Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs

Marija Brkić[1], Sanja Seljan[2], and Maja Matetić[1]

[1] University of Rijeka, Department of Informatics, Croatia
mbrkic@uniri.hr; maja.matetic@ri.t-com.hr
[2] University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information Sciences, Croatia
sseljan@ffzg.hr

Abstract. This paper presents a bidirectional machine translation evaluation study for the Croatian-English and English-Croatian language pairs. Translations from Croatian into English have been obtained in four different domains from four online machine translation services, i.e. Google Translate, Stars21, Translation Guide and InterTran. These translations have been evaluated by three different automatic accuracy metrics, i.e. F-measure, BLEU and NIST, as well as by human evaluators. Evaluations are based on a single reference per sentence. In the reverse translation direction, Google Translate output has been analyzed in the same manner. System level correlation between F-measure, BLEU, NIST and human assessments is given and the significance of the results is discussed.

Keywords: online MT (machine translation), manual evaluation, automatic evaluation, F-measure, BLEU, NIST

1 Introduction

A large-scale experiment which measures how strongly 26 automatic metrics correlate with human assessments of translation quality for five European languages is presented in [1]. The aim of this work is to evaluate the online available machine translation (MT) services for the Croatian-English language pair and vice versa, and to see how well the selected automatic evaluation metrics, which are unforgiving for morphological errors, correlate with human assessments.

Evaluation methods can be manual or automatic. Nevertheless, both categories are extremely subjective [2]. The quality of automatic measures can only be determined by comparison to human assessments [3]. Human assessments are considered gold standard for evaluation. However, they are expensive with

respect to time and money [4]. Automatic metrics have many advantages compared to human assessments. Besides the fact that they are useful for comparing the performance of different systems on a common translation task, they are extremely useful during system development because they are fast and have low-cost [5]. The correlation between two metrics is usually computed using the Pearson correlation coefficient.

The automatic evaluation scores guide the development of the MT system based on concrete performance improvements [5]. The system is tested multiple times on one distinct set of test sentences, either for adjusting parameter settings or for examining the effects of system-design changes [6]. An example that follows illustrates this process in the field of statistical MT (SMT). The basic idea behind phrase-based SMT is to segment source sentences into phrases, translate each phrase and compose target sentences from these phrase translations. In other words, there are three components that contribute to producing the best possible translation—the phrase translation table, the reordering sub-model and the language sub-model [7]. If only lower quality bilingual data is available, the system needs to rely more on the monolingual target language sub-model, meaning that sometimes a sub-model needs to be given more weight [8]. The automatic evaluation scores guide the setting of weights.

The second section of this paper highlights desirable properties of any automatic evaluation metric and focuses on the three metrics most heavily used in MT community. In the third section a detailed description of the conducted study is given and the results are presented. Section four discusses the obtained results. The findings and directions for future work are summarized in the conclusion.

2 MT Evaluation

MT evaluation should be able to determine semantic equivalence or similarity between sentences, which makes it a hard problem. This is evident from the fact that any number of different translators translates the very same sentence differently. Besides determining semantic equivalence, desirable properties of an evaluation metric are that it is tunable, meaningful, consistent, correct, reliable, general, and has low cost [7]. A metric is tunable if system performance can be optimized towards it. If it gives intuitive interpretation of translation quality, it is meaningful. A consistent metric gives the same results by repeated usage, i.e. inter-annotator agreement. If better systems are ranked higher, then the metric is also correct [7]. MT systems that score similarly also perform similarly if the metric is reliable. Furthermore, a metric should be as sensitive as possible to differences in MT quality between different systems, and between different versions of the same system. Finally, if a metric is applicable to different MT tasks in a wide range of domains and scenarios, it has appropriate generalization power [5]. Turian et al. add reliability on shorter texts as another desirable

property. However, MT evaluation metrics are usually less reliable on shorter translations. As the most important criterion, they point out the ability to rank the systems the same way human evaluators would rank them.

