, referring, for example, to features such as grammaticality and clarity. This quality evaluation perspective is known as fluency. When judging fluency, the source text is not relevant. The evaluators have access to only the translation being judged and not the source data. Fluency requires an expert fluent only in the target language. On the other hand, source text adherence is judged to the source text norms and meaning, in terms of how well the target text represents the informational content of the source text. It is known as accuracy. The evaluators have access to the source text and translations being judged. Frequently, the context of a sentence is also taken into account. The evaluators must be bilingual in both the source and target languages. The adequacy and fluency are usually judged on a 5-point scale, as given in **Table 3**.

Human evaluation is time-consuming and expensive. It is also inherently subjective. To alleviate the problem of subjectivity, more experts are usually asked to evaluate the translations in the same evaluation set, and their evaluations are, finally, justified statistically.

4.2 Automatic evaluation

MT systems are rarely static, and they tend to be improved over time as resources grow and bugs are fixed. The evaluation needs to be repeated many times. Automatic evaluation metrics are cost-free alternatives to human evaluation. They are used commonly during the development of MT systems to estimate improvement. They are also applicable to compare different MT systems. While using automatic metrics to judge the translation quality, it is important to understand

Adequacy		Fluency	
5	All meaning	5	Flawless language
4	Most meaning	4	Good language
3	Much meaning	3	Non-native language
2	Little meaning	2	Disfluent language
1	None	1	Incomprehensible

Table 3.
Numeric scale for judging adequacy and fluency.

what their scores mean. They rely on the idea that MT quality in itself should approach human quality. Automatic metrics depend on the availability of human reference translation. They evaluate the output of MT systems by comparing it to the reference translation. As there is a great variability even in human translation, it is important to have more human reference translations for each machine-translated sentence to be evaluated. Evaluation metrics then provide evaluation scores based on the most similar reference translation.

Standard and recently proposed automatic metrics for MT evaluation will be discussed in the continuation of this section. Statistical correlation coefficients are used to see how close automatic evaluation is to manual judgements. Three correlation coefficients will be described later in the section. Machine translation, coupled with subsequent post-editing, has become a widely accepted method in the translation industry. This type of translation workflow will be discussed at the end of this section.

5. Basic metrics for translation evaluation in MT

An obvious method for evaluation is to look at the translation and judge by hand, whether it is correct or not. To get reliable judgements, the evaluators should be appropriately qualified. From a practical point of view, manual evaluation, performed by translation experts, is expensive and takes time. What is needed are automatic metrics that are quick and cheap to use and approximate human judgements accurately. De facto metrics, used in the MT community, are BLEU, NIST, METEOR, and TER. All these metrics need reference translations because they compare the MT output with reference translations and provide comparison scores. If reference translations are available, these metrics can be used to evaluate the output of any number of systems quickly, without the need for human intervention. Let us take an example where the reference translation is “Dve leti že študiram angleščino” and the MT output to be evaluated is “Angleščino študiram dve leti”. If we compute the precision, we get:

$$precision = \frac{correct}{length_o} = \frac{4}{4} = 100\%. \tag{3}$$

correct counts the number of correctly translated words, and *length_o* is the length of machine translation output. For the same example, the recall is:

$$recall = \frac{correct}{length_r} = \frac{4}{5} = 80\%. \tag{4}$$

$length_r$ is the length of reference translation. F-measure results in:

$$F_1 = \frac{precision \cdot recall}{\frac{precision+recall}{2}} = 89\%. \tag{5}$$

Based on given measures, the quality of translation is good, as reordering is not penalised. It is not always a good decision. For example, the MT output “Dve angleščino študiram leti” will get the same evaluation result, even though the translation is disfluent.

