Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 547–551
ISBN 978-83-60810-27-9
ISSN 1896-7094

Matura Evaluation Experiment Based on Human Evaluation of Machine Translation

Aleksandra Wojak and Filip Graliński
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland
Email: aleksandra.wojak@wp.eu, filipg@amu.edu.pl

***Abstract*—A Web-based system for human evaluation of machine translation is presented in this paper. The system is based on comprehension tests similar to the ones used in Polish *matura* (secondary school-leaving) examinations. The results of preliminary experiments for Polish-English and English-Polish machine translation evaluation are presented and discussed.**

I. Introduction

THE SUCCESS of Statistical Machine Translation, well illustrated by the popularity of Google translation tools, has a positive impact on the development of the whole Machine Translation discipline. This phenomenon brings about the need for comparing the quality of MT tools and systems that keep appearing, both in the academic field and on the commercial market.

In [1] Papineni et al. introduced the BLEU metrics that counts well-translated n-grams (sequences of n words). Improvements of the metrics have been proposed by Doddington [2] (NIST) and Lavie and Agarwal [3] — METEOR. The common feature of those approaches is that evaluation is executed fully automatically, by comparing the text translated by an MT system to the reference translation, prepared by a human.

One of the advantages of automatic evaluation is that it can be used for training. For example, Tenerowicz [4] uses the METEOR metrics in a genetic algorithm that trains the probabilistic grammar used in a parser of the Translatica MT system.

The drawback is its weak correlation with human evaluation. Turian et al. [5] claim that the most popular MT evaluation metrics, BLEU and NIST, fail to correlate well with human judgements of translation quality. In the experiment of Tenerowicz [4], a significant number of translations, improved by the METEOR measure, were estimated as worse translations by linguists.

One of the reasons behind it is the following common feature of automatic evaluation tools: they assign points for parts of sentences, even if the whole sentences are not comprehensible. Points are not assigned for translation adequacy if it is not mirrored by appropriate word strings (although METEOR tries to overcome this drawback by scoring synonyms).

On the other hand, in human simple evaluation (ranking quality of translations or choosing the best one from a given set, when a source sentence is known) evaluators' knowledge of sentence meanings affects the measurement results.

We propose an idea of human evaluation with evaluators being not aware of the source sentence. Evaluators are supposed to give answers to a prepared set of questions, knowing only the target text translated automatically. The evaluation resembles the comprehension test for the Polish *matura* (i.e. secondary school-leaving examination) foreign language exam.

The *Matura* Evaluation obviously measures the comprehensibility of the translated text as well as its adequacy. The latter is achieved by preparing the test questions based solely on the source test.

Please note that our approach does not require evaluators to be native speakers or experts on the target language.

The idea to use comprehension tests for machine translation evaluation is not new [6] [7]. What is new is the use of Web-based application for such purposes.

II. Experiment summary

Matura Evaluation is an experiment for human-based evaluation of machine translation. Its main idea was to compare intelligibility and adequacy of different translations of the same source text, with the correctness of answers being the measurement criteria.

The experiment was performed using two directions of translation between Polish and English. Several source texts were translated by each translation system under test. Source texts came with about 10 questions. Each experiment participant was presented with a random translation of a random text along with a relevant set of questions. The participants were expected to answer these questions using the information provided in the translated text.

It is assumed that if the source text is translated correctly by an MT system, i.e. all the information from the source text can be found also in the translation in an intelligible form, then the participant should easily find the correct answer. On the other hand, if the sentence meaning is changed in

the translation, the experiment participant obtains the wrong information and, hence, chooses the wrong answer. It is also likely that the relevant information was translated correctly, yet it cannot be inferred from the whole sentence/paragraph where the answer is to be found; or that the translation is difficult to understand. In such a situation the participant is supposed to mark "Translation impossible to understand" as an answer.

III. Texts and Translations Used in the Experiment

Translations were done in two directions between English and Polish. Nine source texts were used in the experiment: five in Polish and four in English. Each of them was translated by each of the three MT systems tested: Google Translate (*http://translate.google.com/*), Kompas (http://www.kompas.info.pl/) and Translatica (http://www.translatica.pl/).

