PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 115 OCTOBER 2020 129–142

Extending Ptakopět for Machine Translation User Interaction Experiments

Vilém Zouhar, Michal Novák

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

The problems of outbound translation, machine translation user confidence and user interaction are not yet fully explored. The goal of the online modular system Ptakopět is to provide tools for studying these phenomena. Ptakopět is a proof-of-concept system for examining user interaction with enhanced machine translation. It can be used either for actual translation or running experiments on human annotators. In this article, we aim to describe its main components and to show how to use Ptakopět for further research. We also share tips for running experiments and setting up a similar online annotation environment.

Ptakopět was already used for outbound machine translation experiments, and we cover the results of the latest experiment in a demonstration to show the research potential of this tool. We show quantitatively that even though backward translation improves machine-translation user experience, it mainly increases users' confidence and not the translation quality.

1. Introduction

Internet users often find themselves in need to produce a text in a foreign language they do not speak perfectly. This poses a problem as the users are not able to validate the machine translation result. Our goal is to explore this user–computer interaction and to demonstrate what tools may help the users in these scenarios, increasing their confidence in the produced translations.

For this purpose, we describe Ptakopět, a system which can help with outbound translation. Moreover, it can be extended by other tools and offers an environment for examining usage strategies of outbound translation.

Corresponding author: zouhar@ufal.mff.cuni.cz

Cite as: Vilém Zouhar, Michal Novák. Extending Ptakopět for Machine Translation User Interaction Experiments. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 129–142. doi: 10.14712/00326585.008.

1.1. Machine Translation Usage

Machine translation usage can be for some purposes broadly divided into inbound and outbound translation. In inbound translation, mostly gisting, we are the recipients of a message in a foreign–language and it is our responsibility to understand it correctly. It is typically reading websites and e-mails in a foreign language. This use case is characterized by lower quality requirements and the translation not being distributed further.

In outbound translation, the direction of the message is from us to someone else and it is our responsibility to ensure that the message is grammatical- and content-wise correct. An example here is communication by e-mail or filling in foreign language forms. The quality standard here is higher than in gisting.

Reasonable users would not blindly trust the output of a publicly available MT service. Further, they would not paste it into an e-mail and would not send it to someone. In both cases, inbound and outbound translation, feedback on quality is needed. This is true especially in outbound translation because small grammatical errors in inbound translation do not prevent understanding of the message. They, however, do matter in outbound translation because ungrammatical messages could lead to being perceived as unprofessional. This feedback on translation quality should tell users if the translation is correct and if not, which parts contain errors.

The goal is also to increase the users' confidence in machine translation, which, however, cannot be done just by always reporting that everything is correct. To build trust, the whole complex MT service needs to look reliable, that is, to report on adequate occasions that the MT failed and what to do to fix it.

1.2. Existing Approaches

The most rudimentary form of outbound translation solution, especially when the target language is completely unknown to the users, is to perform a manual roundtrip translation (machine translating the result of the forward translation back to the original language). This relies heavily on the assumption that a potential error would only happen in the forward translation and never in the backward translation. This is sometimes not the case. New errors can happen in the backward translation as well and in some cases, the new error may revert the original one. This is shown in the last row (English MT) in Figure 1.

Orthogonal to this is automatic MT quality estimation (QE). The goal of this task is to determine which parts of the forward translation are poorly translated. It is done on word-, phrase-, sentence- or document-level. Companies such as Memsource[1] and Unbabel[2] use QE models (Kepler et al., 2019) to automatically decide which texts

[1] memsource.com/blog/2018/10/01/machine-translation-quality-estimation-memsources-latest-ai-powered-feature/

[2] unbabel.com/blog/unbabel-translation-quality-systems/

svírá úhel *forms an angle*	$\xrightarrow{\text{de}}$	Er schließt den Winkel. *he closes the angle*	$\xrightarrow{\text{cs}}$	Zavírá úhel. *closes the angle*
svírá úhel *forms an angle*	$\xrightarrow{\text{fr}}$	Sait l'angle *knows the angle*	$\xrightarrow{\text{cs}}$	Zná úhel *knows the angle*
svírá úhel *forms an angle*	$\xrightarrow{\text{en}}$	grips the angle *grips the angle*	$\xrightarrow{\text{cs}}$	svírá úhel *forms an angle*