BLEU [10] measures the overlap of unigrams (single words) and high-order n -grams between MT output and reference translations. It is defined as follows:

$$BLEU = \min\left(1, \frac{length_o}{length_r}\right) \left(\prod_{i=1}^4 precision_i\right)^{\frac{1}{4}}. \tag{6}$$

The main component of BLEU is n -gram precision, i.e. $precision_i$. It is calculated as the ratio between matched n -grams and the total number of n -grams in the evaluated translation. Precision is calculated separately for each n -gram order, and the precisions are combined via a geometric averaging. The highest n -gram order is defined commonly to be four (four words in a sequence). Higher-order n -grams are used as an indirect measure of a translations level of grammatical well-formedness. The BLEU metric computes the modified precision score, weighted by the brevity penalty, which punishes sentences that are shorter than the reference. The final scores range from 0 to 1. **Table 4** contains the calculation of BLEU score for our example.

BLEU is typically computed over the entire corpus, not single sentences. It is important to point out that very few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1, as there is great variability of possible correct translations. In this sense, it is also important to note that having more reference translations per sentence is highly welcome. It will increase the BLEU score. NIST [11] is a close derivate of BLEU.

Both metrics, BLEU and NIST, focus only on n -gram precision and disregard recall. Recall is the ratio between matched n -grams and the total number of n -grams in the reference translation. METEOR metric [12] combines precision and recall. The authors of METEOR argue that the brevity penalty in BLEU does not compensate adequately for the lack of recall. METEOR computes a score only for unigram

Metric	Score
1-gram precision	$\frac{4}{4}$
2-gram precision	$\frac{1}{3}$
3-gram precision	$\frac{0}{2}$
4-gram precision	$\frac{0}{1}$
Brevity penalty	$\frac{4}{5}$
BLEU	0%

Table 4. BLEU score computation for the MT output “Angleščino študiram dve leti”, if the reference is “Dve leti že študiram angleščino”.

Reference:			dve	leti	že	študiram	angleščino
Output:	angleščino	študiram	dve	leti			
Edit:	I	I	M	M	D	D	D
$WER = \frac{0.5 \cdot 2 + 0.5 \cdot 3}{5} = 50\%$							

Table 5.
WER computation for the MT output “Angleščino študiram dve leti”, if the reference is “Dve leti že študiram angleščino”.

matching. Matching is done in three stages. The first stage is exact matching. Strings are aligned, which are identical in the reference and the translation. Words that are not matched are stemmed in the second stage. Stemming is the process of reducing inflected words to their word stem by cutting off the ends of words. Words with the same morphological root are aligned after stemming. In the last stage, unaligned words which are found to be synonyms are aligned, according to WordNet. WordNet [13] is a large lexical database of synonyms (called synsets). In WordNet, synsets are interlinked by means of conceptual-semantic and lexical relations. METEOR does not use higher-order n -grams, as n -gram counts do not require an explicit word-to-word matching. In METEOR, an explicit measure of the level of grammaticality is used. It captures directly how good the structure of the matched words in the machine translation is in relation to the reference.

Word error rate (WER) metric was first used to evaluate automatic speech recognition. It counts the minimum number of edits needed to change the evaluated translation so that it matches the references exactly, normalised by the average length of the references. The minimum number of edits is also called Levenshtein distance. Possible edits are insertion (I), deletion (D), and substitution (S) of single words. Matched words are denoted with M . Different edits can have different weights. For example, substitution is usually weighted at unity, but deletion and insertion are both weighted at 0.5:

$$WER = \frac{S + 0.5 \cdot D + 0.5 \cdot I}{length_r}. \tag{7}$$

$$TER = \frac{S + D + I + Shift}{length_r}. \tag{8}$$

Table 5 contains the calculation of WER for our example. Translation edit rate (TER) metric [14] is a derivate from the WER. It uses an additional edit step, namely, shifts of word sequences (*Shift*). A shift moves a contiguous sequence of words within the evaluated translation to another location within the translation. All edits have equal cost. If more than one reference is available, and since the minimum number of edits needed to modify the translation is called for, only the number of edits to the closest reference is measured. TER is normalised by the average length of the reference. Position-independent error rate (PER) is another derivate from WER, which treats the reference and translation output as bags of words. Words from the translation are aligned to words in the reference, ignoring the position.

