# Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain

Sanja Seljan; Ivan Dunđer
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Zagreb, Croatia
sanja.seljan@ffzg.hr; ivandunder@gmail.com

*Abstract* — **Automatic quality evaluation of machine translation systems has become an important issue in the field of natural language processing, due to raised interest and needs of industry and everyday users. Development of online machine translation systems is also important for less-resourced languages, as they enable basic information transfer and communication. Although the quality of free online automatic translation systems is not perfect, it is important to assure acceptable quality. As human evaluation is time-consuming, expensive and subjective, automatic quality evaluation metrics try to approach and approximate human evaluation as much as possible. In this paper, several automatic quality metrics will be utilised, in order to assess the quality of specific machine translated text. Namely, the research is performed on sociological-philosophical-spiritual domain, resulting from the digitisation process of a scientific publication written in Croatian and English. The quality evaluation results are discussed and further analysis is proposed.**

***Keywords – automatic quality evaluation; machine translation; BLEU; NIST; METEOR; GTM; English-Croatian; Croatian-English; sociological-philosophical-spiritual domain.***

## I. INTRODUCTION

Machine translation evaluation has become a hot topic of interest to numerous researchers and projects, usually when comparing several machine translation systems with the same test set or when evaluating one system through different phases, such as automatic evaluation, correlation of automatic evaluation scores with human evaluation etc. [1].

Lately, extensive evaluation of machine translation quality was conducted focusing on online machine translation systems, commercial or integrated systems applying statistical machine translation, sometimes combined with other machine translation approaches [2].

Statistical machine translation systems rely on huge amounts of parallel data, which is sometimes inconvenient for less-resourced languages or between different types of languages, and which requires more detailed quality error analysis and evaluation, in order to improve the performance of a machine translation system [3].

Some authors point out the context of machine translation evaluation relating quality, purpose and context, trying to establish a coherent evaluation approach [4]. Automatic evaluation for less-resourced, but morphologically rich languages is a topic of interest of numerous researchers and organisations, since the results could be useful to professional translators, translation industry, researchers and to everyday users. The main advantages of automatic evaluation metrics are speed, cost and objectiveness. They perform always in the same way, and besides being tuneable, they can provide meaningful, consistent, correct and reliable information on the level of machine translation quality [5].

Human evaluation, on the other hand, is considered to be the "gold standard", but it is subjective, tedious and more expensive.

## II. RELATED WORK

Evaluation of machine translation was mainly performed in the legislation, technical or general domain, due to available bilingual corpora [6]. Domains such as sociology, philosophy or religion are rarely investigated, as acquiring necessary corpora and a specific reference translation represents a notable problem.

However, one research paper describes translation in cross-language information access performed by machine translation, supplemented by domain-specific phrase dictionaries, which were automatically mined from the online Wikipedia in the domain of cultural heritage [7]. Queries were translated from and into English, Spanish and Italian and then evaluated using human annotations.

In a research, the idea was to evaluate machine translations produced by Google Translate for the English-Swedish language pair and for the fictional and non-fictional texts, including samples of law documents, commercial company reports, social science texts (religion, welfare, astronomy) and medicine [8]. Evaluation is carried out with the BLEU metric, showing that law texts gained double of average BLEU scores.

Evaluation of machine translation shows better scores for about 20% when two reference sets are used, and up to 29% for three reference sets, with regard to differentiating short and long sentences [9]. Besides sentence length, other problems in the machine translation process were investigated, such as specific terminology, anaphora and ambiguity [8].

Another research describes the importance of the translation domain, which influences the quality of machine

translation output. Therefore, domain knowledge and specific terminology translations have been added [10]. The research is conducted with the SYSTRAN translation system, which uses the transfer translation approach for Chinese-English, English-French, French-English and Russian-English language pairs.

Research on machine translation of religious texts by Google Translate for English-Urdu and Arabic-Urdu has also been conducted [11]. Evaluation of cross-language information retrieval using machine translation in the domain of sociology is presented in [12] for English, French, German and Italian.