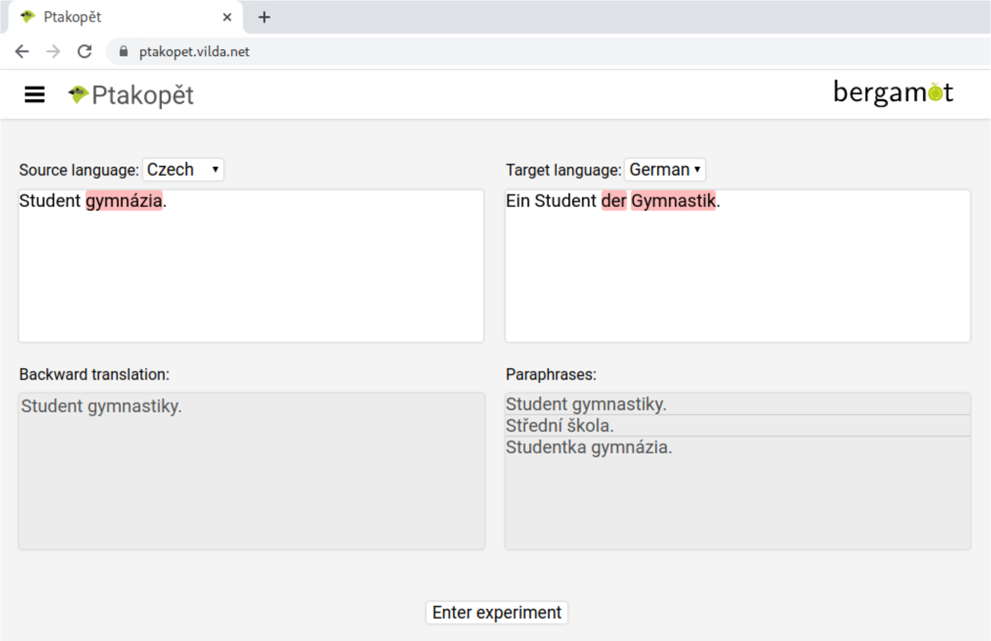
Figure 1. Example of error masking in backward translation in English MT compared to German and French MT in which the forward translation error is revealed. Based on a figure from Zouhar (2020). Capitalization and punctuation preserved from the MT output.

need to be post-edited and to what extent. Such tools are, however, missing in publicly available machine translation services, such as Microsoft Bing Translator, Google Translate or DeepL. The last two services provide alternatives to words and phrases, respectively. Showing alternative translations can also lead to higher user confidence in the system, but it does not help in case the users do not know the target language at all.

1.3. Source Complexity Application

In the context of outbound translation, we would also like to let the users know which parts of the input they should reformulate and focus on to make the output better. Highlighting poorly translated words is the most straightforward application of QE. By using word-alignment between the source text and the translation, we can then estimate which words in the input map to the problematic words in the output.

However, for some MT errors, it is not a specific word that worsens the translation but rather problems in agreements or syntactic structures. Source tokens selected by the described approach are not always responsible for the wrong translation and substituting them may not lead to an improvement. Niehues and Pham (2019) try to model source complexity by comparing the inputs to the training data seen by the MT, which leads to better results than mapping QE to source.



Source language: Czech

Student gymnázia.

Target language: German

Ein Student der Gymnastik.

Backward translation:

Student gymnastiky.

Paraphrases:

Student gymnastiky.
Střední škola.
Studentka gymnázia.

Figure 2. Ptakopět is used to translate a simple Czech noun phrase to German. QE highlights parts of both source and target that were translated incorrectly. Figure from Zouhar (2020).

2. Ptakopět

Ptakopět is a system for outbound machine translation and the exploration of user strategies. It was first presented in Zouhar and Bojar (2020); Zouhar (2020). It is publicly available,[3] the code is open-source[4] and a brief user and technical documentation is also available.[5]

Two other versions of Ptakopět (old-1 and old-2)[6] predate the current version. Their focus was purely outbound translation on the Internet. The current and final version of Ptakopět broadens the focus from only providing an outbound translation tool to also offering a system for analyzing user strategies in dealing with machine translation.

