BLEU Evaluation of Machine-Translated English-Croatian Legislation

Sanja Seljan[1], Tomislav Vičić[2], Marija Brkić[3]
[1]University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Ivana Lučića 3, 10000 Zagreb, Croatia
[2]Freelance translator
10000 Zagreb, Croatia
[3]University of Rijeka, Department of Informatics
Omladinska 14, 51000 Rijeka, Croatia
E-mail: sanja.seljan@ffzg.hr, ssimonsays@gmail.com, mbrkic@uniri.hr

Abstract

This paper presents work on the evaluation of online available machine translation (MT) service, i.e. Google Translate, for English-Croatian language pair in the domain of legislation. The total set of 200 sentences, for which three reference translations are provided, is divided into short and long sentences. Human evaluation is performed by native speakers, using the criteria of adequacy and fluency. For measuring the reliability of agreement among raters, Fleiss' kappa metric is used. Human evaluation is enriched by error analysis, in order to examine the influence of error types on fluency and adequacy, and to use it in further research. Translation errors are divided into several categories: non-translated words, word omissions, unnecessarily translated words, morphological errors, lexical errors, syntactic errors and incorrect punctuation. The automatic evaluation metric BLEU is calculated with regard to a single and multiple reference translations. System level Pearson's correlation between BLEU scores based on a single and multiple reference translations is given, as well as correlation between short and long sentences BLEU scores, and correlation between the criteria of fluency and adequacy and each error category.

Keywords: BLEU metric, English-Croatian legislation, human evaluation

1. Introduction

Evaluation of machine translation (MT) web services has gained considerable attention lately, because of their more widespread usage in accessing information in a foreign language by students, researchers, patients, teachers, everyday users, etc. Comparisons between human and different automatic metrics, error analysis, suggestions for improvement have become a logical follow-up.

Although results of MT web services oscillate from "laughably bad" to "a tremendous success" (Hampshire, 2010), most of them aim to achieve reasonably good quality (although the notion of "good quality" is a question *per se*). An assessment of machine translated text is important for product designers, professional translators and post-editors, project managers, private users, as well as in education and research. The issue of "good" translation is often discussed, as well as the consensus on the agreement on various evaluation criteria (fluency, adequacy, meaning, severity, usefulness, etc.) and subjective evaluation approach.

Evaluation in MT research and product design can be done with the aim of measuring system performance (Giménez and Màrquez, 2010; Lavie and Agarwal, 2007) or with the aim of identifying weak points and/or adjusting parameter settings of different MT systems or of a single system through different phases (Denkowski and Lavie, 2010a; Agarwal and Lavie, 2008). Moreover, the identification of weak points might contribute to quality improvement, especially for less-resourced languages and languages with rich morphology.

2. Related Work

Google Translate (GT), being a free web service, is included in almost every research on MT evaluation, especially because it offers translation from and into less widely spoken languages.

In the study presented by Khanna et al. (2011), a text from one pamphlet on the importance of health care for people with limited English proficiency is selected. The text is GT-translated from English into Spanish and then compared with human professional translation.

The study presented by Shen (2010) compares three web translation services – GT, i.e. a statistically-based translation engine, Bing (Microsoft) Translator, i.e. a hybrid statistical engine with language specific rules, and Yahoo Babelfish, i.e. a traditional rule-based translation engine. While GT is preferred for longer sentences, and language combinations for which huge amount of bilingual data is provided, Microsoft Bing Translator and Yahoo Babelfish give better results on phrases having less than 140 characters and on some specific language pairs (e.g. Bing Translator on Spanish, German, and Italian; Babelfish for East Asian languages).

In Garcia-Santiago and Olvera-Lobo (2010) the quality of translating questions from German and French into Spanish by several MT services (GT, ProMT and WorldLingo) is analyzed.

Dis Brandt (2011) presents evaluation of three popular web services (GT, Inter Tran, Tungutorg) for translation from Icelandic into English.

In the study presented by Kit and Wong (2008), several web translation services (Babel Fish, GT, ProMT, SDL free translator, Systran, WorldLingo), used by law library users for translating from 13 languages into English, are discussed.

According to the research presented by Seljan, Brkić and Kučiš (2011) GT is a preferred online translation service for Croatian language. It shows better results in the Croatian-English direction (in the domains of football, law, and monitors) than in the English-Croatian direction (in the city description domain).