6. Advanced metrics for translation evaluation in MT

Although BLEU, NIST, METEOR, and TER metrics are used most frequently in the evaluation of MT quality, new metrics emerge almost every year. There is a

metrics-shared task, held annually at the WMT Conference where new evaluation metrics are proposed [15, 16, 17]. Those which exhibit high correlation with human judgement will be presented from the pool of recently defined metrics.

CDER [18] is a more advanced metric that is concerned with edits and Levenshtein distance. It calculates the distance between two strings e_i^I and $e_i^{\sim L}$ using the auxiliary quantity $D(i, l)$, defined as:

$$D(i, l) := d_{CD}(e_1^i, e_1^{\sim l}). \quad (9)$$

$$D(0, 0) = 0, \quad (10)$$

$$D(i, l) = \min \left\{ \begin{array}{l} D(i-1, l-1) + c_{SUB}(e_i, \tilde{e}_l), \\ D(i-1, l) + 1, \\ D(i, l-1) + 1, \\ \min_{i'} D(i', l) + 1. \end{array} \right\} \quad (11)$$

In addition to classical edit operations (i.e. insertion, deletion, and substitution), it models block reordering explicitly as an additional edit operation. As a further improvement, it introduces word dependent substitution costs $c_{SUB}(e_i, \tilde{e}_l)$. The observation that the substitution of a word with a similar one is likely to affect translation quality less than the substitution with a completely different word is accounted for in a metric score.

Tolerant BLEU [19] and LeBLEU [20] are derivatives of BLEU with a relaxation of the strict word n -gram matching that is used in standard BLEU. Tolerant BLEU applies a specific distance measure that requires an exact match only in the middle of words, not in words as a whole. LeBLEU uses a distance measure based on characters. It also facilitates a fuzzy matching of longer chunks of text that allows, for example, to match two independent words with a compound.

CharacTER [21] is a derivate of TER:

$$CharacTER = \frac{S_{char} + D_{char} + I_{char} + Shift\ Cost}{length_{ochar}}. \quad (12)$$

It calculates the edit rate on character level, whereas shift edits are still performed on word level. First, a technique for word-level shifts is performed, words are then split into characters, and, finally, the edit distance is calculated based on characters, and *Shift Cost* is calculated. In addition, the lengths of translations in characters ($length_{ochar}$) instead of references are used for normalising the edit distance, which effectively counters the issue that shorter translations normally achieve lower TER. If we have two translations of different lengths, but with the same edit distance, they will obtain the same TER, as the length of the reference remains unchanged. In the same case, the longer translation will obtain lower TER if the edit distance is normalised by the length of translation.

METEOR universal [22] is a derivate of METEOR. It adds the fourth stage to matching. It is paraphrase matching. For each target phrase e_1 , all source phrases f that e_1 translates are found. Each alternate phrase ($e_2 \neq e_1$) that translates f is considered a paraphrase with probability $P(f|e_1) \cdot P(e_2|f)$. The cumulative probability of e_2 being a paraphrase of e_1 is the sum over all possible pivot phrases f :

$$P(e_2|e_1) = \sum P(f|e_1) \cdot P(e_2|f). \quad (13)$$

Phrases are matched if they are listed as paraphrases in a language appropriate paraphrase table. Paraphrases are extracted automatically from the parallel corpora used to train statistical MT systems.

The BEER [23] metric provides a linear combination of different features:

$$BEER(h, r) = \sum_i w_i \times \phi_i(h, r) \quad (14)$$

where h is the system output and r is the reference. Each feature $\phi_i(h, r)$ has a weight w_i assigned to it. The first group of features consists of adequacy features. These features use precision, recall, and F1-score for different counts. F1-score is the harmonic mean of precision and recall multiplied by the constant of 2. The constant of 2 scales the F1-score to 1 when both recall and precision are 1. In BEER, function words and content words (nonfunction words) are counted separately. By differentiating function and nonfunction words, a better estimate is obtained of which words are more important and are less. The most important adequacy feature is a count of matching character n -grams. Using it, the translations are considered partially correct even if they did not get the morphology completely right. Character n -grams of order 6 are used. The second group of features comprises ordering features. Word order is evaluated by presenting the reordering as a permutation and calculating the distance to the ideal monotone permutation. Permutation trees are used to estimate long-distance reordering.