## III. RESEARCH

The following subsections describe the digitisation process and data set, and discuss the research methods, quality evaluation metrics and used tools.

### A. Digitisation

For the purpose of this research, a book of abstracts from a scientific conference containing mutual translations in Croatian and English was digitised with a scanner. Digitisation represents the systematic recording, storing and processing of content using digital cameras, scanners and computers [13]. It is the process of creating a digital representation of an object, image, document or a signal, and allows them to be stored, displayed, disseminated and manipulated on a computer. In order to digitise the mutual bitexts, a HP Scanjet G3110 flatbed scanner was used, set to 300 dpi and grayscale scanning. Scanned abstracts were in A5 format and text was written in Times New Roman font, size 10, standard black font colour on white paper.

Afterwards, optical character recognition (OCR) was carried out for extracting, editing, searching and repurposing data from the scanned book of abstracts. In this research Abby Fine Reader 8.0.0.677 was used as the OCR software, which identifies text by analysing the structure of the object that needs to be digitised, by dividing it into structural elements and by distinguishing characters through comparison with a set of pattern images stored in a database and built-in dictionaries. During optical character recognition, errors are inevitable, and the induced noise is a serious challenge to subsequent processes that attempt to make use of such data [14].

### B. Data set

The book of abstracts, which consisted of 41 abstracts, was digitised and afterwards processed with OCR, then manually corrected and later on used as the reference set for machine translation, i.e. gold standard. The book contained very specific abstracts of full scientific papers in the fields of sociology, psychology, theology and philosophy with emphasis on several topics, such as human dignity, religion, dialogue, freedom, peace, responsibility, family and community, philosophical and sociological reflections.

The texts were compiled into a parallel bilingual sentence-aligned Croatian-English bitext consisting of mutual translations relating to the sociological-philosophical-spiritual domain. The process of preparing the data set included the digitisation of printed material, applying OCR techniques and its evaluation. All segments were sentence-aligned and a translation memory was created. The format of such a translation memory is ideal for further research in statistical terminology and collocation extraction, evaluation and analysis.

The following table shows the data set statistics (Table I). The data set consisted of 369 segments (sentences) and 107264 characters in total. The longest segment was 80 words long in Croatian, and 88 words in English, whereas 3677 distinct words appeared in Croatian, and 2782 in English. On average, English abstracts were composed of 10.6% more characters, and 19% more words. Specificity of terminology is also reflected in the large number of hapax legomena, which also indicates a variety of different topics in the digitised book of abstracts.

TABLE I. DATA SET STATISTICS

| Data set | Language | |
|---|---|---|
| | ***Croatian*** | ***English*** |
| No. of characters | 50631 | 56633 |
| No. of words | 7340 | 9062 |
| No. of segments | 369 | 369 |
| No. of abstracts | 41 | 41 |
| Max. of words per segment | 80 | 88 |
| Min. words per segment | 1 | 1 |
| Distinct words | 3677 | 2782 |
| Words that appear only once (hapax legomena) | 2837 (77.16%) | 1892 (68.01%) |
| Words that appear twice (dis legomena) | 424 (11.53%) | 389 (13.98%) |
| Words that appear three times (tris legomena) | 165 (4.49%) | 159 (5.72%) |
| Words that appear more than three times | 251 (6.83%) | 342 (12.29%) |
| Arithmetical mean of characters per abstract | 1234.90 | 1381.29 |
| Arithmetical mean of characters per segment | 137.21 | 153.48 |
| Arithmetical mean of words per abstract | 179.02 | 221.02 |
| Arithmetical mean of words per segment | 19.89 | 24.56 |
| No. of OCR errors in total | 66 | 67 |
| Arithmetical mean of OCR errors per abstract | 1.61 | 1.63 |

In total, 133 OCR errors occurred in the digitisation process. Typical errors during optical character recognition were misrecognitions of characters, missing whitespace characters or apostrophes, various forms of substitution errors, as well as space deletion and insertion errors. The most frequent OCR errors in Croatian were substitution errors (e.g. (l) → (i)) and space deletion, where two words were erroneously unified (e.g. (U postkomunističkim) → (Upostkomunističkim)). The most frequent OCR errors in English were substitution errors (e.g. («) → ((()) and missing apostrophes ('). OCR errors have an impact on later-stage processing and data usability, therefore, all scanned texts were manually post-edited afterwards.