[3] ptakopet.vilda.net

[4] github.com/zouharvi/ptakopet

[5] ptakopet.vilda.net/docs

[6] github.com/zouharvi/ptakopet-old

2.1. Usage

The system offers backward translation, quality estimation, source complexity and paraphrases to help with outbound translation. All of these modules are demonstrated in Figure 2. The top left text area is used for input, the top right for MT output, the bottom left for backward translation and the bottom right for input paraphrases.

The output of the quality estimation is used to highlight erroneous words in the translated output. These words are mapped to the input using word-alignment to estimate problematic source words, which are then also highlighted.

In Figure 2 we see that for the input *student gymnázia* (*grammar school student*) the output *ein student der Gymnastik* (*a student of gymnastics*) appeared instead of *ein Gymnasiast* (*a grammar school student*). Without knowing any German and just by looking at the output, the users could get a false sense of having received a correct translation because the output generally looks like a valid German sentence and the lexemes look similar to what would someone expect in this translation.

Fortunately, this translation error manifested itself in the backward translation and was also detected by quality estimation, which got projected to the source sentence. The affected erroneous parts were highlighted red. The users are now informed that in order to make the translation correct, they must change the last word in the input. For that, the paraphraser module offers several possibilities.[7]

The system is connected to many backends which provide machine translation, quality estimation, word alignment, tokenization, paraphrasing and logging. Naturally, not all backends work with every language pair, and some are more suited for specific testing needs. A menu is shown after clicking the button in the top left corner in Figure 2. It contains settings for switching between the available backends. In Section 3.2 we show how experiment definitions interact with these settings.

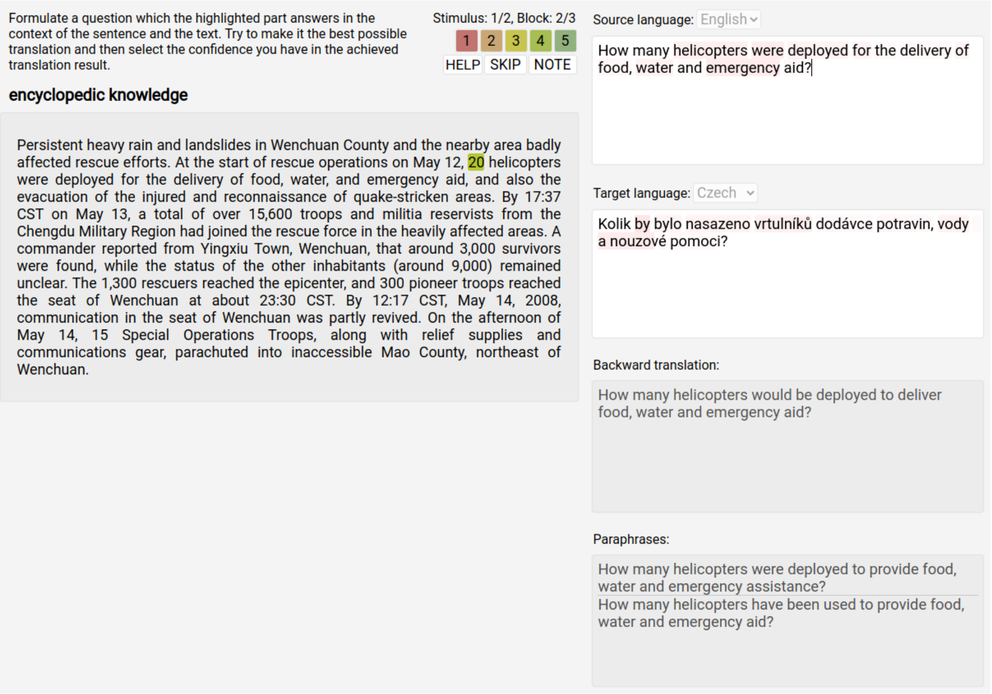
2.2. Architecture

The system is composed of three parts: server, frontend and experiment design and data processing suite. The Ptakopět server[8] is used to provide some of the backend services for the frontend, such as quality estimation or tokenization and has no special role compared to other backends except for logs collection.

The frontend is written in TypeScript and is designed to be highly modular. As a result, adding new backend wrappers can be done easily by implementing a single function. Some of these wrappers may not even use the network to resolve requests and compute the result locally as is done, for example, with one tokenization backend wrapper.

[7] Unfortunately none of the three paraphrases suggested in Figure 2 is helpful, each for a different reason.

