Automatic and Human Evaluation on English-Croatian Legislative Test Set

Marija Brkić[1], Sanja Seljan[2], and Tomislav Vičić[3]

[1] University of Rijeka, Department of Informatics, Omladinska 14, 51000 Rijeka, Croatia, mbrkic@uniri.hr
[2] Faculty of Humanities and Social Sciences, Department of Information Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia, sanja.seljan@ffzg.hr
[3] Freelance translator, ssimonsays@gmail.com

Abstract. This paper presents work on the manual and automatic evaluation of the online available machine translation (MT) service Google Translate, for the English-Croatian language pair in legislation and general domains. The experimental study is conducted on the test set of 200 sentences in total. Human evaluation is performed by native speakers, using the criteria of fluency and adequacy, and it is enriched by error analysis. Automatic evaluation is performed on a single reference set by using the following metrics: BLEU, NIST, F-measure and WER. The influence of lowercasing, tokenization and punctuation is discussed. Pearson's correlation between automatic metrics is given, as well as correlation between the two criteria, fluency and adequacy, and automatic metrics.

1 Introduction

Evaluation of machine translation (MT) is an extremely demanding task. Besides being time-consuming and subjective, there is no uniform opinion on "good quality" translation. However, the human translation, i.e. reference translation, is considered to be a "gold standard". There may be more than one reference translation set. Automatic evaluation metrics rely on different approaches, which all aim at performing evaluation as close as possible to human evaluation. The goal of evaluation can be comparing outputs of a single MT system through different phases, i.e. testing different parameter settings or system changes; comparing different systems based on different approaches; comparing similar systems, etc. Evaluation can be performed within a domain or across different domains. Automatic evaluation for morphologically rich under-resourced languages presents a domain of interest for researchers, educators and everyday users, especially when the language is to become one of official EU languages.

2 Related work

A number of studies have explored correlation between human and automatic evaluation and conducted error analysis, especially for widely spoken languages.

Qualitative analysis of MT output on a test set might point out some important general or domain-specific linguistic phenomena, especially when dealing with morphologically rich languages. In [12] the importance of qualitative view and the need for error analysis of MT output is pointed out. In [6] the complexity of MT evaluation is discussed and a framework for MT evaluation is defined, which relates the quality model to the purpose and the context, enabling evaluators to define usage context out of which a relevant quality model is generated. The main purpose is creating a coherent picture of various quality characteristics and metrics, providing a common descriptive framework and vocabulary, and unifying the evaluation process. [5] suggests a classification system of MT errors designed more for MT users than for MT developers. Error categories can be ranked according to the level of importance they have in the eyes of users, with regard to, for example, improvability and intelligibility. In [14] the relationship between automatic evaluation metrics (WER, PER, BLEU, and NIST) and errors found in translation is discussed. Errors are split into five classes: missing words, word order, incorrect words, unknown words and punctuation errors. The relationship between BLEU as an automatic evaluation measure and the expert human knowledge about the errors is discussed in [4]. Their results point to the fact that linguistic errors might have more influence on perceptual evaluation than other errors. Callison-Burch et al. in [1] evaluate MT output for 8 language pairs and conduct human evaluation in order to obtain different systems ranking and higher-level analysis of the evaluation process, and to calculate correlation of automatic metrics with human evaluation. Correlation between human evaluation of MT output and automatic evaluation metrics, i.e. BLEU and NIST, is explored in [2].

3 Evaluation metrics

Four automatic metrics presented in subsequent sections are widely used in MT evaluation. However, there are not many researches on the evaluation of Croatian MT output, whereas Croatian is a highly inflected less widely spoken language that belongs to a group of Slavic languages. In Croatian, each lemma has many word forms, i.e. on average 10 different word forms for nouns, denoting case, number, gender and person. In this experimental study, GT-translated text has been evaluated by native speakers, errors have been analyzed, and, finally, correlation between automatic metrics separately, as well as between automatic metrics and human evaluation is given.

3.1 BLEU

Bilingual Evaluation Understudy (BLEU) is based on matching candidate n-grams with n-grams of the reference translation [11]. Scores are calculated for each sentence, and then aggregated over the whole test set. The algorithm calculates modified precisions in order to avoid MT over-generation of n-grams. For each candidate translation n-gram, BLEU takes into account the maximum

number of times the n-gram appears in a single reference translation, i.e. the total count of each n-gram is clipped by its maximum reference count. The clipped counts are summed together and divided by the total number of n-grams in the candidate translation. Unigram precisions account for adequacy, while n-gram precisions account for fluency. In order to avoid too short candidates, the multiplicative brevity penalty factor is introduced. Some of the critiques directed towards BLEU are that it does not take into account the relative relevance of words, the overall grammatical coherence, it is quite unintuitive, and relies on the whole test set in order to correlate well with human judgments [8].