### C. Tools and methods

All machine translations for both directions (Croatian-English and English-Croatian) were generated by the freely available online machine translation service, Google Translate (https://translate.google.com/). Automatic machine translation quality evaluation is performed for both directions by the following metrics: BLUE (BiLingual Evaluation Understudy), NIST (National Institute of Standards and Technology), METEOR (Metric for Evaluation of Translation with Explicit ORdering) and GTM (General Text Matcher).

The basic idea behind the mentioned metrics is to calculate the matching of automatic and reference translations. The metrics are based on overlapping of the same surface forms, which is not suitable for languages with rich morphology and relatively free word order. Some of the metrics are based on fixed word order (METEOR, GTM), which is also not suitable for languages with relatively free word order, such as Croatian. BLEU is more order-independent, whereas METEOR introduces linguistic knowledge for n-grams having the same lemma, or for synonym matches.

GTM and METEOR are based on precision and recall, while BLEU and NIST are based on precision and to compensate recall, BLEU introduced brevity penalty [15]. Metrics are mainly focused on evaluation of adequacy, as it gives information to what extent the meaning in the translation is preserved, and penalise translations with missing words, affecting recall.

The BLEU metric, proposed by IBM, represents a standard for machine translation evaluation [16]. It matches machine translation n-grams with n-grams of its reference translation, and counts the number of matches on the sentence level, typically for 1-4 n-grams. For each n-gram it assigns the same weights, which is one of the main defaults of this metric. This metric is based on the same surface forms, accepting only complete matches, and does not take into account words having the same lemma. BLEU also assigns a brevity penalty score, which is given to automatic translations shorter than the reference translation. It allows evaluation of multiple reference translations as well.

The NIST metric is based on BLEU with some modifications [17]. While BLEU is based on n-gram precision assigning an equal weight to each word, NIST calculates information weight for each word, i.e. higher scores are given to more rare n-grams which are considered as more informative n-grams. It differs also from BLEU in brevity penalty calculation, where small differences in translation length do not impact the overall score. Stemming is significantly beneficial to BLEU and NIST [18].

METEOR metric modifies BLEU in the way that it gives more emphasis to recall than to precision [19]. This metric incorporates linguistic knowledge, taking into account the same lemma and synonym matches, which is suitable for languages with rich morphology [20]. This metric, like GTM, favours longer matches in the same order. It uses fragmentation penalty, which reduces F-measure if there are no bigrams or longer matches [21]. This metric is calculated at the sentence or segment level, while BLEU metric is usually computed at the corpus level. It cannot implement language knowledge from several references into the score, but gives scores for each reference translation.

GTM metric computes the correct number of unigrams and favours longer matches, based on precision (the number of correct words, divided by the generated machine translation system output-length) and recall (the number of correct words, divided by the reference-length) and calculates the F-measure [22]. This metric computes unigrams, i.e. the correct number of unigram matches referring to non-repeated words in the output and in the reference translation. This metric favours n-grams in the correct order and assigns them higher weights [23].

Apart from the mentioned disadvantages of automatic evaluation metrics, there are other numerous defaults, such as, aspect of evaluation, difficulties with the interpretation and meaning of scores, ignoring the importance of words, not addressing grammatical coherence etc. [23].

