S. Seljan, I. Dunđer

MACHINE TRANSLATION AND AUTOMATIC EVALUATION OF ENGLISH/RUSSIAN-CROATIAN

Abstract. In this research, a specific data set was machine translated by two publicly available machine translation services, Google Translate and Yandex.Translate. Machine translations were performed for two language pairs: English-Croatian and Russian-Croatian. Afterwards, automatic quality evaluation of the machine translated data set was carried out. Several automatic metrics were used: BLEU, NIST, METEOR and GTM, in order to evaluate machine translations relating to the domain of city description, for each language pair and for each machine translation service.

Keywords. *Automatic evaluation, machine translation, English-Croatian, Russian-Croatian, public machine translation service.*

1. Introduction

Automatic evaluation of machine translation is a topic of interest of numerous researches, using various automatic metrics tending to approach as much as possible to human quality assessments. While human evaluation is considered to be «gold standard», it is a subjective, long-term and tiring task. On the other hand, automatic metrics are low-cost, quick, tuneable (if the system performance can be optimised towards), meaningful (giving intuitive interpretation of translation quality), consistent (always giving the same results for repeated usage), correct (better systems ranked higher), reliable, general but also specific to machine translation system properties, as described in [Koehn 2010]. Automatic metrics can be used for comparing the performance of different systems on a common translation task, or for evaluation of different phases during a system development, for machine translation system ranking etc. Automatic metrics use one or more reference translations, sometimes preceded by case information and punctuation removal, tokenisation, by joining numbers (e.g. in phone numbers) and by special treatment of non-ASCII tokens (e.g. Croatian words with diacritics, Russian words with accents). Automatic evaluations are often correlated with human judgements, aiming to approach as much as possible to human quality assessment. Although a number of studies analysed machine translation output for widely spoken languages, not much research was performed for less spoken languages, such as Croatian. Croatian language, belonging to the group of Slavic languages, is a morphologically rich language with relatively free word order, asking for correct cases, number, gender, various types of agreements. In Croatian, each lemma has on average 10 different word forms for nouns, denoting case, number, gender and person which causes lower machine translation quality, especially when translating from less complex languages (e.g. English). This research represents continuation of the work presented in Seljan and Dunđer [2015], where human evaluation was performed for the same data set as in this research. Human evaluation of machine translated sentences was performed using a five point scale for the criteria of fluency and adequacy, developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistics Data Consortium [LDC 2005]. Evaluation was enriched by six categories of error analysis, correlating error type and evaluation criteria.

In this research several automatic metrics were used: BLEU, NIST, METEOR and GTM in order to evaluate two publicly available machine translation services, Google Translate and Yandex.Translate, for two language pairs: one non-closely related language pair (English-Croatian) and one closely related Slavic language pair (Russian-Croatian). For each language pair and each machine translation tool, 100 sentences from the domain of city description were evaluated. At the end conclusions are given regarding correlation with human assessment of the same test set and suggestions for further research proposed.

2. Related work

Callison-Burch et al. [2007] evaluated machine translation output for 8 language pairs, carried out extensive human evaluation and measured intra- and inter-annotator agreement, performed machine translation system ranking and higher-level analysis of the evaluation process, correlated automatic and human judgements as well. In the research 11 automatic metrics were used as a mean to test correlation: METEOR, BLEU, GTM, TER, ParaEval, Dependency overlap, Semantic role overlap, WER and Maximum correlation training on adequacy and fluency. In Callison-Burch [2010] the research on 26 automatic metrics correlating with human assessments for five European languages was presented. Some work has already been made for automatic evaluation of Croatian language. The paper presented by Brkić et al. [2013] reports on machine translation evaluation of 200 sentences for English-Croatian in the legislative domain using automatic metric correlation (BLEU, NIST, F-measure and WER). WER and F-measure, as well as BLEU and NIST showed significant correlation. Although there was no statistically significant correlation between human judgements and automatic metrics, NIST has shown better correlation with the criterion of adequacy. In the paper presented by Seljan et al. [2012] the automatic evaluation metric BLEU was calculated with regard to a single and multiple reference translations, correlating with short and long sentences, analysing the criteria of fluency and adequacy with each error category. The average human score on short sentences was 3,48 and on long sentences 3,00 (5 being best). BLEU score for short sentences was 0,25, for long sentences 0,20, with regard to a single reference set, i.e. 0,32 and 0,26 respectively, with regard to three reference sets. In human evaluation, long sentences have gained on average a 16% lower grade (3,00) than short sentences (3,48), and on average 22% lower BLEU score with regard to one, two, and three reference sets. Correlation between human evaluation and different error types was analysed.