3.2 NIST

National Institute of Standards and Technology (NIST) is based on BLEU metric, but it introduces some changes. While BLEU gives the same weight to each n-gram in the candidate translation, NIST calculates how informative that n-gram is, namely the rarer the n-gram appears, the more informative it is, and more weight will be given to it. NIST also differs from BLEU in the calculation of brevity penalty factor, which does not influence result as much as the one in BLEU [3].

3.3 F-measure

F-measure is widely used not only in MT, but also in information and document retrieval. This is the measure of accuracy which takes into account precision and recall, namely F-measure is a weighted average of both. It ranges from 0 to 1, 1 being the best value [10].

3.4 WER

Word Error Rate (WER) is a reference translation length-normalized Levenshtein distance [9]. Borrowed from speech recognition, it is one of the first metrics applied to statistical machine translation (SMT). Levenshtein distance can be defined as the minimum number of insertions, deletions and substitutions needed on a candidate or hypothesis translation so that it matches the reference translation [8]. WER is often criticized for being too harsh on word order. Namely, it does not allow any reordering [13]. If a candidate is exactly the same as its reference translation, WER equals to 0. Furthermore, it can be even bigger than 1 if a candidate is longer than its reference translation.

4 Experimental study

4.1 Testset descriptions

One part of the research has been conducted on English-Croatian parallel corpora of legislative documents, available at `http://eur-lex.europa.eu/` and

http://ccvista.taiex.be/. However, some additional editing has been deemed necessary for documents containing mostly tables and formulas, not usable for analysis, as well as typos and misspellings. For the purpose of analysis a total of 100 source sentences have been extracted, together with their reference translations. MT translation candidates have been obtained from Google Translate (GT) service, which has Croatian language support among others. Another part of the research has been conducted on the test set compiled from professional translations in different domains, i.e. religion, psychology, education, etc. 100 sentences have been extracted. The test set descriptions are given in Table 1.

Table 1. # of words in testset descriptions

	source	reference	translation
Testset 1	2.121	1.700	1.725
Testset 2	1.660	1.467	1.440

4.2 Human evaluation

Human evaluation has been performed according to the criteria of fluency and adequacy, through an online survey. The survey has consisted of two polls for each criterion. Possible evaluation grades for fluency have been: Incomprehensible (1), Disfluent (2), Non-native (3), Good (4), Flawless (5). Adequacy evaluation grades having been: None (1), Little (2), Much (3), Most (4), All information preserved (5). The average obtained grade is 3.03 for fluency and 3.04 for adequacy on testset 1, and 3.30 for fluency and 3.67 for adequacy on testset 2.

4.3 Error analysis

GT-translated sentences have been compared to the reference sentences. Although there have been many cases of several types of errors in a single sentence, the following errors have been distinguished: not translated/omitted words, surplus words in a translation, morphological errors/suffixes, lexical errors – wrong translation, syntactic errors – word order, and punctuation errors. The analysis has shown the highest number of morphological errors (on average 1.45 per sentence in testset 1 and 1.87 in testset 2), while other types of errors have been less represented. The next most represented error category has been that of lexical errors (on average 0.73 errors per sentence in testset 1 and 0.59 in testset 2), not translated words 0.41 errors per sentence in testset 1 and 0.4 in testset 2) and syntactic errors (0.48 errors per sentence in testset 1 and 0.47 in testset 2). The categories with the smallest number of errors detected have been surplus words (0.29 per sentence in testset 1 and 0.26 in testset 2) and punctuation errors (0.17 per sentence in testset 1 and 0.01 in testset 2).

4.4 Results

While in the first part of the experiment automatic scores have been configured to include case information, in the second part of the experiment case information has been omitted (Tables 2 and 3). The prefix l denotes case-insensitive part of the evaluation. The confidence intervals for BLEU and NIST have been calculated by bootstrapping and all the scores lie within the 95% interval [7].