[8] github.com/zouharvi/ptakopet-server



Formulate a question which the highlighted part answers in the context of the sentence and the text. Try to make it the best possible translation and then select the confidence you have in the achieved translation result.

Stimulus: 1/2, Block: 2/3

encyclopedic knowledge

Persistent heavy rain and landslides in Wenchuan County and the nearby area badly affected rescue efforts. At the start of rescue operations on May 12, 20 helicopters were deployed for the delivery of food, water, and emergency aid, and also the evacuation of the injured and reconnaissance of quake-stricken areas. By 17:37 CST on May 13, a total of over 15,600 troops and militia reservists from the Chengdu Military Region had joined the rescue force in the heavily affected areas. A commander reported from Yingxiu Town, Wenchuan, that around 3,000 survivors were found, while the status of the other inhabitants (around 9,000) remained unclear. The 1,300 rescuers reached the epicenter, and 300 pioneer troops reached the seat of Wenchuan at about 23:30 CST. By 12:17 CST, May 14, 2008, communication in the seat of Wenchuan was partly revived. On the afternoon of May 14, 15 Special Operations Troops, along with relief supplies and communications gear, parachuted into inaccessible Mao County, northeast of Wenchuan.

Source language: English

How many helicopters were deployed for the delivery of food, water and emergency aid?

Target language: Czech

Kolik by bylo nasazeno vrtulníků dodávce potravin, vody a nouzové pomoci?

Backward translation:

How many helicopters would be deployed to deliver food, water and emergency aid?

Paraphrases:

How many helicopters were deployed to provide food, water and emergency assistance?

How many helicopters have been used to provide food, water and emergency aid?

Figure 3. An example stimulus is presented in Ptakopět. Based on this, the users are required to produce a translation in for them an unknown language with the help of offered tools. The target output is content-wise correct, but it is ungrammatical, because the forward MT ommited a preposition "k" ("to" or "for").

3. Deploying and Running Experiments

In this section we demonstrate how experiments in Ptakopět look like and how to design them from the technical point of view.

3.1. Usage

Experiments in Ptakopět are done in the form of showing stimuli to the users and asking them to finish the stimuli with the help of Ptakopět. A stimulus is anything that incentivizes users to produce a text in a foreign language. In Zouhar and Bojar (2020) the stimuli was reporting issues to an IT helpdesk, inquiring into administrative issues and answering encyclopedic questions from the question–answering dataset SQuAD

(Rajpurkar et al., 2018). Technically a stimulus can be any HTML entity, such as text, an image or any more complex web form.

An example stimulus based on SQuAD is shown in Figure 3. A specific piece of information is highlighted to which the users are expected to formulate a question in a foreign language using Ptakopět. In this scenario, we assume that the users speak English and do not speak Czech, so they are not able to evaluate the produced forward translation manually. After they are done working with this current stimulus, they select the level of confidence they have of the produced translation. Multiple events are logged during their work, notably: incoming forward/backward translation, quality estimation, source complexity and paraphrases.

For experiments it is sometimes also desirable to change the configuration settings. This way we can for example enable quality estimation only sometimes (or change the backend) and see whether this change affects the confidence and translation quality.

3.2. Experiment design

An experiment is defined in a single JSON file, which gets loaded when a user tries to log in using *Enter experiment* in Figure 2. Every experiment participant has an assigned user ID (`UID`) by which they log in and are referenced in the experiment definition. We use the concept of *baked queues*. We determine the sequence of stimuli together with their specific configurations for every user in advance as opposed to choosing a random stimulus during the experiment. This way we can check beforehand that the generated queues cover for example every stimulus with a specific number of configurations.

We also present the stimuli in so called *blocks*. They are used only for psychological management purposes so that it is easier for users to split their work into several phases. Users are notified by an alert box every time they completed a block. Our data confirms that there is no connection between the work quality and position of the stimulus in the block.

The experiment definition contains:

- **baked queue:** an array of arrays of stimuli for every user (baked queues in blocks)
- **stimuli dictionary:** a string (valid HTML) for every stimulus identified by stimulus ID (`SID`)
- **configuration rules:** an array of regex rules and changes in settings that get applied when the given stimulus matches the rule's regex

Since the same stimulus can appear in different combinations for different users, the baked queue references stimuli by `SID` Extended (`SIDE`). It is nothing more than a SID with a suffix, separated by a special token `#` (e.g. `p105#bt.y.pp.n.qe.y`). It does get considered when looking up the stimuli content in the stimuli dictionary.

