PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 95 APRIL 2011 77–86

Quiz-Based Evaluation of Machine Translation

Jan Berka[a b], Martin Černý[a], Ondřej Bojar[b]

[a] Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering
[b] Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

This paper proposes a new method of manual evaluation for statistical machine translation, the so-called quiz-based evaluation, estimating whether people are able to extract information from machine-translated texts reliably. We apply the method to two commercial and two experimental MT systems that participated in WMT 2010 in English-to-Czech translation. We report inter-annotator agreement for the evaluation as well as the outcomes of the individual systems. The quiz-based evaluation suggests rather different ranking of the systems compared to the WMT 2010 manual and automatic metrics. We also see that overall, MT quality is becoming acceptable for obtaining information from the text: about 80% of questions can be answered correctly given only machine-translated text.

1. Introduction

There are many ways for evaluating the quality of machine translation, from automatic metrics like BLEU (Papineni et al., 2002) or METEOR (Lavie and Denkowski, 2009), to different kinds of human judgement (manual evaluation) (Callison-Burch et al., 2010).

These methods are based on the question "Is this a plausible translation of the text?" We propose a different manual evaluation method, which asks a slightly different question: "Does the translated text provide all the information of the original?" This follows the idea, that in many real-life situations like reading the news or getting travel directions we do not need to have a totally correct translation—we just need to now what happened or where to go.

Our proposed quiz-based evaluation method is centered around yes/no questions. We start by collecting naturally occurring text snippets in English, manually equip

Corresponding author: bojar@ufal.mff.cuni.cz
Cite as: Jan Berka, Martin Černý, Ondřej Bojar. Quiz-Based Evaluation of Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 95, 2011, pp. 77–86. doi: 10.2478/v10108-011-0006-1.

them with a set of yes/no questions (in Czech) and translate them using four MT systems to Czech. The translated texts are then handed to annotators, who see only one of the machine translations and answer the questions. We measure the quality of translation by the number of correctly answered questions.

2. Preparation of Texts and Questions

For the experiment, we collected English texts from various sources, written hopefully by native speakers.[1] These texts covered four topic domains:

- Directions description – these texts provided information of a location of a certain place, or described a route to somewhere,
- News – this topic contained snippets from newspaper articles about politics and economy,
- Meeting – texts from this domain contained information about places, times and subjects of different meetings,
- Quizes – short quiz-like questions in the fields of mathematics, physics or software engineering.

These topics cover a large variety of common texts, from which the reader needs usually only the core information. The grammatical correctness of the MT output is not important as long as the meaning is not disrupted.

The collected texts had also three different lengths from one sentence texts to texts with two and three sentences. This distribution of texts allowed us to examine, whether some topics are harder to translate and if success of the translation (from the point of view of quiz-based evaluation) depends on text length.

We managed to collect a total of 132 texts with close to uniform distribution of topic domains and lengths.

In the next step, we created three questions with answers yes or no for each text. This meant the total of 396 different questions for evaluation of the machine translation systems. Figure 1 shows four single-sentence sample texts and the corresponding questions (in Czech with an English gloss).

After the texts were collected and questions were prepared, we did a final preannotation check of the "golden" answers (answers deemed correct by authors of the questions). In this process, 78 answers were changed, 12 of them with no change of the actual value and changing only the uncertainty indicator (in situations when it was natural and right for an annotator to be unsure). 8 questions were completely removed. We ended up with 376 questions with the following distribution of golden answers: 191 yes, 170 no, 15 can't tell.

Texts were then translated by four different machine translation systems (see Section 3). Each annotator was given a set of 132 texts with the corresponding ques-

[1] We always chose web sited in countries where English is the official and majority language. In the current globalized world, the mother tongue of the author can be different.

Topics	Texts and questions
Directions	Follow the red arrows to the registration desk.
	Jsou šipky zelené? Are the arrows green? Ukáže cestu asistent? Will an assistent show you the way to registration desk? Does the registration take place right by the entrance? Probíhá registrace hned u vchodu?
News	The Chinese government is considering legislation that would make eating cats and dogs illegal.
	Je v Číně zakázáno jíst psy? Is dog eating banned in China? Uvažuje čínská vláda o zákazu pojídání psů a koček? Is government considering a ban of dog and cat eating? Jí v Číně psi často kočky? Do dogs in China often eat cats?
Meetings	The University of York Filmmaking Society meets every Monday at 6.30pm at L/047.
	Existuje na univerzitě v Yorku spolek filmařů? Does a filmmaking society exist on University of York? Je v Yorku zřejmě filmová univerzita? Is a film university in York? Konají se schůzky každé pondělí? Do the meetings take place every monday?
Quiz	A equals two thirds and B equals free fifths.
	Je A větší než B? Is A greater than B? Jsou A a B stejně velké? Does A equal B? Je B menší než 1? Is B less than 1?