2.1 Automatic Evaluation

An automatic evaluation metric should besides the quality of translation satisfy some extra requirements. It should be fast, easily integrated into the existing workflow, customizable, and its memory requirements must not go beyond memory available on the machines under consideration [7]. Before scoring with an automatic evaluation metric, the translated text and the reference translations, i.e. one or more human translations of the same sentence, are conditioned to improve the efficiency of the algorithm, e.g. case information is removed, numerical information is kept together as single words, punctuation is tokenized into separate words, and adjacent non-ASCII words are concatenated into single words [9]. All automatic evaluation metrics use one or more reference translations. These reference translations are used for comparison with the MT output or hypothesis translations. Automatic metric is considered better if it has higher degree of correlation with human assessments [7]. There are numerous automatic metrics, such as Word Error Rate (WER) [6], Position-independent Word Error Rate (PER) [4], Translation Edit Rate (TER) [3], F-measure [10], Bilingual Evaluation Understudy (BLEU) [11], NIST [9], ROUGE [12], and METEOR [5]. One of the first automatic evaluation methods applied to SMT, WER, which is borrowed from speech recognition, is based on the Levenshtein distance [7]. WER, TER and PER are error measures, while the rest of the metrics fall into the category of accuracy measures [4]. The metrics differ in the way they measure similarity. However, the hypothesis translation which is closer to reference translation is ranked better by all of the metrics [2].

The rest of this section describes the three selected fully-automatic accuracy evaluation metrics, F-measure, BLEU and NIST. Due to the limited scope of this work, the remaining metrics have been excluded from further study, and will be dealt with in the future. METEOR has been excluded due to the lack of necessary language tools.

F-measure. The measures that are typically used in evaluation are precision, recall and F-measure (1). Precision is the percentage of generated words that are actually correct. Recall stands for the percentage of words that are generated and that are actually found in the reference translation. F-measure is the harmonic mean of recall and precision [7]. It is also known as the F1-measure because precision and recall are evenly weighted.

$$F\text{-}measure = \frac{precision \times recall}{(precision + recall)/2} \quad . \tag{1}$$

BLEU. BLEU ranks MT output according to a weighted average of the number of n-gram overlaps with the reference translations [2]. It is based on the modified unigram precision, which relies on the notion that a reference word should be considered exhausted after a matching hypothesis word is identified. The total count of each hypothesis word is clipped by the maximum number of times the word appears in any of the reference translations. These clipped counts are added and divided by the total number of hypothesis words. Modified n-gram precision is computed by analogy. Modified unigram precision accounts for adequacy, while modified n-gram precision accounts for fluency. A modified precision score, p_n, for the entire corpus, is calculated as in (2).

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} count_{clip}(n\text{-}gram)}{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} count(n\text{-}gram)} \quad . \tag{2}$$

The clipped n-gram counts for all the sentences are added and divided by the number of hypothesis n-grams in the test corpus. N-gram matches are, therefore, computed at the sentence level, but the modified n-gram precision is the fraction of n-grams matched in the entire corpus. A weighted linear average of the modified precisions enables combining of the modified precisions for various n-gram sizes. However, the modified unigram precision is much larger than the modified bigram precision, etc. In order to take this exponential decay into account, a weighted average of the logarithm of the modified precisions is calculated. The brevity penalty (3) is a multiplicative factor which penalizes hypotheses shorter than their reference translations. It is computed over the entire corpus to allow some freedom at the sentence level [11]. Main critiques directed towards this metric are that it ignores the relative relevance of words, it does not address the overall grammatical coherence, the actual BLEU scores are meaningless, and human BLEU scores are barely higher than that of an MT system, although the translations are of much higher quality [7]. Furthermore, BLEU is quite unintuitive and relies upon a large number of sentences in order to correlate with human assessments [3].

$$BP = \begin{cases} 1 & \text{if } outputLength > referenceLength \\ e^{(1-\frac{referenceLength}{outputLength})} & \text{if } outputLength \leq referenceLength \end{cases} \quad . \tag{3}$$