Texts used in the experiment differed in topic and level of difficulty: some of them were supposed to be more specialised (e.g. summary of the *System of Education Act*), other more general (e.g. an article from Wikipedia on *Alice's Adventures in Wonderland*), yet another were parts of literary works (*Little Prince* by A. de Saint-Exupery and *The Deluge* by H. Sienkiewicz). The aim of this diversity was to compare the results for translations of different types of texts. It is well known that it is much more difficult for an MT system (and for human translator as well) to translate literary works than other types of texts, because they contain a large number of metaphors, which cannot be translated literally.

We decided to use real texts, not artificially crafted for the purposes of machine translation evaluation. The following texts were used in the experiment:

1) Polish source texts:
   - article from Wikipedia about the book *Alice's Adventures in Wonderland* (1581 words)
   - part of the first chapter of *The Deluge* by H. Sienkiewicz (2128 words)
   - an article *Aesthetics of the Pythagoreans* (1575 words)
   - summary of the *System of Education Act* (1659 words)
   - an article *Vanishing Venice* (2748 words)
2) English source texts:
   - English translation of the first chapter of *Little Prince* by A. de Saint-Exupery (1753 words)
   - an article about *Greater Poland Uprising* (826 words)
   - an article about the history of St. Patrick's Day celebrations (767 words)
   - an article by Paul Graham *What You Wish, You'd Known* (5083 words)

IV. Questions

Questions were based on the source texts, but written in the target language of the translations. About ten questions based on the source text were prepared in the target language. There were three/four variant answers prepared for each question but only one of them was correct.

Questions were supposed to check if some precise information from the source text had been preserved during translation. Therefore a very specific information was usually expected as an answer to each question.

Various types of questions were prepared for the experiment. Some of them were supposed to check if a word with multiple meanings was translated correctly.

For example, in the text about *St. Patrick's Day* there was a question:
Why did the experiment fail in Savannah?
which the answer to could be found in the following paragraph:
"[...] in 1961, Savannah mayor Tom Woolley had plans for a green river. Due to rough waters on March 17, the experiment failed[...]".
The relevant answer was the correct translation of the word *rough*. There were three answer variants:

- *surowy*
- *szorstki*
- *wzburzony*

All of the answers are different (and, in general, correct) translations of the word *rough* into Polish. However, only the third translation fits the context. It turned out that only one MT system tested (Translatica) translated this word using the correct meaning.

Other questions checked if the meaning of the sentence, possibly with more than one negation word, was not changed (some MT systems have problems with complex negative sentences). It sometimes happens that two negation words, related to two different words in the source text, appear one after another in the translation, thus changing the meaning of the sentence or making it impossible to understand. A sentence from the *System of Education Act* is an example of negation-related problems in translation. The sentence started with the clause *If the child didn't go to nursery school*, which was translated correctly by Kompas and Translatica. However, Google Translate did not manage to translate this sentence correctly. In its translation, the output sentence started with *If the child went to kindergarten*, which totally changed the meaning of the sentence.

Another type of questions was supposed to check the adequacy of translation of compound sentences, especially relative clauses – if the logical relation between parts of the sentences remains the same after translating the source text. However, these relations were usually preserved in translations.

Preparing the questions for such an experiment is quite a challenge, because they should check various aspects of translations. Moreover, it is impossible to check if every sentence is translated correctly. Due to the time limit imposed on the participants, there had to be a limited number of questions to each text. In this experiment we decided that ten questions for each text would be enough to check the general understanding and some chosen specific information.

V. Participants

The *Matura* Evaluation experiment was carried out through the Internet. It was prepared in the form of web application created in Silverlight, so persons taking part in the experiment could access it through the website. Participants were provided with random translations of randomly selected source texts and the corresponding questions. They were supposed to select answers based on the information from the given translation.

The majority of participants taking part in the experiment were students (mainly from the Faculty of Mathematics and Computer Science, but not only). All of them were educated enough to be able to find the correct answer in the text if it was translated clearly and correctly enough. Of course every person has different reading comprehension skills and different deduction abilities, so this experiment should be conducted on a large number of participants for credible and meaningful results. Sometimes it also could happen that the participant knew the answer to the question even without reading the text. It was due to the fact that texts used here were not written for the purpose of this experiment. On the contrary, they consisted of well known fragments of literature works (*The Deluge*, *Little Prince*), articles describing problems which could be known to the participants (Greater Poland Uprising, aesthetic of Pythagoreans etc.). The aim of this experiment was to check the quality of translations, not the knowledge of people taking part in it. Therefore participants were asked to choose answers according to the given text, not their previous knowledge or guesses.