Figure 1. Examples of one-sentence texts and their corresponding questions

tions. We tried to get annotations of all topics, lengths and MT systems uniformly distributed, but not every annotator completed the task. In total we obtained a set of 1891 annotated texts, with the distribution of topics and lengths as shown in Table 1 and MT systems as shown in Table 2.

The use of yes/no questions slightly affected the possibilities of questioning, but allowed us to process the answers automatically. The annotators were given 6 possible answers to choose from:

- yes, denoted by annotation mark 'y',
- probably yes (marked as 'Y'),
- no ('n'),
- probably no ('N'),
- can't tell (based on the text), marked as 'x',
- don't understand the question ('X').

Except for 'X', the capital letter was used to indicate that the annotator was not sure.

	1 sentence	2 sentences	3 sentences	All lengths
Directions	10.4%	8.2%	8.5%	27.1%
Meetings	7.0%	6.1%	7.1%	20.1%
News	10.3%	10.0%	8.8%	29.1%
Quizes	8.5%	9.6%	5.6%	23.7%
All topics	36.2%	33.9%	30.0%	

Table 1. Topic domains and lengths distribution in annotated texts

	Sentences	Google	CU-Bojar	PCTrans	Tectomt
Directions	1	23.2%	25.8%	29.0%	22.1%
	2	25.0%	23.7%	27.0%	24.3%
	3	29.3%	18.5%	23.6%	28.7%
Meetings	1	24.5%	32.0%	21.1%	22.5%
	2	23.0%	24.8%	23.9%	28.3%
	3	26.0%	22.7%	30.7%	20.7%
News	1	23.6%	25.7%	26.7%	24.1%
	2	24.2%	30.8%	23.6%	21.4%
	3	23.0%	26.7%	25.5%	24.9%
Quizes	1	25.3%	18.5%	26.5%	29.6%
	2	27.4%	21.2%	20.7%	30.7%
	3	24.3%	27.2%	22.3%	26.2%

Table 2. MT systems distribution in annotated texts (with respect to topic domains and text lengths)

3. Brief Overview of Examined Systems

In this paper, we consider 4 systems from WMT10. It is a small subset of all the systems present, but they represent a wide range of technologies.

Google Translate is a commercial statistical MT system trained on unspecified amount of parallel and monolingual texts.

PC Translator is a Czech commercial system developed primarily for English-to-Czech translation.

TectoMT is a system following the analysis-transfer-synthesis scenario with the transfer implemented at a deep syntactic layer, based on the theory of Functional Generative Description (Sgall et al., 1986) as implemented in the Prague Dependency Treebank (Hajič et al., 2006). For TectoMT, the tectogrammatical layer was further simplified (Žabokrtský et al., 2008). We use the WMT10 version of TectoMT (Žabokrtský et al., 2010).

CU-Bojar is an experimental phrase-based system based on Moses[2] (Koehn et al., 2007), tuned for English-to-Czech translation (Bojar and Kos, 2010).

4. Results

4.1. Intra-annotator Agreement

In order to estimate intra-annotator agreement, some texts and the corresponding questions in the set of 132 texts given to each annotator were duplicated. The annotators were volunteers with no benefit from consistent results, so we didn't worry they would search their previous answers to answer repeated questions identically. In fact, they even didn't know that they have identical texts in their set.

However, the voluntary character of the annotation has also caused troubles, because we got only very few data for the intra-annotator agreement. Only 4 annotators answered questions about two identical texts, with the average intra-annotator agreement of 92%.

About two months after the annotation, one of the annotators answered once again all the questions from his set of texts, providing a dataset of 393 answered questions. From the comparison of his new and old answers we estimate the intra-annotator agreement as 78.9%.