Lastly, the structure contains rules which get applied to the settings whenever the current stimulus matches the regex. For example a rule with the regex `^.*#.*qe\.y.*$`

would match the previous SIDE and could then turn the QE on. Multiple settings can be applied at the same time. We find this pattern to be powerful because it allows us to encode the configuration in user baked queues.

4. Results

The pilot experiment in Zouhar and Bojar (2020) suggested that working with an enhanced machine translation system increases production quality. Unfortunately, we then did not collect self-reported user confidence and had every module (except for the paraphraser, which was not part of the experiment) always enabled.

In a small follow-up experiment, we asked 10 annotators to work with Ptakopět on web-form stimuli (e-commerce domain).[9] An excerpt of an online form with a highlighted field was shown (similarly as in Figure 3). The annotators were then asked to fill this field in the target language. We used English for the source language and Czech for the target language. Out of all annotators, 7 knew no Czech, 2 knew Russian (similar to Czech in some aspects) and 1 knew very little Czech. The users were each shown 70 stimuli and they reported their confidence in the produced translation on a scale from 1 (worst) to 5 (best). The 70 stimuli were shared across all users, but one stimulus was seen by different users with different configurations. The configuration was not constant for one user.

We used two MT systems: (1) low–quality, trained on a subsample of 5 million sentence pairs from CzEng 1.7 Bojar et al. (2016), and (2) high–quality, winning MT model of Czech–English News Translation in WMT 2019 (Popel et al., 2019) trained on over 120 million authentic and backtranslated sentence pairs in total.

The enhancement modules comprised backward translation, quality estimation and paraphrases. Backward translation was provided by the abovementioned MT systems trained in the opposite direction. Quality estimation is supplied by a binary supervised classifier, whose word-level predictions are based on glass-box confidence indicators extracted from the output of a neural MT model. Previously, this approach has been successfully employed in sentence-level quality estimation (Fomicheva et al., 2020). We trained the classifier on texts associated with the stimuli collected for this experiment. Paraphrasing was performed using a round-trip translation from English to English through 41 mainly European pivot languages, producing one paraphrase for each pivot language. The MT system used here is based on the Transformer model (Vaswani et al., 2017) sharing the encoder and the decoder for all languages. The annotators were presented only with a selection of paraphrases yielded by 10 higher-resourced pivot languages. The modules were turned on and off during this experiment to see how they affect the confidence and quality.

[9] This experiment is a part of a wider-range experiment in cooperation with University of Edinburgh and University of Sheffield which is still in progress. The complete results of this experiment will be presented in a separate publication.

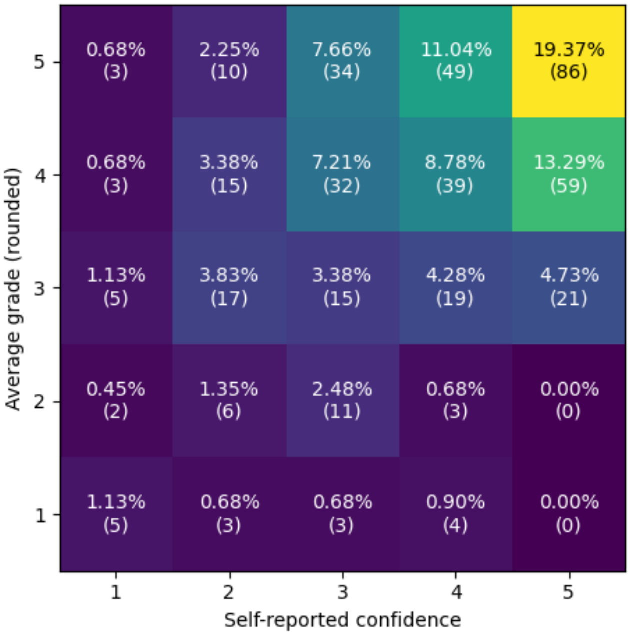


Figure 4. Heatmap of the average rounded grade and self-reported user confidence showing the distribution (all numbers sum to 1). First row in each cells shows the percentage and the second the number of instances. Both axes are 1 (worst) to 5 (best).