## IV. RESULTS AND DISCUSSION

The following table shows the results of automatic machine translation quality evaluation metrics (Table II). BLEU scores range from 0 (no overlapping with reference translation) to 1 (perfect overlapping with reference translation), whereas scores over 0.3 generally reflect understandable translations, and scores over 0.5 reflect good and fluent translations [24]. METEOR scores are usually higher than BLEU scores and reflect understandable translation when higher than 0.5, and good and fluent translation when scored higher than 0.7 [24]. NIST scores be 0 or higher, and have no fixed maximum, whereas GTM scores can range from 0 to 1. Different metric scores provide an overall overview of the machine translation quality with regard to various aspects of evaluation and can be correlated.

TABLE II. RESULTS OF AUTOMATIC QUALITY EVALUATION METRICS

| **Machine translation direction** | **Automatic machine translation quality evaluation metrics** (higher is better) | | | |
|---|---|---|---|---|
| | ***BLEU*** | ***NIST*** | ***METEOR*** | ***GTM*** |
| English-Croatian | 0.1656 | 4.6527 | 0.1976 | 0.3348 |
| Croatian-English | 0.2383 | 5.8686 | 0.2439 | 0.5044 |

Overall, results of automatic quality assessment show better scores for Croatian-English direction for 20-30%. This is mainly due to fact that Croatian language is highly flective with rich morphology. Furthermore, metrics which rely on word matching penalise the word types that appear in form of different tokens. As the data set belongs to the sociological-philosophical-spiritual domain, it contains specific terminology for which Google Translate does not provide a correct translation. Namely, such terminology is unlikely to appear in the correct context in the language and translation models of the analysed machine translation system. The fact that the data set contains 77.16% of hapax legomena in Croatian and 68.01% in English also points to infrequently used terminology. BLEU metric, which ignores the word relevance, penalises a machine translation that is shorter than the reference translation and counts the words having the same

surface form. In this research BLEU score is relatively low for English-Croatian (0.17) when compared to the Croatian-English direction (0.24). Generally, translating from morphologically rich languages to less rich languages results in better BLEU scores. NIST metric which is sensitive to more informative n-grams which occur less frequently, gives the following results: 4.65 for English-Croatian and 5.87 for Croatian-English. The metric METEOR shows scores close to BLEU metric for English-Croatian (0.20) and for Croatian-English (0.24). Although METEOR counts matches at the stem level, in this research the raw data set was used, which was not lowercased and tokenised. The results of GTM metric, which computes F-measure, are as follows: 0.33 for English-Croatian and 0.50 for Croatian-English. The GTM score for English-Croatian is lower than for Croatian-English due to morphological variants of the same lemma, which causes lower scores due to non-matching of words belonging to the same lemma but with different morphological suffixes.

## V. CONCLUSIONS

In this research, a book of scientific abstracts was digitised with a scanner, subsequently processed with OCR, later on post-edited, and eventually used as a gold standard for machine translation. Automatic machine translations were generated by Google Translate and afterwards evaluated by means of several metrics and for both directions (Croatian-English and English-Croatian). The results for translation into Croatian are less scored due to specific terminology that is not widely used on the internet and therefore not available in the correct context in the machine translation models, morphological richness of the Croatian language, long sentences, relatively free word order and grammatical case agreement. Namely, this causes decreased scores since several types of the same lemma, which are not identical with the reference morphological variant, count as mismatches. Overall, results of the automatic machine translation quality evaluation for BLEU, NIST, METEOR and GTM are better for the Croatian-English direction (20-30% better). Further research on automatic quality evaluation would include more extensive evaluation applying other metrics, text lemmatisation, lowercasing, tokenisation and enlargement of the data set, using multiple reference sentences.

## REFERENCES


[1] D. R. Amancio, M. G. V. Nunes, O. N. Oliveira Jr., T. A. S. Pardo, L. Antiqueira and L. da F. Costa, "Using metrics from complex networks to evaluate MT," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 1, 2011, pp. 131-142.

[2] S. Hampshire and C. Porta Salvia, "Translation and the internet: Evaluating the quality of free online machine translators," Quaderns: revista de traducció, no. 17, 2010, pp. 197-209.