4.2. Inter-annotator Agreement

In order to estimate the inter-annotator agreement, each translated text with corresponding questions was present in several sets given to independent annotators. The inter-annotator agreement between two annotators x, y was then computed as:

$$\mathrm{IAA}(x,y) = \frac{\text{number of identically answered questions}}{\text{number of common questions}} \tag{1}$$

The overall inter-annotator agreement as the average of $\mathrm{IAA}(x,y)$:

$$\mathrm{IAA} = \frac{\sum_x \sum_{y \neq x} \mathrm{IAA}(x,y)}{2 \cdot \text{number of all couples of different annotators}} \tag{2}$$

From the results we estimate the overall inter-annotator agreement as 66% taking uncertainty into account and 74.2% without it (i.e. accepting e.g. 'y' and 'Y' as the same answer).

[2] http://www.statmt.org/moses

4.3. Success Rates

This section provides the overall results of the four examined MT systems. It also shows, how the success rate depends on and varies with topic domains and text lengths.

First, let us discuss the possibilities of what should be considered a correct answer. The main question is, whether to accept answers 'Y' and 'N' as correct, when the golden answers are 'y' and 'n', or in other words: do we accept an unsure but otherwise correct answer? We decided to accept these answers as correct, as they meant that the reader of the translated text indeed got the information, only not so explicit as it was in the original text.

Another question is how to handle answers 'x' ("can't tell from the text") and 'X' ("don't understand the question"). We took 'x' as an ordinary answer, counting as correct only when the golden answer was also 'x'. Answers 'X' were not taken into account, because they indicated a problem of understanding the question, not the text.

We evaluated the dataset using all the interpretation possibilities and observed differences only in the absolute values but never in overall trends (e.g. the winning MT system). Therefore we present only the judgment strategy described above.

The dataset for evaluation of the four examined MT systems consists of 5588 answers to questions about 1905 text instances as provided by the total of 18 different annotators. 61 answers were not included in final statistics because they were 'X'.

The success rates are computed as follows:

$$\text{Success rate} = \frac{\text{Number of correct answers}}{\text{Number of all answers}} \cdot 100\,\% \tag{3}$$

The overall success rate was 79.5%.

Table 3 shows the success rates for individual MT systems with respect to topic domain and number of sentences in translated texts. Each cell in the table (except the "Overall" row) is based on 115.1 answers on average (standard deviation 26.4, minimum 69, maximum 170 answers).

Tables 4 and 5 show the overall success rates of all examined MT systems with respect to text length and then topic domain.

4.4. Discussion

The results document that the overall success rate is slightly higher than our estimate of intra-annotator and inter-annotator agreement. We have thus probably reached the limits of this type of evaluation. The main good news is that overall, our MT systems allowed to answer nearly 80% of questions correctly. In many practical situations, this success rate can be sufficient. For getting or meeting somewhere, the users should be more cautious as the success rate dropped to 76.59%.

Topic	Text length	Google	CU-Bojar	PC Translator	TectoMT
Directions	1	**81.1%**	72.5%	80.8%	78.4%
	2	77.9%	75.9%	76.4%	**79.3%**
	3	83.3%	68.6%	**85.0%**	79.0%
Meetings	1	**80.2%**	68.4%	64.2%	78.5%
	2	**83.3%**	73.8%	73.8%	75.0%
	3	77.0%	79.5%	**84.7%**	79.5%
News	1	**91.1%**	81.1%	87.8%	89.7%
	2	78.2%	**82.9**%	81.8%	76.7%
	3	75.7%	75.4%	69.7%	**81.1%**
Quizes	1	75.2%	69.9%	82.5%	**84.1%**
	2	78.6%	80.5%	84.4%	**89.1%**
	3	81.1%	76.2%	79.7%	**81.3%**
Overall		80.3%	75.9%	80.0%	**81.5%**

Table 3. Success rates for examined MT systems. Best in each row in bold.

Text length	Success rate
1 sentence	79.93%
2 sentences	79.74%
3 sentences	78.64%

Table 4. Overall success rates for different text lengths

As we see from Tables 4 and 5, the success rates drop only slightly with increasing length of translated texts. The rates of different topic domains are also very close, with the news topic being the most successful. This could be caused by the annotators already knowing some of the information from local media or by the fact that most of the systems are designed to handle "generic text" and compete in shared translation tasks like WMT which are set in the news domain.

Table 6 compares our ranking of systems to various metrics used in the WMT10 evaluation campaign (Callison-Burch et al., 2010). The figures indicate that various manual evaluations provide rather different results. Users of MT systems should therefore evaluate system candidates specifically for the translation task where the systems will eventually serve.