The results were then graded by 3 native Czechs on the scale from 1 (worst) to 5 (best). In Figure 4, we show the heatmap for self-reported confidence and translation quality scores. The distribution mass is concentrated in the upper right part of the graph (both high confidence and good translation quality) with very few outliers where the confidence did not match the translation quality grade.

The Fleiss' kappa between the native Czech speakers was 0.36 and the average Pearson correlation coefficient was 0.68. The Pearson correlation coefficient between the average self-reported confidence and average translation quality grade was 0.38.

The relationship between the number of tokens and confidence and quality scores is shown in Figure 5. The translation quality decreases with source sentence length. This is expected because longer sentences are usually more complex and harder to translate. The confidence follows a similar pattern, only with slightly more noise. This same trend could be the result of users correctly identifying errors in long translations using the provided tools. Their judgement could also be based on their apriori knowledge and experience with MT systems which perform poorly on longer sentences.

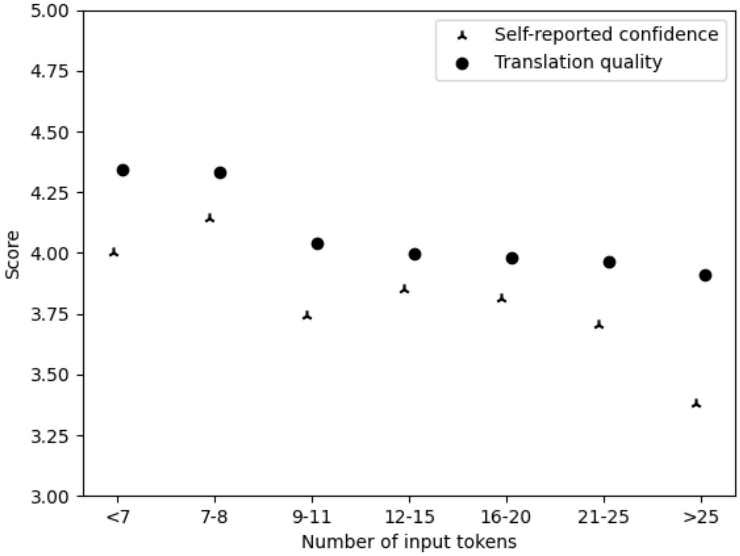


Figure 5. Relationship of the input sentence length and received confidence and translation quality score. Both decline with increased source length.

	Low-quality MT	High-quality MT
Translation quality	3.91	4.18
Self-reported confidence	3.70	3.92
Time-spent	86.95s	77.45s
Interactions	13.17	11.31

Table 1. The effect of MT quality on the average stimulus confidence, translation quality, spent time and the number of user interactions.

Differences in scores, spent time and the number of interactions[10] per MT model are shown in Table 1. With the higher quality of the model, the confidence and translation quality increased by 8.22% and 6.91%, respectively. This means that increasing MT model quality had a higher effect on the translation quality than on the confidence. Users spent on average 9.5 more seconds with the lower–quality MT. This is not because server responses took longer to complete for the lower–quality MT[11] – the number of interactions for the lower–quality MT was also higher. This means that the users used the interface more even though they had not been told what MT model was responding to their translation requests.

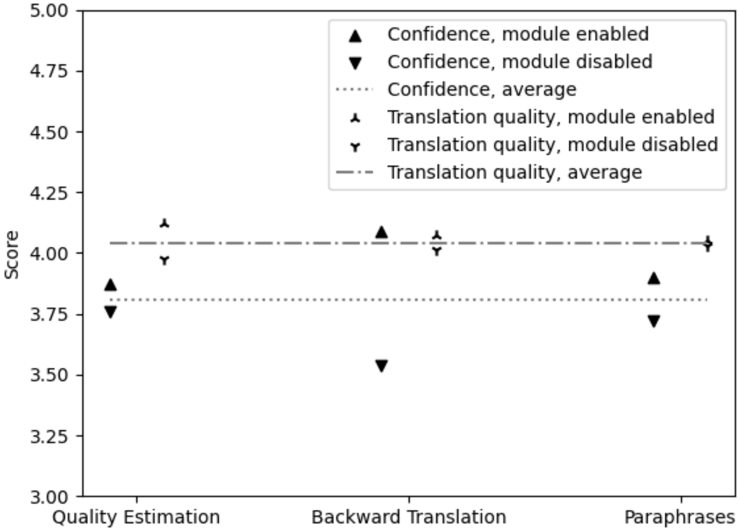


Figure 6. The effect of module presence on confidence and translation quality. Modules help in all cases but to varying degrees.