ChrF [24] is another, even simpler, metric based on character n -grams. It computes an F-score, based on precision and recall, using character n -grams:

$$ChrF\beta = (1 + \beta^2) \frac{ChrP \cdot ChrR}{\beta^2 \cdot ChrP + ChrR} \quad (15)$$

$ChrP$ and $ChrR$ are character n -gram precision and recall, averaged over all n -grams. $ChrP$ is the percentage of n -grams in the translation, which have a counterpart in the reference. $ChrR$ is the percentage of character n -grams in the reference, which are also present in the translation. In the final score, the parameter is used, which gives more importance to recall than to precision. In [25], the optimal value for β is found to be 2, and the metric is called chrF2. In this metric, a recall has two times more importance than precision. WordF2 is a similar metric, where words are used instead of characters. Different weightings for n -grams were also investigated. Uniform weights are the most promising for machine translation evaluation.

DREEM [26] is a new metric based on distributed representations of words and sentences generated by deep neural networks. Neural networks are models that imitate human brains to recognise patterns in sequences. DREEM employs three different types of word and sentence representations: One-hot representations, distributed word representations learned from a neural network model, and distributed sentence representations computed with a recursive autoencoder. The final score is the cosine similarity of the representation of the translation and the reference, multiplied with a length penalty.

RATATOUILLE [27] is a metric combination of BLEU, BEER, METEOR, and few more metrics, out of which METEOR-WSD is a novel contribution. METEOR-WSD is an extension of METEOR that includes synonym mappings.

In this section, state-of-the-art MT evaluation metrics were investigated briefly. Only the most important characteristics of them were exposed. For a more elaborate description of each metric, the reader is advised to use the provided references to literature.

It should be noted that despite the well-known problems with BLEU, and the availability of many other metrics, MT system developers have continued to use BLEU as the primary measure of translation quality.

Today, different MT systems are available for use in practice. Usually, the qualities of different MT systems are compared between themselves by computing the translation quality scores on a predetermined evaluation set. The question arises whether, if there is a difference in quality on the evaluation set, one can be ensured that different MT systems indeed own different system quality. A difference in quality on an evaluation set may be just the result of happenstance. Research work on the statistical significance test for MT evaluation was done by Koehn [28], and the bootstrap resampling method is proposed to compute the statistical significance intervals for evaluation metrics on evaluation data. Statistical significance usually refers to the notions of the p-value, the probability that the observed difference in quality will occur by chance given the null hypothesis.

7. Correlation between automatic and human evaluation

Human judgements of translation quality are usually trusted as the gold standard, and the aim of an automatic evaluation metric is to produce quality estimates that are as close as possible to human judgements. As there are many different evaluation metrics, the user needs to decide which automatic evaluation metric he trusts the most. Correlation coefficients are used commonly to measure the closeness of automatic metric scores and manual judgements. Manual MT quality judgements on a number of test data are needed for comparison. Correlation coefficients are then computed on system level and/or segment level.

System-level comparison is done to compare different MT systems in general. First, each system gets a cumulative rank that reflects how high the annotators ranked that system. The metric scores of systems are also converted into ranks, and then the Spearman's rank correlation coefficient ρ is computed as [16]:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}. \quad (16)$$

d_i is the difference between the annotator's rank and metric's rank for system i . The number of systems is denoted with n . The possible values of ρ range between 1 (where all systems are ranked in identical order) and -1 (where the systems are ranked in the reverse order). Metrics with values of Spearman's ρ closer to 1 are better. The Spearman's correlation coefficient ρ is sometimes too harsh [17]: If a metric disagrees with humans in ranking two systems of a very similar quality, the ρ coefficient penalises this equally as if the systems were very distant in their quality. Pearson's correlation coefficient r is sometimes preferred [17]. It measures the strength of the linear relationship between a metric's scores and human scores:

$$r = \frac{\sum_{i=1}^n (H_i - \bar{H}) \cdot (M_i - \bar{M})}{\sqrt{\sum_{i=1}^n (H_i - \bar{H})^2} \sqrt{\sum_{i=1}^n (M_i - \bar{M})^2}}. \quad (17)$$

H is the vector of annotator's scores of all systems, and M is the vector of the corresponding scores as predicted by the given metric. \bar{H} and \bar{M} are their means, respectively. **Figure 4** shows Pearson's correlations of selected system-level metrics and MT systems built for different language pairs [15]. We can see that in majority of cases metrics correlate well with human judgements.