In terms of allowing to correctly answer questions in our examined four domains, TectoMT seems to be the best. It is therefore somewhat surprising that TectoMT was the worst in terms of "Edit deemed acceptable", i.e. the percentage of post-edits of the output carried out without seeing the source or reference that an independent

Topic	Success rate
Directions	78.44%
Meetings	76.59%
News	81.33%
Quizes	80.87%

Table 5. Overall success rates for different topic domains

Metric	Google	CU-Bojar	PC Translator	TectoMT
$\geq$ others (WMT10 official)	**70.4**	65.6	62.1	60.1
$>$ others	49.1	45.0	**49.4**	44.1
Edits deemed acceptable [%]	**55**	40	43	34
Quiz-based evaluation [%]	80.3	75.9	80.0	**81.5**
BLEU	**0.16**	0.15	0.10	0.12
NIST	**5.46**	5.30	4.44	5.10

Table 6. Manual and automatic scores of the MT systems. Best in bold. We report WMT manual evaluations (comparison with other systems and acceptability of post-editing) and the overall result of our quiz-based evaluation.

annotator then validated as to preserve the original input. The discrepancy can have several reasons, e.g. TectoMT performing better on a wider range of text domains than the news domain of WMT10, or our quiz-based evaluation asking about some "core" information from the sentences whereas the acceptability of edits requires all details to be preserved.

Overall, the most fluent output is produced by Google (with respect to the WMT official score based on the percentage of sentences where the system was manually ranked equal or better than other systems as well as with respect to the acceptability of edits). Google ends up being the second in our quiz-based evaluation. PC Translator was often a winner alone, clearly distinct from others, because it scored best in "$>$ others".

The most surprising is the result of CU-Bojar: while second in the official "$\geq$ others", it scores much worse in all other comparisons. CU-Bojar is probably often incomparably similar to the best system but if observed alone, it does not preserve valuable information as good as other systems.

5. Conclusion

In this paper we described a novel technique for manual evaluation of machine translation quality. The presented method, called quiz-based evaluation, is based on

annotators' reading of machine-translated texts and answering questions on information available in original texts. The presented method was used for evaluating four English-to-Czech MT systems participating in WMT10 (Callison-Burch et al., 2010) on short texts in four different topic domains.

The results indicate a completely different order of the evaluated systems compared to both automatic and manual evaluation methods as used in WMT10. The results also suggest that the success rate of machine translation mildly decreases with increasing text length, although our texts were too short (one to three sentences) for a reliable observation. The success rates of various topic domains were also very close, with translations of news being the most successful.

The overall success rate was 79.5%, meaning that on average, machine translation allowed our annotators to answer four of five questions correctly. This suggests a fairly high practical usability of modern machine translation systems.

Acknowledgement

The work on this project was supported by the grants P406/10/P259, P406/11/1499, and the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003 of the Czech Republic).

We are grateful to all our student collaborators who provided us with the texts, questions as well as the evaluated annotations.

Bibliography

Bojar, Ondřej and Kamil Kos. 2010 Failures in English-Czech Phrase-Based MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 60–66, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-1705`.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 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*, WMT '10, pages 17–53, Morristown, NJ, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8. URL `http://portal.acm.org/citation.cfm?id=1868850.1868853`.

Hajič, Jan, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková Razímová. Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4, 2006.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Morristown, NJ, USA, 2007. Association

for Computational Linguistics. URL http://portal.acm.org/citation.cfm?id=1557769.1557821.

Lavie, A. and M.J. Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115, 2009. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-77954763029&partnerID=40&md5=38249c2daa847f4657c08f5f051a1b6e.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135. URL http://dx.doi.org/10.3115/1073083.1073135.

Sgall, P., F. Hajičová, and J. Panevová. *The Meaning of Sentence and Its Semantic and Pragmatic Aspects*. Academia, Prague, Czechoslovakia, 1986. ISBN 90-277-1838-5.

Žabokrtský, Zdeněk, Jan Ptáček, and Petr Pajas. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 167–170, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-09-1. URL http://portal.acm.org/citation.cfm?id=1626394.1626419.

Žabokrtský, Zdeněk, Martin Popel, and David Mareček. Maximum entropy translation model in dependency-based mt framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 207–212, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W10-1731.

Address for correspondence:
Ondřej Bojar
bojar@ufal.mff.cuni.cz
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
11800 Praha, Czech Republic