All the participants were Polish native speakers. As the experiment tested the translations between Polish and English in both directions, every participant was supposed to define their English skills prior to its beginning. Texts in English were given only to participants who described their English skills as *good* or *medium*. All the other participants were provided with texts in Polish. Of course the ideal situation would be to give English translations to English native speakers to be sure that if they choose the wrong answer, it is because the text is translated wrongly, and not that the participant's reading comprehension skills in English are too poor. However, we wanted to test how an automatically translated text is perceived by source language native speakers (commercial Polish-English MT systems are usually reviewed in the press or on the Internet by Polish native speakers rather than English native speakers). Therefore, the following solution was used: an additional answer variant was added to each question - *My English skills are not good enough to provide the correct answer to this question*. It was done in order to prevent a participant from guessing the correct answer or choosing the option *translation impossible to understand*, while the translation could be actually quite good but in English too advanced for the participant to understand.

VI. Results of the Experiment

The experiment results are presented in Table I. In the top row of the table the average results (i.e. the percentage of correct answers) obtained by each of the tested MT systems are displayed. As we can observe, all the systems received quite high rates: from 65.45% up to 74.81%. From these figures we can deduce that the translations produced were generally quite understandable, because in average every participant was able to answer correctly about six – seven questions out of ten. This result is quite optimistic, because it implies that an average translation was in about 70% understandable and adequate in reference to its source text.

All the average results presented in Table I are counted using weighted arithmetic mean. Weights depend on a number of times a specific translation was used in the experiment (as mentioned before, translations and texts were chosen randomly for each experiment). Table II indicates how many times each translation of each text was used in the experiment.

When we compare the results obtained by each of the MT systems in both directions of the translations tested in this experiment, we can notice quite a difference. Generally translations from Polish into English received higher marks than translations in the opposite direction. This is even more interesting if we keep in mind that all the experiment participants were Polish native speakers and, hence, able to better understand texts written in their mother tongue, Polish, even after translation. However, the assumption turned out to be false. Polish translations were not only more difficult to understand then the English translations, but texts translated into Polish more often contained wrong information. This could be because English is more difficult to parse than Polish and Polish is more difficult to synthesise than English (because of complex morphosyntactic agreements) and therefore it is more difficult for an MT system to generate a correct and understandable sentence in Polish than in English.

VII. Result Analysis

The interesting fact is that for some texts translation results differ significantly between MT systems. The most essential difference can be observed between the translations into Polish, e.g. Polish translation of *Greater Poland Uprising* translated by Google Translate obtained 73.33% (the best score for translation of this text), while the same text translated by Translatica received an average mark of 33.33%. Quite the opposite results were obtained by these MT systems as far as translation of the article about *St. Patrick's Day* is concerned: Google Translate received the lowest mark for this translation: 37.50%, while Translatica 84.00%. These both results are quite objective, because the translations mentioned above were used in almost the same number of experiments: translation of *Greater Poland Uprising* by Google Translate: 5 times, by Translatica: 6; translation of *St. Patrick's Day* by Google Translate: 4 times, the same amount by Translatica. Translation from Polish into English did not differ so significantly. The largest differences in marks occurred in translation of the most specialised text – *System of Education Act*. Again Translatica translated it in the best way, obtaining 89.47% score, while Google Translate only 66.67%. However, these results cannot be compared in a very credible way, because translation by

Translatica was used 6 times in the experiment, while the translation by Google Translate only 3.

If we want to go deeper in our analysis, we can compare the number of wrong answers which were given to each question after reading translations generated by each MT system. There were two types of answers considered as "wrong": an answer which was not the correct one and an answer saying that *translation is impossible to understand*. The number of such answers was counted for each text and for each translation separately. The most interesting were situations in which one question was answered almost always correctly when using one translation, and the wrong answer was provided based on another translation. Usually it implied that the translation of the paragraph/sentence with information needed to answer the question was much worse in the second case.

An example question with sentences from different translations to illustrate such a situation comes from the article about *Alice's Adventures in Wonderland*. Question no. 8 was not answered correctly by anyone using translation by Google Translate, and it was answered correctly by 3 out of 4 participants using the translation created by Kompas:

Question 8: Who was the author of the first [Polish] translation closest in meaning to the original text?