[3] S. Stymne, "Pre- and postprocessing for statistical machine translation into germanic languages," Proceedings of the ACL-HLT 2011 Student Session, 2011, pp. 12-17.

[4] E. Hovy, M. King and A. Popescu-Belis, "Principles of context-based machine translation evaluation," Machine Translation, vol. 17, 2002, pp. 43-75.

[5] P. Koehn, "What is a better translation? Reflections on six years of running evaluation campaigns," Tralogy 2011, 2011, p. 9, available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/tralogy11.pdf

[6] C. Kit and T. M. Wong, "Comparative evaluation of online machine translation systems with legal texts," Law Library Journal, vol. 100, no. 2, 2008, pp. 299-321.

[7] G. J. F. Jones, F. Fantino, E. Newman, and Y. Zhang, "Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia," ceedings of th Second International Workshop on "Cross Lingual Information Access", 2008, pp. 34-41.

[8] J. Salimi, "Machine Translation Of Fictional And Non-fictional Texts," Stockholm University Library, 2014, p. 16, available at: http://www.diva-portal.org/smash/get/diva2:737887/FULLTEXT01.pdf

[9] S. Seljan, T. Vičić and M. Brkić, "BLEU evaluation of machine-translated English-Croatian legislation," Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2143-2148.

[10] E. D. Lange and J. Yang, "Automatic domain recognition for machine translation", Proceedings of the MT Summit VII, 1999, pp. 641-645.

[11] T. T. Soomro, G. Ahmad and M. Usman, "Google Translation service issues: Religious text perspective," Journal of Global Research in Computer Science, vol. 4, no. 8, 2013, pp. 40-43.

[12] M. Braschler, D. Harman, M. Hess, M. Kluck, C. Peters and P. Schäuble, "The evaluation of systems for cross-language information retrieval," Proceedings of the Second International Conference on Language Ressources and Evaluation (LREC-2000), 2000, p. 6.

[13] D. Lopresti, "Optical character recognition errors and their effects on natural language processing," International Journal on Document Analysis and Recognition, vol. 12, no. 3, 2009, pp. 141-151.

[14] J. Smolčić and A. Valešić, "Legal contexts of digitization and preservation of written heritage," Proceedings of the INFuture2009 – Digital Resources and Knowledge Sharing Conference, 2009, pp. 87-94.

[15] C. Callison-Burch, M. Osborne and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 249-256.

[16] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311-318.

[17] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," Proceedings of the Second Conference on Human Language Technology, 2002, pp. 128-132.

[18] A. Lavie, K. Sagae and S. Jayaraman, "The significance of Recall in Automatic Metrics for MT Evaluation," in Machine Translation: From Real Users to Research, R. E. Frederking and K. B. Taylor, Eds. Berlin, Heidelberg: Springer, 2004, pp. 134-143.

[19] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics, 2005, pp. 65-72.

[20] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," Proceedings of the Sixth Workshop on Statistical Machine Translation, (ACL), 2011, pp. 85-91.

[21] A. Agarwal and A. Lavie, "METEOR, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output," Proceedings of the ACL 2008 Workshop on Statistical Machine Translation, 2008, pp. 115-118.

[22] Automated Community Content Editing PorTal (ACCEPT), "Analysis of existing metrics and proposal for a task-oriented metric," European Community's FP7 project deliverable, 2012, available at: http://cordis.europa.eu/docs/projects/cnect/9/288769/080/deliverables/001-D91AnalysisofexistingmetricsandproposalofataskorientedmetricI.pdf

[23] J. P. Turian, L. Shen and I. D. Melamed, "Evaluation of machine translation and its evaluation", Proceedings of the 9th Machine Translation Summit, 2003, pp. 386-393.

[24] A. Lavie, "Evaluating the Output of Machine Translation Systems," AMTA Tutorial, 2010, p. 86, available at: http://amta2010.amtaweb.org/AMTA/papers/6-04-LavieMTEvaluation.pdf