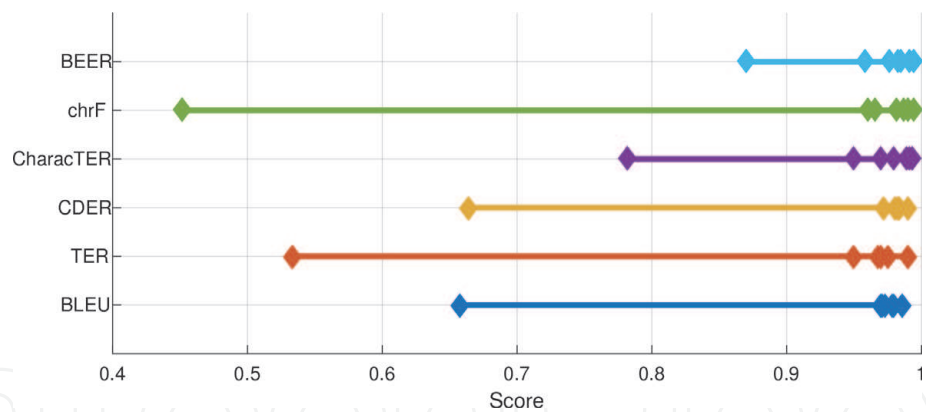


Figure 4.
Pearson's correlation coefficient r for selected evaluation metrics used for different MT systems.

The quality of a metric's segment level scores is usually measured by means of Kendall's τ rank correlation coefficient [17]. Let $r(\cdot)$ denotes annotator's rank and $m(\cdot)$ metric's rank. To compute Kendall's τ , the annotators rank all the translations of each segment from the best to the worst. Pairs (a, b) are then built where one system's translation $r(a)$ of a particular segment is judged to be (strictly) better than the other system's translation $r(b)$:

$$Pairs := \{(a, b) \mid r(a) < r(b)\}. \tag{18}$$

Afterwards, all concordant and discordant pairs are counted:

$$\begin{aligned} Con &:= \{(a, b) \in Pairs \mid m(a) > m(b)\}, \\ Dis &:= \{(a, b) \in Pairs \mid m(a) < m(b)\}. \end{aligned} \tag{19}$$

In a concordant pair, a human annotator and an automatic metric agree in ranking, and in a discordant pair, they disagree. Finally, Kendall's τ is computed as:

$$\tau = \frac{|Con| - |Dis|}{|Con| + |Dis|}. \tag{20}$$

τ value is between -1 (a metric always predicted a different order than humans did) and 1 (a metric always predicted the same order as humans). Metrics with higher τ are better.

In this section, no analysis of correlations for automatic metrics is presented, as it depends on many parameters. In general, all evaluation metrics presented in this chapter correlate well with human judgements. It is only worth mentioning that, for inflected languages, metrics that work on character level correlate better with human judgements than metrics that work only on word level.

8. Post-editing machine translation

In recent years, MT has become accepted more widely in the translation industry [29]. The most common workflow involves the use of machine-translated text as a raw translation that is corrected or post-edited by a translator. Post-editing (PE) tools and practices for such workflows are being developed in large multilingual organisations, such as the European Commission [30]. The researchers in [31] report that 30% of the companies in the translation industry currently use MT. The majority (70%) of the MT users combine MT with PE at least some of the time.

Post-editing MT is attractive because it has been shown to be faster than human translation. It is faster than translation from scratch and even faster than translation assisted by a translation memory [32]. Speed is not the only factor that should be taken into account when assessing the post-editing process. More recent studies have looked at ways of determining post-editing effort. In [33], three levels of post-editing effort are defined: Temporal effort, cognitive effort, and technical effort. The temporal effort is the time needed to post-edit a given text, cognitive effort is the activation of cognitive processes during post-editing, and the technical effort means the operations such as insertions and deletions that are performed during post-editing. All three levels of post-editing effort are influenced greatly by the translation quality. The use of PE and MT also raises the question about the quality of final translations. Has the quality improved, or is it worse?