GT, a free MT web service, is provided by Google Inc. GT initially used Systran-based translator. Many state-of-the-art MT systems use rule-based approach, e.g. Systran, which requires a long-term work of linguists and information scientists on grammars and vocabularies. GT employs statistical approach and relies on huge quantities of monolingual texts in the target language and of aligned bilingual texts. It applies machine learning techniques to build a translation model. GT translates between more than 60 languages. Translation from and into Croatian was introduced in May 2008.

3. Automatic Evaluation

Automatic evaluation metrics compare a machine translated text to a reference translation. Their primary task is high correlation with human evaluation. Human evaluation is considered a "gold standard", however, it is a time-consuming, very subjective and expensive task. Automatic evaluation metrics are generally fast, cheap, and have minimal human labour requirements. There is no need for human bilingual speakers. However, currently used metrics do not differentiate well between very similar MT systems and give more reliable results on the whole test set than on individual sentences.

One of the most popular automatic evaluation metrics is BLEU – Bilingual Evaluation Understudy, proposed by IBM (Papineni et al., 2002), which actually represents a standard for MT evaluation. BLEU matches translation *n*-grams with *n*-grams of its reference translation, and counts the number of matches on the sentence level. These sentence counts are aggregated over the whole test set. The matches are not dependent on the position in a sentence. Adequacy is accounted for in word precision, while fluency is accounted for in *n*-gram precision. Recall is compensated by brevity penalty factor. The final BLEU score is the geometric average of modified *n*-gram precisions. BLEU scores range from 0 to 1. According to Denkowski and Lavie (2010b) in AMTA Evaluation Tutorial, BLEU scores above 0.30 generally reflect understandable translations and BLEU scores above 0.50 reflect good and fluent translations. BLEU metric, being statistically-based and language independent, does not take into account morphological variants of a word, which is an important issue for inflective languages. This metric requires exact word matches, with all matches being equally weighted.

Due to BLEU score low correlation with human adequacy and fluency judgments, Chiang et al. (2008) and Callison-Burch et al. (2006) recommend using BLEU for comparing similar systems or different versions of the same system, i.e. for what it was primarily designed.

For the above stated reasons, an evaluation of translations from English into Croatian, a morphologically rich language, with multiple reference sets is conducted. Automatic metric scores are compared to human evaluation scores. Due to the need for qualitative evaluation, human evaluation is enriched by error analysis, which might be integrated into statistical approaches (Monti et al., 2011).

4. Experimental Study

4.1 Test Set Description

This research has been conducted on already existing English – Croatian parallel corpora of legislative documents, namely http://eur-lex.europa.eu/ and http://ccvista.taiex.be/. These legislative documents are grouped according to the year of issue, and contain "duplicates" (with minor amendments, corrections, etc.). In total 200 unique source and reference translation pairs of different length and content have been chosen. However, some pre-processing has been deemed necessary (on the Croatian side), regarding typos, misspellings and other common mistakes that somehow persist despite the reviews. Furthermore, additional pre-processing has been done on documents containing mostly tables and formulas, not usable for analysis.

Out of total 200 source sentences, two groups have been distinguished – 100 short sentences (21 words or less) and 100 long sentences (between and including 22 and 61 words). For each English sentence, three Croatian reference translations have been provided, the first translation being the "official" one (Ref1). MT translations have been obtained from GT. The statistical data on the average number of words in the test set is given in Table 1.

# of sentences	Source	Ref1	Ref2	Ref3	GT
100 short	14.74	12.73	11.87	11.71	12.48
100 long	32.24	27.92	24.83	24.54	26.37
200	23.49	20.33	18.13	18.13	19.43

Table 1: Test set statistics.

The fact that Croatian is morphologically rich, unlike English, reflects in the obvious difference in the number of words in translations, compared to source sentences. On the other hand, each additional reference translation reduces the number of words by getting rid of redundancies, characteristic for legislative expressions, while preserving the meaning in full, as well as the legislative tone.

4.2. Human Evaluation

4.2.1. Profile of Evaluators

The percentage of students in the total number of evaluators is 88.64%, out of which 86.36% are on the final year of their undergraduate studies, and 13.5% are attending graduate studies. The remaining 11.36% of evaluators have finished their studies, mostly 0-7 years ago.

The self-evaluation of English language knowledge according to the Common European Framework of Reference for Languages is as follows – 0% have self-evaluated themselves for the level A1, 4.55% for A2, 15.91% for B1, 47.73% for B2, 22.73% for C1 and 9.09 for C2. The average self-evaluation grade in Croatian as their native language is 4 (on a 1-5 scale).