ReferenceLength is the test corpus effective reference length (sum of the best match lengths which are the closest reference translation lengths; if there are two lengths equally close, the shorter one is taken), and *outputLength* is the total length of the hypothesis translation corpus. Finally, the BLEU metric is defined as in (4) [9].

$$BLEU = \exp\left\{ \sum_{n=1}^{N} w_n \log p_n - \max\left(0, \frac{L^*_{ref}}{L_{sys}} - 1\right) \right\} \quad . \tag{4}$$

NIST. Since IBM showed a strong correlation between BLEU scores and human assessments of translation quality, DARPA commissioned NIST to develop an MT evaluation facility based on the IBM work. Since BLEU uses a geometric mean of co-occurrences over N, the score is equally sensitive to proportional differences in co-occurrence for all N. This might lead to counterproductive variance due to low co-occurrences for the larger values of N. This problem is overcome by using an arithmetic mean of n-gram counts. Furthermore, n-grams that are more informative, i.e. that occur less frequently, deserve more weight (5) [9]. The formula for calculating NIST score is given in (6), where the ratio used in minimization stands for the number of words in the translation being scored and the average number of words in a reference translation, averaged over all reference translations. Factor β is chosen to make the brevity penalty factor 0.5 when the number of words in the system output is two-thirds of the average number of words in the reference translation, and when N equals to 5. A change in the brevity penalty factor is made to minimize the impact of small variations in the length of a translation [9].

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{\#w_1 \dots w_{n-1}}{\#w_1 \dots w_n} \right) \quad . \tag{5}$$

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum_{co\text{-}occur} Info(w_1 \dots w_n)}{\sum_{output}(1)} \right\} \times \exp \left\{ \beta \log^2 \left[min \left(\frac{L_{sys}}{\overline{L}_{ref}}, 1 \right) \right] \right\} \quad . \tag{6}$$

3 Experimental Study

The study has been divided into two parts. In the first part translations from Croatian into English have been obtained from four online MT services, i.e. Google Translate, Stars21, Translation Guide and InterTran. Google Translate currently supports translation between 57 languages. Croatian has been supported since 2008. Stars21 for the Croatian-English language pair offers services from Google Translate, InterTran or TranStar, which is again powered by Google Translate. In spite of that, the reported results for Google Translate and TranStar are not exactly the same, which can most probably be explained by different pre- or postprocessing techniques used by these services. The service offered by Translation Guide is powered by InterTran, which, on the other hand, is powered by WordTran and NeuroTran. WordTran consists of word-by-word and phrase-by-phrase translations. NeuroTran is a rule-based system which takes care of inflections and word-order. Although they are powered by the same engine, the translations obtained from Translation Guide and InterTran differ somewhat. To put it more precisely, Translation Guide has much higher percentage of untranslated words, especially words with diacritics. This can again be explained by different pre- or postprocessing techniques.

Source texts are short excerpts from four different domains, i.e. city, law, football and monitors. These excerpts contain 9, 9, 7 and 9 sentences, respectively.

The obtained translations have been evaluated by three different automatic metrics, i.e. F-measure, BLEU and NIST, as well as by 48 translators on a 1-5 scale according to two criteria, fluency and adequacy. The evaluators have scored the MT output according to the fluency criterion without seeing the source sentences or reference translations. Next, they have scored the output according to the adequacy criterion with regard to the source sentences. Since the two criteria are usually related, we have taken the average of the two.

According to human assessments, TranStar proves to be the best system with the score of 4.66 when the score is averaged over all domains, and Translation Guide the worst with the score of 1.17. Google Translate performs slightly worse with the score of 4.62, and beats TranStar only in the football domain, which is also the domain with the highest score achieved.

The same study has been conducted in the reverse direction, but has included only the popular Google Translate system. Excerpts from all four domains contain 9 sentences each. Altogether 50 human assessments have been collected and the obtained score averaged over all four domains is 4.29. The best scored domain is the city domain, and the worst scored is the football domain.

Prior to running automatic evaluation, lowercasing and tokenization have been done. All of the calculations are based on a single reference per sentence. The results are presented in the subsequent subsections.