As PE effort is related strongly to MT quality, derivatives of standard quality metrics are developed, which are concerned more with PE effort. Human-mediated translation error rate (HTER) [14] is a human-in-the-loop variant of TER. Instead of a reference, post-edited translation is used in the comparison. HTER centres on what edits are to be made to convert a translation into its post-edited version. It is computed as the ratio between the number of edit steps and the number of words in the post-edited version. HTER can be used as a measure of technical PE effort: The fewer changes necessary to convert the translation into its post-edited version, the less the effort required from the translator.

HTER is concerned more with the final translation and not the process. In [34] a metric called actual edit rate (AER) is proposed, which measures the translator's actual edit operations, which may involve more complex tasks, for example, applying corrections to previously post-edited parts of the text.

A study on PE of MT confirmed the relation between HTER and MT qualities [34]. An increase in HTER was evident as the quality of the MT system decreased. In contrast, they did not establish any significant association between AER and MT qualities. Keyboard activity may not be as sensitive to MT quality as PE time. They also found a linear relationship between MT quality and post-editing speed. MT quality was measured by the BLEU score of the system. The increase of BLEU score by one point resulted in a decrease of post-editing speed of about 0.16 seconds/word post-editing time. Their study also shows the correlation between the quality of machine translation output and the quality after post-editing. They confirmed that worse translation almost always leads to worse result after post-editing. As the use of MT and PE workflows has increased, there is a growing demand for expertise in PE skills. The research on and teaching of skills specific to post-editing has become necessary. The authors in [31] emphasise the impact of "familiarity with translation technology" on the employability of future translators.

9. Conclusion

Machine translation is being used by millions of people on a daily basis. This chapter discusses different MT approaches that were developed over time. Currently, the most promising approach is neural machine translation. Although effective, it also suffers some issues, such as scaling to larger vocabularies of words and the slow speed of training the models. Researchers continue to work on solving the problems and making translation a better service accessible to everyone.

The second part of the chapter describes how machine translation output is evaluated. The main characteristics of human and automatic MT evaluation were outlined. Human evaluation of MT output remains crucial to look for ideas to improve MT systems still further. On the other hand, automatic MT evaluation is

cheap and fast. In the chapter, traditional and advanced metrics for automatic MT evaluation were presented. Despite the well-known problems with BLEU, and the availability of many other metrics, MT system developers have continued to use BLEU as the primary measure of translation quality.

MT quality is continually improving. Despite that, there are still a number of flaws in machine translation output. To make the translation correct, post-editing machine translation output is proposed to be integrated into the translation processes. It is discussed at the end of the chapter.

Future research in MT will be devoted to neural machine translation. It is still not very well understood. Its inner workings are commonly seen as a black-box, which works as the neurons of the human brain. As the computing power NMT requires becomes more widely available, many different configurations can be examined to further improve the accuracy of machine translation.

Future effort in machine translation evaluation will be directed toward character-based metrics which show the highest correlation with human judgement at the system and segment levels.

Human translators are worried to be replaced by machines. Machine translation, no matter how sophisticated, cannot match the accuracy of people. Human translators are also an important segment in MT evolution not only as post-editors but also as teachers for MT systems to become better and better.

Acknowledgements


The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0069).