Regarding their experience in translating, 72.73% of evaluators translate for private purposes, 6.82% professionally, 9.05% still do not have professional experience, but are in the preparation process (therefore high level of language proficiency), and 11.36% are not in the translation business.

Regarding their experience in the use of translation tools, 60% of evaluators have already had experience in the use of free web translation services (GT, Systran, Babel Fish), 6% in the use of translation memories (SDL, Atril, Word Fast) and 6% of evaluators combine professional and free translation tools. 25.4% of evaluators still translate in the classic way, by directly typing in a text editor.

Out of the total number of evaluators who use translation technology, 60% would like to take specialization courses, and 32% have already taken courses on the use of translation tools.

When translating unknown words or syntactic structures, 40.19% use a web service, 28.04% hard copy of a dictionary, 21.50% an electronic dictionary, 5.54% a translation memory, and 3.74% a terminology database and a glossary.

4.2.2. Adequacy and Fluency

Human evaluation has been performed by native speakers of Croatian language on a 1-5 scale using the criteria of fluency and adequacy. An online survey has been prepared for separate evaluation of both, fluency and adequacy, for short, as well as long sentences in sets of 25, whereas the total number of sentences has been 200. The survey consists of 4 polls per group and per each evaluation criterion.

Fluency refers to the grammaticality and sounding "natural", while adequacy checks whether any part of a message has been lost or distorted. The evaluation of the fluency criterion has been made on the following scale: Incomprehensible (1), Barely enough comprehensible (2), So-so; in-between good and bad (3), Very good (4), Impeccable (5).

For evaluating adequacy, the following evaluation grades have been offered: Insufficient/inadequate/wrong information (1), Barely enough information (2), Intermediate level of information preserved (3), Very good but not complete (4), Complete information preserved (5).

As presented in Table 2, short sentences have obtained higher average grades than the long ones according to both criteria (about 20% higher for fluency and about 15% higher for adequacy).

# of sentences	Criterion		Average
	Fluency	Adequacy	
100 short	3.40	3.56	3.48
100 long	2.86	3.13	3.00
Average	3.13	3.35	3.24

Table 2: Average human fluency and adequacy criteria.

4.2.3. Fleiss' Kappa

Fleiss' kappa is a measure used for assessing the inter-rater agreement (1).

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \quad (1)$$

The nominator calculates the degree of agreement actually achieved above chance, and the denominator the degree of agreement attainable above chance. The score is standardized to lie on a -1 to 1 scale, where 1 indicates perfect inter-rater agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance. The interpretation of values is given in Table 3. The results are presented in Table 4. Fleiss' kappa shows almost perfect agreement for the criteria of fluency, for all sentences. The evaluation according to the criterion of adequacy shows substantial level of inter-rater agreement.

κ	Interpretation
<0	poor agreement
0.01 – 0.20	slight agreement
0.21 – 0.40	fair agreement
0.41 – 0.60	moderate agreement
0.61 – 0.80	substantial agreement
0.80 – 1.00	almost perfect agreement

Table 3: Interpretation of Fleiss' kappa values.

κ	Fluency	Adequacy	Average
Short sentences	0.90	0.67	0.785
Long sentences	0.85	0.72	0.785

Table 4: Fleiss kappa on human evaluation.

4.2.4. Error Analysis

Human evaluation is enriched by error analysis, in order to examine the influence of error types on fluency and adequacy, and to use it in further research. In the process of error analysis two professional translators have been engaged (they have not participated in the first part of the study), whose evaluation has proven exactly the same for all 200 sentences. They have reported the number of errors in the output of GT, compared to the first professional reference set. The error categories and error examples are given in Table 5, and the total number of errors per category is given in Table 6.

Errors from several different categories often appear in the same sentence. As presented in Table 6, there is by far the highest number of morphological errors, i.e. on average 2.26 errors per sentence. Short sentences have on average 1.24 morphological errors per sentence, while this number doubles in long sentences, i.e. 3.28 errors per sentence. Out of other categories, there is about 1 error or less per sentence. The error categories in the descending order according to the number of errors are as follows – morphological errors, lexical errors, syntactic errors, surplus of words, omissions and not translated words, and, lastly, punctuation.