- *Chuck Connors (1965) – The first translation in line with the original (Google Translate)*
- *Maciej Słomczyński (1965) – the first translation corresponding to the original (Kompas)*

Correct answer to this question is "Maciej Słomczyński". Of course no one reading translation by Google Translate would be able to answer this question correctly because of changed name (*Maciej Słomczyński* was translated into *Chuck Connors*).

Another interesting example comes from the second chapter of *Little Prince* and a question *Over how many parts of the world did the author fly?* All the participants reading this text translated by Google Translate (5 persons) gave the wrong answer, and all the participants using translation made by Translatica (6 persons) answered correctly. The assumption that something was wrong with the translation of Google turned out to be correct. The original sentence *I have flown a little over all parts of the world* was translated by Google into *Mam lotu mało w stosunku do wszystkich części świata*, what gives the wrong understanding that the author has flown *not much*.

VIII. Conclusions

The experiment called *Matura Evaluation* turns out to be a good method of human-based evaluation of machine translation. Results obtained in this experiment show the correspondence with the quality of translations. However, such an experiment has to fulfil some requirements for its results to be credible. First of all, a large number of participants must take part in the experiment. Moreover, all tested translations should be used by similar number of participants. There are also some requirements regarding the experiment preparation: the texts should be correctly selected, different in style and difficulty to enable the comparison of translations of different types of texts.The questions should be clear, prepared based only on source texts for the results of experiment to be objective. All answers should be easy to find in the source text (and, in consequence, in the correct translations), because this experiment does not check the participants' reading comprehension skills, but the quality of translation.

Acknowledgment

The paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No 003/R/T00/2008/05).

References

[1] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[2] G. Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics," in *Proceedings of the Human Language Technology (Notebook)*, San Diego, CA, 2002, pp. 128–132.

[3] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 228–231. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0734

[4] Z. Tenerowicz, "Zastosowanie obliczeń ewolucyjnych w przetwarzaniu j ezyka naturalnego (Using evolutionary computation in Natural Lanuage Processing)," Master's thesis, Adam Mickiewicz University, Poznań, 2010.

[5] J. Turian, L. Shen, and I. D. Melamed, "Evaluation of machine translation and its evaluation," in *In Proceedings of MT Summit IX*, 2003, pp. 386–393.

[6] M. Tomita, M. Shirai, J. Tsutsumi, M. Matsumura, and Y. Yoshikawa, "Evaluation of MT Systems by TOEFL," in *Proceedings of the Theoretical and Methodological Implications of Machine Translation*, 1993.

[7] M. Fuji, "Evaluation Experiment for Reading Comprehension of Machine Translation Outputs," in *Proceedings of Machine Translation Summit VII*, 1999, pp. 285–289.

TABLE I
EXPERIMENT RESULTS

	Google Tr.	Kompas	Translatica	Weighted avg
Weighted average result	**65.34%**	**71.98%**	**74.16%**	**70.96%**
PL → EN translation	78.17%	80.60%	89.25%	83.64%
EN → PL translation	58.22%	62.21%	59.83%	59.96%
Polish → English translations				
Alice in Wonderland	80.56%	70.83%	88.89%	81.41%
Vanishing Venice	80%	90%	93.33%	90.00%
Pythagorean aesthetics	100%	94.44%	90.7%	92.80%
System of Education Act ...	66.67%	80.43%	89.47%	78.11%
The Deluge – chapter I	80%	78.95%	83.33%	80.64%
English → Polish translations				
Little Prince	56.25%	64.41%	51.67%	57.51%
St. Patrick's Day	37.50%	65%	84%	63.64%
What You Wish, ...	62.50%	62.50%	75%	66.67%
Greater Poland Uprising	73.33%	55.56%	33.33%	53.84%

TABLE II
NUMBER OF EXPERIMENTS PERFORMED

	Google Translate	Kompas	Translatica
Alice in Wonderland	3	4	6
Vanishing Venice	1	2	3
Aesthetics of the Pythagoreans	1	2	5
System of Education Act	3	5	2
The Deluge – chapter I	2	4	3
Total Polish → English	10	17	19
Little Prince	5	6	6
St. Patrick's Day	4	2	5
What You Wish, You'd Known	4	4	4
Greater Poland Uprising	5	3	5
Total English → Polish	18	15	20
Total	28	32	39