3. Research

Data set consisted of 400 sentences in total: 100 for each language pair (Eng-Cro and Rus-Cro) and for both online machine translation services (Google Translate and Yandex.Translate). English source sentences had in average 20,9 words, while Russian and Croatian sentences were equally long, on average 17,6 words. Machine translated sentences were approximately equally long in all cases, ranging from 17,1-18,1 words. The longest and the shortest English sentence consisted of 36; 7 words, Russian 33; 6 words, Croatian 35; 6 words. The text was taken from the tourist brochures on the city of Zagreb, capital of Croatia. For each machine translation service and for two language pairs, the same set of sentences was evaluated by automatic metrics. All together 400 sentences were analysed: 100 sentences for English-Croatian by Google Translate,100 sentences for Russian-Croatian by Google Translate, 100 sentences for English-Croatian by Yandex.Translate, 100 sentences for Russian-Croatian by Yandex.Translate.

4. Automatic evaluation metrics

Numerous automatic metrics have been used in the evaluation of machine translated text: BLEU, NIST, F-measure (GTM) and METEOR. WER (Word Error Rate), PER (Position-independent Word Error Rate) and TER (Translation Edit Rate) are error measures, while the rest of the metrics fall into the category of accuracy measures. 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.

BLEU (BiLingual Evaluation Understudy) is the most widely used automatic evaluation metric [Doddington 2002, Coughlin 2003], showing that it underestimated the quality of rule-based machine translation systems [Koehn and Monz 2005]. BLEU matches machine translated n-grams with n-grams of its reference translation, and counts the number of matches at the sentence level, typically for 1-4 n-grams [Papineni 2002]. 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 also evaluation with multiple reference translations. Unigram precisions account for adequacy, while n-gram precisions account for fluency. Some of the critiques directed towards BLEU are that it does not take into account the relative relevance of words, overall grammatical coherence, that it is quite unintuitive and generally relies on the whole test set [Koehn 2010].

NIST metric, based on BLEU introduced some modifications [Doddington 2002]. NIST gives information weight for each word, i.e. higher scores to more rare n-grams which are considered as more informative n-grams. NIST differs also from BLEU in brevity penalty calculation, where small differences in translation length do not impact the overall score.

METEOR metric modifies BLEU in the way that it gives more emphasis to recall than to precision [Banerjee and Lavie 2005]. This metric incorporates linguistic knowledge, taking into account the same lemma and synonym matches, which is suitable for languages with rich morphology [Denkowski and Lavie 2011].

GTM metric computes the correct number of unigrams and favours longer matches, is based on precision (number of correct words, divided by generated machine translation system output-length) and recall (number of correct words, divided by reference-length) and calculates the F-measure. 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.

5. Results

Results of automatic evaluation metrics are given in Table 2. Regarding METEOR, all possible matches between two sentences were identified according to the matcher "exact", i.e. words were matched if their surface forms were identical. In other words, synonyms (WordNet), paraphrases and stems were not used. Also, METEOR parameters were not tuned in this research. When comparing language pairs, better results are obtained for closely-related language pair Rus-Cro for both tools. In human annotation presented in Seljan and Dunđer [2015], average grade for Russian-Croatian was 3.1 and for English-Croatian 3.065.