Error category	Error example / elaboration
Not translated / omitted words	*Administration requiring the ships* translated as *Administracija zahtijeva brodova* instead of *Uprava koja od brodova zahtijeva* ili *Administracija koja zahtijeva od brodova*
Surplus of words in translation	*There may be cases* translated as *Postoji svibanj biti slučajevi* instead of *U nekim slučajevima* ili *Postoje slučajevi* (there are also morphological and lexical errors in this example)
Morphological errors / suffixes	*Decisions … should be taken unanimously* translated as *Odluke … mora biti donesena jednoglasno* instead of *Odluke … moraju biti donesene jednoglasno*
Lexical errors – wrong translation	*There may be cases* translated as *Postoji svibanj biti slučajevi* instead of *U nekim slučajevima* ili *Postoje slučajevi*
Syntactic errors – word order	*Steps should therefore be taken* translated as *Koraci stoga treba poduzeti* instead of *Stoga treba poduzeti korake*
Punctuation errors	very rare; sometimes the comma was omitted or set on the wrong place

Table 5: Error categories and error examples.

	Average number of errors per category					
# of sentences	Omissions	Surplus	Morphological	Lexical	Syntactic	Punctuation
100 short	0.27	0.27	1.24	0.73	0.5	0.09
100 long	0.59	0.61	3.28	1.19	1.17	0.37
200	0.43	0.44	2.26	0.96	0.84	0.23

Table 6: Error categories and number of errors per category.

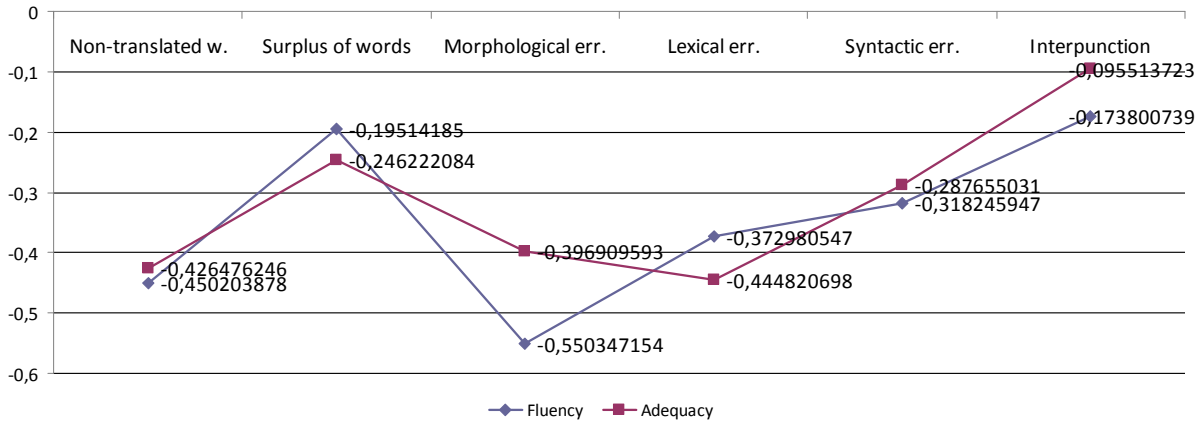

Figure 1: Correlation between fluency and adequacy criteria and error type.

4.3. Results

BLEU evaluation metric has given higher scores on short sentences. BLEU score calculated with regard to the first professional reference set is higher than for the other two reference sets, as given in Table 7. However, BLEU score gets much higher when all three reference sets are taken into account.

Due to the morphological richness of Croatian, relatively low BLEU score is obtained. Namely, in Croatian each noun has approximately 10 different word forms, which BLEU considers to be 10 different words, and not 10 different word forms of a single lemma.

By comparing BLEU scores on short sentences, we observe that the score is augmented by 19.5% when two reference sets are taken into account, instead of only one reference set. With three reference sets, the score is augmented by 27.5%. By comparing long sentences BLEU scores, the result is augmented by 22.3% when two reference sets are taken into account, instead of one. With three reference sets, the result is augmented by 29.0%. The Pearson's correlation between short sentences BLEU scores and long sentences BLEU scores, with regard to the number of reference sets, is 0.997.

	Short sentences	Long sentences
Ref 1	0.2500	0.2009
Ref 2	0.1540	0.1539
Ref 3	0.1421	0.1498
Ref 1 & Ref 2	0.2984	0.2468
Ref 1 & Ref 2 & Ref 3	0.3186	0.2592

Table 7: BLEU scores with a single and multiple reference sets.

Figure 1 shows correlation between the criteria of fluency and adequacy and different types of errors on 200 sentences from the domain of legislation translated from English into Croatian by GT. The highest negative correlation has been determined between the criteria of fluency and the number of morphological errors (-0.55), followed by non-translated and omitted words (-0.45), i.e. the greater number of these errors, the lower fluency grades. The criteria of adequacy is mostly affected by lexical errors (-0.44), closely followed by non-translated and omitted words (-0.43). The human scores normalized on a 0-1 scale, and automatics scores with respect to the number of reference translations are presented in Figure 2.