Table 2. Automatic evaluation scores on testset 1 with respect to lowercasing, tokenization and punctuation removal

	no-preprocessing	tokenization	tok. and punct. removal
WER	76.12	57.20	58.78
lWER	75.76	56.50	57.62
F-measure	35.13	57.16	54.32
lF-measure	35.78	58.16	55.42
BLEU	33.70	33.64	31.61
lBLEU	34.32	34.25	32.19
NIST	6.2586	6.2539	6.0314
lNIST	6.3321	6.3271	6.1098

Table 3. Automatic evaluation scores on testset 2 with respect to lowercasing, tokenization and punctuation removal

	no-preprocessing	tokenization	tok. and punct. removal
WER	66.55	59.60	62.31
lWER	66.22	59.30	62.13
F-measure	47.74	55.82	51.89
lF-measure	48.89	56.83	53.11
BLEU	31.11	31.06	26.57
lBLEU	31.60	31.55	26.98
NIST	6.2628	6.2629	5.8507
lNIST	6.3432	6.3432	5.9309

5 Discussion

Before scoring with an automatic metric, the translated set and the reference set are usually preprocessed in order to improve the efficacy of the scoring algorithm [3]. Preprocessing usually implies lowercasing and tokenization. In addition to

these two steps, we have added punctuation removal, and explored how these aspects affect the scores according to four automatic metrics. Lowercasing has systematically improved scores slightly. While tokenization has had enormous beneficial effect on WER and F-measure scores, especially for testset 1, i.e. the WER scores have dropped down for about 20 points, the F-measure scores have gone up for about 22 points, BLEU and NIST scores have slightly deteriorated. This is due to the fact that the script used for calculating these scores performs internal tokenization which proved to be more beneficial than the one performed explicitly. Removing punctuation has had detrimental effect on all the scores, which has been expected because punctuation is translated more correctly. WER as an error measure has increased for more than 1 point compared to the tokenized testset 1 score, and for about 3 points on testset 2 score, irrespective of the case-sensitivity. The other three metrics scores have decreased, even more so on testset 2.

Pearson's correlation between WER and F-measure, as far as tokenization effects on true cased and lowercased test set are concerned, has proven statistically significant according to a two-tailed test at 0.05 significance level. As far as punctuation and tokenization is concerned, correlation between WER and F-measure, in addition to the correlation between BLEU and NIST, has proven statistically significant. Furthermore, WER and F-measure scores without punctuation have still beaten the baseline scores, i.e. the scores without tokenization and with punctuation included.

When all three aspects are taken into consideration, only WER and F-measure, as well as BLEU and NIST significantly correlate. WER and F-measure completely agree on the rankings of preprocessing techniques, while NIST seems to be less sensitive to tokenization when compared to BLEU.

The results indicate that when calculating WER and F-measure, an important pre-processing step should be tokenization, followed by lowercasing. As far as BLEU and NIST are concerned, lowercasing has proven to be of the biggest importance. However, all the above findings should be checked against correlation with human judgments.

For that purpose, we have divided our test sets into 5 different test sets, each containing 40 sentences, and calculated the correlation between human and automatic scores, with the above described aspects taken into consideration. None of the calculated correlations is statistically significant. We have also observed that NIST correlates much better with human adequacy, than human fluency scores, as in [3]. In our future work, we intend to explore correlations with human judgments in more detail.

References

[1] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. (meta-)evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, 2007.

[2] D. Coughlin. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX*, pages 63–70, 2003.

[3] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

[4] M. Farrús Cabeceran, M. Ruiz Costa-Jussà, J.B. Mariño Acebal, J.A. Rodríguez Fonollosa, et al. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of EAMT*, pages 52–57, 2010.

[5] M. Flanagan. Error classification for mt evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 65–72, 1994.

[6] E. Hovy, M. King, and A. Popescu-Belis. Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75, 2002.

[7] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, volume 4, pages 388–395, 2004.

[8] P. Koehn. *Statistical Machine Translation*, volume 11. Cambridge University Press, 2010.

[9] G. Leusch, N. Ueffing, H. Ney, et al. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit IX*, pages 33–40, 2003.

[10] I.D. Melamed, R. Green, and J.P. Turian. Precision and recall of machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*, pages 61–63. Association for Computational Linguistics, 2003.

[11] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[12] S. Stymne. Blast: A tool for error analysis of machine translation output. *Proc. of the 49th ACL, HLT, Systems Demonstrations*, pages 56–61, 2011.

[13] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated dp based search for statistical translation. In *European Conf. on Speech Communication and Technology*, pages 2667–2670, 1997.

[14] D. Vilar, J. Xu, L.F. d'Haro, and H. Ney. Error analysis of statistical machine translation output. In *Proceedings of LREC*, pages 697–702, 2006.