Table 2. Results of automatic machine translation quality evaluation (higher is better)

	BLEU	**NIST**	**METEOR**	**GTM**
GT Eng-Cro	**0.1259**	**3.7606**	**0.1730**	**0.4249**
GT Rus-Cro	0.1536	3.9997	0.1896	0.4510
YT Eng-Cro	0.0946	3.0601	0.1405	0.3646
YT Rus-Cro	**0.2206**	**4.7198**	**0.2349**	**0.5215**

When comparing two MT services in average for both language pairs and across all metrics, Yandex.Translate obtained slightly better results, in average ranging from 0.2-13%. Yandex.Translate obtained better results for Russian-Croatian, ranging from 16-44%. Google Translate scored better for English-Croatian, having better scores ranging from 17-33%.

6. Conclusion

In the research, automatic metrics were used to evaluate all together 400 sentences, i.e. 100 per language pair and per machine translation service. The results showed slightly better scores for closely-related language pair Russian-Croatian for both tools. When comparing machine translation services in average, Yandex.Translate obtained slightly better results, ranging from 0.2-13%. Google Translate obtained better results for English-Croatian language pair and Yandex.Translate for Russian-Croatian, both on all automatic metrics (BLEU, NIST, METEOR and GTM). The results correlate with average human evaluations. However, automatic metrics in this research did not take into account the lemmas, which play an important part when translating into morphologically rich languages, assigning therefore complete errors for words having different word forms. Results are also lower due to one reference set and relatively long sentences, non-nominative cases and some specific terminology. Further research would include more automatic metrics, larger test sets and analysis of correlation between human and automatic evaluation.

References

1. *Banerjee S., Lavie, A.* (2005), 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. Ann Arbor, 2005, pp. 65–72.

2. *Brkić, M., Seljan, S., Vičić, T.* (2013), Automatic and Human Evaluation on English-Croatian Legislative Test Set, Lecture Notes in Computer Science – LNCS, Vol. 7817, pp. 311–317.

3. *Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.* (2007), (Meta-) Evaluation of Machine Translation. Proceedings of 2nd Workshop on Statistical Machine Translation. Prague, 2007, pp. 136–158.

4. *Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., Zaidan, O.* (2010), Findings of Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. Proceedings of the Joint 5th Workshop on SMT and Metrics MATR (ACL). Uppsala, 2010, pp. 17–53.

5. *Coughlin, D.* (2003), Correlating automated and human assessments of machine translation quality. Proceedings of MT Summit IX. New Orleans, 2003, pp. 23–27.

6. *Denkowski, M., Lavie, A.* (2011), Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. Proceedings of the Sixth Workshop on Statistical Machine Translation. Edinburgh, 2011, pp. 85–91.

7. *Doddington, G.* (2002), Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the Second international conference on Human Language Technology Research. San Francisco, 2002, pp. 138–145.

8. *Koehn, P.* (2010), Statistical Machine Translation, Cambridge University Press, Cambridge.

9. *Koehn, P., Monz, C.* (2005), Shared task: Statistical machine translation between European languages. Proceedings of ACL 2005 Workshop on Parallel Text Translation. Arbor, A. 2005, pp. 119–124.

10. *LDC.* (2005), Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

11. *Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.* (2002), Bleu: A method for automatic evaluation of machine translation. Proceedings of ACL 2002. Philadelphia, 2002, pp. 311–318.

12. *Seljan, S., Dunđer, I.* (2015), Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. Advances in Intelligent Systems and Computing, Springer, 2015, pp. 1089-1098.

13. *Seljan, S., Vičić, T., Brkić, M.* (2012), BLEU Evaluation of Machine-Translated English-Croatian Legislation. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, 2012, pp. 2143–2148.

Sanja Seljan, Ivan Dunđer
University of Zagreb (Croatia).
Faculty of Humanities and Social Sciences.
E-mail: sanja.seljan@ffzg.hr, ivandunder@gmail.com