3.1 F-measure

F-measure ranges from 0 to 1. Individual F-measure scores obtained for translations from Croatian into English and vice versa are presented in Table 1. Overall F-measure scores obtained for translations from Croatian into English are presented in Table 2. The overall F-measure score obtained for Google Translate system for the English-Croatian language pair is 0.7224.

Table 1. F-measure obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**0.6945**	0.6301	**0.7999**	**0.8873**
	TranStar	**0.6945**	**0.6475**	0.7857	**0.8873**
	Translation Guide	0.1957	0.1458	0.2281	0.2166
	InterTran	0.3754	0.4082	0.4109	0.3343
EN → CRO	Google Translate	0.8030	0.7599	0.6373	0.6781

Table 2. Overall scores obtained for four different MT systems for the Croatian-English language pair.

CRO → EN	F-measure	BLEU	NIST
Google Translate	0.7348	**0.5383**	7.2234
TranStar	**0.7376**	0.5337	**7.2596**
Translation Guide	0.1907	0.0551	2.4969
InterTran	0.3863	0.0873	3.5205

3.2 BLEU

Possible BLEU scores range from 0 to 1. Individual BLEU scores obtained for translations from Croatian into English and vice versa are presented in Table 3. Overall BLEU scores obtained for translations from Croatian into English are presented in Table 2. The overall BLEU score obtained for Google Translate system for the English-Croatian language pair is 0.5836.

Table 3. BLEU scores obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**0.5050**	0.3719	**0.5941**	**0.7796**
	TranStar	**0.5050**	**0.3957**	0.5402	**0.7796**
	Translation Guide	0.0576	0.0405	0.0708	0.0530
	InterTran	0.1144	0.0814	0.1003	0.0545
EN → CRO	Google Translate	0.7102	0.5577	0.5290	0.5311

3.3 NIST

A NIST score of 0 means that the hypothesis and the reference have no n-grams in common. Higher positive scores suggest better translations. Individual NIST scores obtained for translations from Croatian into English and vice versa are presented in Table 4. Overall NIST scores obtained for translations from Croatian into English are presented in Table 2. The overall NIST score obtained for Google Translate system for the English-Croatian language pair is 7.2016.

Table 4. NIST scores obtained for four different MT systems in four different domains.

		Domain			
	System	City	Law	Football	Monitors
CRO → EN	Google Translate	**5.5714**	4.8921	**6.0526**	**6.7125**
	TranStar	**5.5714**	**5.0639**	5.9828	**6.7125**
	Translation Guide	2.0486	1.7176	2.0723	2.3891
	InterTran	3.1390	2.8658	3.1540	2.6141
EN → CRO	Google Translate	6.2356	6.0487	5.0035	5.2658

4 Discussion

4.1 Croatian-to-English Translation Task

According to all automatic measures, Google Translate and TranStar are best suited for translating technical manuals, i.e. monitors, from Croatian into English, and worst suited for translating legal documents. Surprisingly, human evaluators find Google best suited for translating in the domain of football, and TranStar best suited for translating legal documents for translation from Croatian into English. Both are worst suited for literary descriptions, i.e. city information. All of the metrics, as well as human evaluators almost completely agree on the system rankings. The only disagreement shows BLEU which gives the highest rank to Google Translate instead of TranStar (Fig. 1). System level correlation between F-measure, BLEU, NIST and human assessments for the Croatian-to-English translation task is given in Fig. 2. A correlation of 1 means that there is a positive linear relationship between the two variables, a correlation of -1 means that there is a perfect negative linear relationship between them, and a correlation of 0 means that there is no linear relationship between them. The correlation between F-measure and BLEU is high for Google Translate and TranStar, somewhat lower for Translation Guide, and the lowest for InterTran. The same applies to the correlation between F-measure and NIST. The correlation between BLEU and NIST is extremely high for all but the worst ranking system Translation Guide. The correlation between automatic metrics and human assessments is much lower. The strongest correlation between all three automatic metrics and human assessments shows Google Translate. The least strong correlation between BLEU and human assessments shows, surprisingly, TranStar. The correlation between automatic metrics and human assessments averaged over all the systems for the Croatian-to-English translation task is shown in Table 5. The significance of the correlations has been tested through a two-tailed test at the 0.05 significance level with two degrees of freedom. The results in Table 5 are not statistically significant. The correlation between F-measure and BLEU, as well as F-measure and NIST, is statistically significant