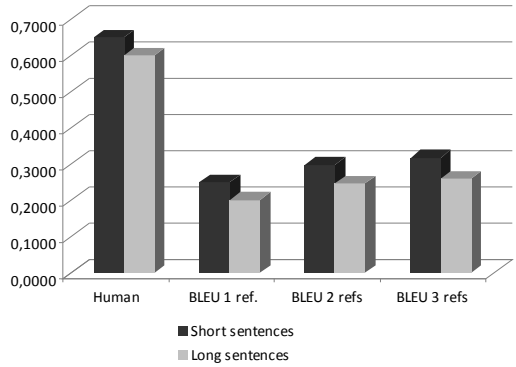

Figure 2: Human and BLEU scores on a 0-1 scale on short and long sentences separately, and with respect to the number of reference sets.

5. Conclusion

Although human evaluation is extremely subjective and time-consuming task, the average Fleiss' kappa of 0.785 shows substantial, almost perfect consistency in evaluation. The average human score on short sentences is 3.48 and on long sentences 3.00. Short sentences BLEU score is 0.25 and long sentences BLEU score is 0.20, with regard to a single reference set, i.e. 0.32 and 0.26 respectively, with regard to three reference sets.

Long sentences have gained on average 16% lower grade in human evaluation (3.00) than short sentences (3.48), and on average 22% lower BLEU score with regard to one, two, and three reference sets (0.29 for short sentences; 0.235 for long sentences). Although human and BLEU evaluation scores differ, the correlation between average BLEU scores for short and long sentences with regard to one, two and three reference sets is 0.996. With two reference sets, BLEU scores have increased for 23.5% and with three reference sets for 6.25% on average, when compared to the scores with regard to a single reference set.

Correlation between human evaluation and different error types shows that fluency is mostly affected by morphological errors (-0.55), followed by non-translated and omitted words. The criterion of adequacy is almost equally affected by lexical errors, and non-translated and omitted words.

6. Acknowledgements

The research presented here is achieved partially within the project LetsMT! that has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement no 250456 and by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grant 130-1300646-0909.

7. References

Agarwal, A., Lavie, A. (2008). METEOR. M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In *Proceedings of ACL 2008 Workshop on Statistical Machine Translation*. pp. 115--118

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics EACL*. pp. 249--256

Chiang, D., DeNeefe, S., Chan, Y.S., Ng, H.T. (2008). Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP*. pp. 610--619.

Denkowski, M., Lavie, A. (2010a). Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies NAACL/HLT*. pp. 250--253.

Denkowski, M., Lavie, A. (2010b). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. In *Proceedings of the Association for Machine Translation in the Americas AMTA*.

Dis Brandt, M. (2011). Developing an Icelandic to English Shallow Transfer Machine Translation System. Ms. Thesis. Reykjavik University.

Khanna, R.R., Karliner, L.S., Eck, M., Vittinghoff, E., Koenig, C.J., Fang, M.C. (2011). Performance of an online translation tool when applied to patient educational material. *Journal of Hospital Medicine,* 6(9), pp. 519--525.

Kit, C., Wong, T. M. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal,* 100(2), pp. 299--321.

Garcia-Santiago, L., Olvera-Lobo, M.D. (2010). Automatic Web Translators as Part of a Multilingual Question-Answering (QA) System: Translation of Questions. *Translation Journal*, 14(1).

Giménez, J., Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94, pp. 77--86.

Hampshire, S., Porta Salvia, C. (2010). Translation and the Internet : Evaluating the Quality of Free Online Machine Translators. *Quaderns: revista de traducció*, (17), pp. 197--209.

Lavie, A., Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation.* pp. 228--231.

Monti, J., Barreiro, A., Elia, A., Marano, F., Napoli, A. (2011). Taking on new challenges in multi-word unit processing for machine translation. In *Proceedings of the 2nd Workshop on Free/Open-Source Rule-Based Machine Translation*. pp. 11--19.

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pp. 311--318.

Seljan, S., Brkić, M., Kučiš V. (2011). Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. In *Proceedings of the 3rd International Conference on the Future of Information Sciences: INFuture2011 - Information Sciences and e-Society*. Zagreb, Croatia, pp. 331--345.

Shen, E. (2010). Comparison of online machine translation tools. *Tcworld*. Retrieved February, 13, 2012, from http://tcworld.info/index.php?id=175