Author details

Mirjam Sepesy Maučec* and Gregor Donaj
Faculty of Electrical Engineering and Computer Science, University of Maribor,
Slovenia

*Address all correspondence to: mirjam.sepesy@um.si

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Doherty S, Gaspari F, Groves D, van Genabith J. Mapping the industry. I: Findings on translation technologies and quality assessment. In: GALA. 2013
- [2] Koehn P, Och FJ, Marcu D. Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1. Association for Computational Linguistics; 2003. pp. 48-54
- [3] Durrani N, Schmid H, Fraser A. A joint sequence translation model with integrated reordering. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. Association for Computational Linguistics; 2011. pp. 1045-1054
- [4] Donaj G, Kačič Z. Language Modeling for Automatic Speech Recognition of Inflective Languages: An Applications-Oriented Approach Using Lexical Data. Springer; 2016
- [5] Donaj G, Kačič Z. Context-dependent factored language models. EURASIP Journal on Audio, Speech, and Music Processing. 2017;2017(1):6
- [6] Maučec MS, Donaj G. Morphosyntactic tags in statistical machine translation of highly inflectional language. In: Proceedings of the Artificial Intelligence and Natural Language Conference (AINL FRUCT); Saint-Petersburg, Russia. 2016. pp. 99-102
- [7] Maučec MS, Brest J. Slavic languages in phrase-based statistical machine translation: A survey. Artificial Intelligence Review. 2019;51(1):77-117
- [8] Maučec MS, Donaj G. Morphology in statistical machine translation from english to highly inflectional language. Information Technology and Control. 2018;47(1):63-74
- [9] Lommel AR, Burchardt A, Uszkoreit H. Multidimensional quality metrics: A flexible system for assessing translation quality. In: Proceedings of ASLIB: Translating and the Computer. Vol. 35. 2013
- [10] Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics; 2002. pp. 311-318
- [11] Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc; 2002. pp. 138-145
- [12] Lavie A, Agarwal A. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics; 2007. pp. 228-231
- [13] Fellbaum C. WordNet. In: Poli R, Healy M, Kameas A, editors. Theory and Applications of Ontology: Computer Applications. Springer; 2010. pp. 231-243
- [14] Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas. Vol. 200. 2006
- [15] Ma Q, Bojar O, Graham Y. Results of the wmt18 metrics shared task: Both

characters and embeddings achieve good performance. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. 2018. pp. 671-688

[16] Macháček M, Bojar O. Results of the WMT13 metrics shared task. In: *Proceedings of the Eighth Workshop on Statistical Machine Translation*. 2013. pp. 45-51

[17] Macháček M, Bojar O. Results of the wmt14 metrics shared task. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014:293-301

[18] Leusch G, Ueffing N, Ney H. Cder: Efficient MT evaluation using block movements. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006

[19] Libovický J, Pecina P. Tolerant bleu: A submission to the wmt14 metrics task. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014. pp. 409-413

[20] Virpioja S, Grönroos S-A, Lebleu: N-gram-based translation evaluation score for morphologically complex languages. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 2015. pp. 411-416

[21] Wang W, Peter J-T, Rosendahl H, Ney H. Character: Translation edit rate on character level. In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Volume 2*. 2016. pp. 505-510

[22] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. 2014. pp. 376-380

[23] Stanojevic M, Sima'an K. Beer: Better evaluation as ranking. In: *Proceedings of the Ninth Workshop on*

Statistical Machine Translation. 2014. pp. 414-419

[24] Popović M. Chrf: Character n-gram f-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 2015: 392-395

[25] Popović M. Chrf deconstructed: Beta parameters and n-gram weights. In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, Volume 2*. 2016. pp. 499-504

[26] Chen B, Guo H, Kuhn R. Multi-level evaluation for machine translation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 2015. pp. 361-365

[27] Marie B, Apidianaki M. Alignment-based sense selection in meteor and the ratatouille recipe. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 2015:385-391

[28] Koehn P. Statistical significance tests for machine translation evaluation. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 2004

[29] Way A. Quality expectations of machine translation. In: *Translation Quality Assessment*. Springer; 2018. pp. 159-178

[30] Bonet J. No rage against the machine. *Languages and Translation*. 2013;6(2)

[31] Gaspari F, Almaghout H, Doherty S. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives*. 2015;23(3):333-358

[32] Plitt M, Masselot F. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*. 2010;93:7-16

[33] Krings HP, Shreve GM. Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes. Vol. 5. Kent State University Press; 2001

[34] Sanchez-Torron M, Koehn P. Machine translation quality and post-editor productivity. In: AMTA 2016 Vol. 2016. p. 16