for Google Translate and TranStar, while the correlation between BLEU and NIST is significant for all but InterTran. None of the automatic metrics significantly correlates with human assessments.

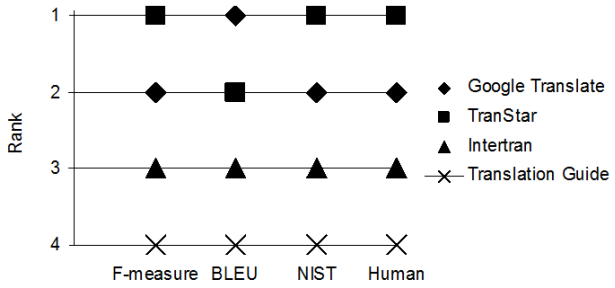

Fig. 1. Ranking of four systems in the Croatian-to-English translation task (Google Translate, TranStar, Translation Guide and InterTran) according to three automatic metrics (BLEU, NIST and F-measure) and human assessments.

Table 5. Correlation between automatic metrics and human assessments averaged over all systems for the Croatian-to-English translation task.

	F-measure	BLEU	NIST
BLEU	0.8490		
NIST	0.8665	0.8441	
Human	0.3182	0.3040	0.2135

4.2 English-to-Croatian Translation Task

The correlation between automatic metrics and human assessments for translations from English into Croatian is given in Table 6. According to the two-tailed significance test at level 0.05 with two degrees of freedom, only the correlation between NIST and F-measure is statistically significant. The correlation for each criterion separately, i.e. fluency and accuracy, is shown in Fig. 3. Lastly, the scores for each metric for both translation directions are shown in Fig. 4. We observe that there is a negative linear relationship between the two translation directions, however, this relationship is not statistically significant.

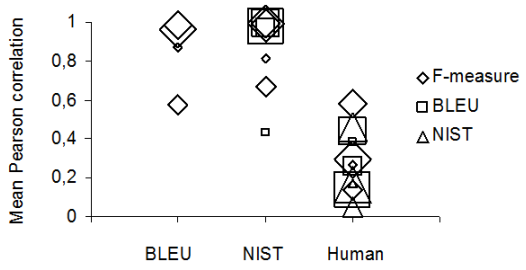

Fig. 2. System-level correlation between F-measure, BLEU, NIST and human assessments in the Croatian-to-English translation task. The size of a shape for a system depends on the majority-agreement-ranking of four systems, where the best system, i.e. TranStar, has the biggest shape.

Table 6. Correlation between automatic metrics and human assessments averaged over all systems for the English-to-Croatian translation task.

	F-measure	BLEU	NIST
BLEU	0.8272		
NIST	0.9933	0.7712	
Human	0.7437	0.6766	0.6809

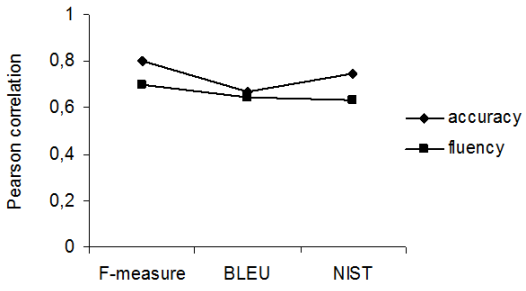

Fig. 3. System-level correlation between automatic metrics and human accuracy and fluency assessments in the English-to-Croatian translation task.

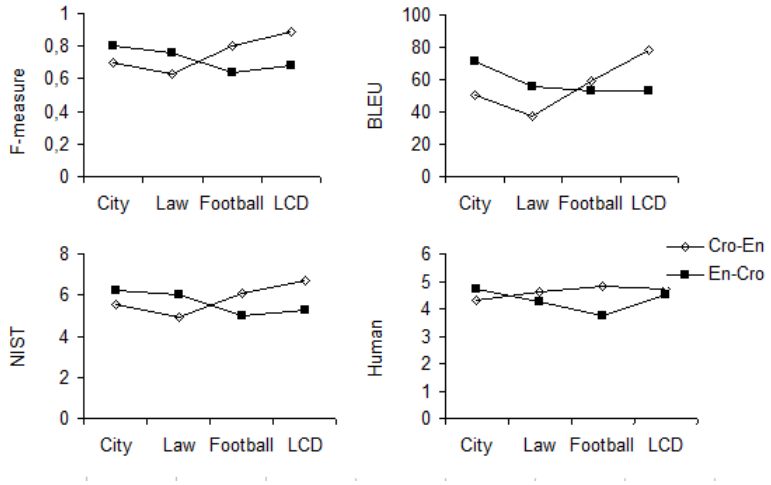

Fig. 4. Google Translate correlation between two translation directions according to three automatic metrics and human assessments.

5 Conclusion

In the first part of the study translations from Croatian into English have been obtained from four MT systems, i.e. Google Translate, TranStar, Translation Guide and InterTran in four different domains, and scored by the three fully-automatic accuracy evaluation metrics. All of the metrics, as well as human evaluators, almost completely agree on the rankings of the systems. The correlation between F-measure and BLEU, as well as F-measure and NIST, is statistically significant for Google Translate and TranStar, while the correlation between BLEU and NIST is significant for all but InterTran. In the reverse translation direction only Google Translate output has been evaluated and the correlation between F-measure and NIST proved to be statistically significant. However, none of the automatic metrics significantly correlates with human assessments. This might be due to the size of the test set, which is limited due to a lack of human evaluators and to the time-consuming nature of manual evaluation task. F-measure has the highest correlation with human assessments for the two highest ranking systems in the Croatian-to-English translation task, as well as for the reverse translation direction. For the worse two systems BLEU correlates with human assessments the best. Although not statistically significant, we observe that there is a negative linear relationship between the two translation directions.

We conclude that adding more reference translations might improve reliability of the automatic metrics, since a hypothesis that might be perfectly correct

is scored badly if it differs a lot from a reference translation. In our future work we will investigate the correlation between the remaining metrics in the field of MT evaluation.

References

1. Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.: Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 17–53. Association for Computational Linguistics, Uppsala, Sweden (2010)
2. Jurafsky, D., Martin, J., Kehler, A.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall (2009)
3. 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, pp. 223–231. Cambridge, Massachusetts (2006)
4. Leusch, G., Ueffing, N., Ney, H.: A Novel String-to-string Distance Measure with Applications to Machine Translation Evaluation. In: Proceedings of MT Summit IX, pp. 240–247. New Orleans, Louisiana (2003)
5. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Ann Arbor, Michigan (2005)
6. Nießen, S., Och, F., Leusch, G., Ney, H.: An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 39–45. Athens, Greece (2000)
7. Koehn, P., Corporation: Statistical Machine Translation, Cambridge University Press (2010)
8. Mauser, A., Hasan, S., Ney, H. Automatic Evaluation Measures for Statistical Machine Translation System Optimization. In: International Conference on Language Resources and Evaluation, pp. 3089–3092. Marrakech, Morroco (2008)
9. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In: Proceedings of the 2nd International Conference on Human Language Technology Research, pp. 138–145. Morgan Kaufmann Publishers Inc., San Francisco, California (2002)
10. Turian, J., Shen, L., Melamed, I.: Evaluation of Machine Translation and Its Evaluation. In: Proceedings of the MT Summit IX, pp. 386–393. New Orleans (2003)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania (2002)
12. Lin, C., Och, F.: Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 605–612. Association for Computational Linguistics, Barcelona, Spain (2004)