[10] This was measured by the number of backward translation requests which are started every time forward translation finishes (started upon source input) or the users manually edit the output.

[11] The average translation request duration for an 8-token sentence was 3.2 seconds.

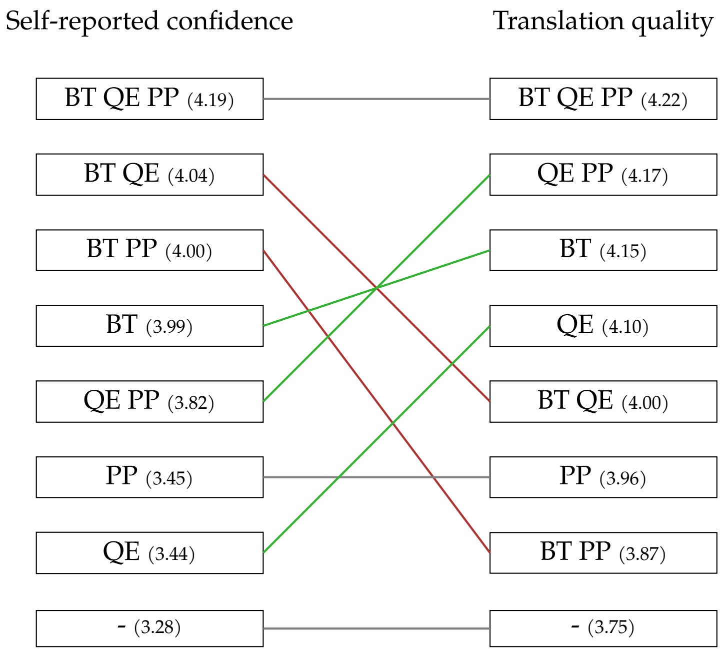


Figure 7. Lists of module configurations sorted by self-reported user confidence (left) and translation quality rated by native Czech speakers (right). BP: backtranslation, QE: quality estimation, PP: paraphrases.

Changing the configuration settings during the experiment helped us in examining which modules helped the most during this task. In Figure 6, we see how the presence or absence of a module affected user confidence and translation quality. In all cases, the presence of a specific module did not worsen the confidence nor the quality. The changes in quality are, however, much less significant than the changes in user confidence. This is most notable in backward translation for which the difference in confidence is 0.55, but the difference in quality is only 0.06.

In Figure 7 we show two columns in comparison. The left column lists configurations sorted by the average self-reported user confidence while the right one lists configurations sorted by the translation quality, respectively. The position of configurations with all modules on (BT QE PP) or off (-) is preserved, but there are many changes in the position of other configurations.

From these figures, we see that any extra module helps in increasing both confidence and translation quality. This refines the previous results that especially backward translation improves machine translation user experience. It does improve it, but it mainly increases users' confidence and not the translation quality.

5. Conclusion

In this article, we described the issue of outbound translation and user confidence in machine translation. We focused on the system Ptakopět and elaborated on the way by which experiments on human annotators are designed in this tool and the design patterns we found useful in the context of online annotation environments.

Finally, in Section 4 we showed some of the results of experiments done using this system. They suggest that enhancements in the form of backward translation, quality estimation and paraphrases help in increasing user confidence more than objective translation quality.

The role of the user is often overlooked in MT research, which is in stark contrast to the fact that there exist tools usable by the users that affect both the confidence and the quality. In future experiments, we would like to extend the functionality of Ptakopět even further to describe the effect of possible enhancement tools for MT rigidly.

Acknowledgements

This project has received funding from the grants H2020-ICT-2018-2-825303 (Bergamot) of the European Union and 19-26934X (NEUREM3) of the Czech Science Foundation.

We also used language resources developed and/or stored and/or distributed by the LINDAT-Clarin and LINDAT/CLARIAH-CZ projects of the Ministry of Education of the Czech Republic (projects no. LM2010013 and LM2018101).

Bibliography

Bojar, Ondřej, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Artificial Intelligence, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London, 2016. Springer International Publishing. doi: 10.1007/978-3-319-45510-5_27.

Fomicheva, Marina, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8(0):539–555, 2020. doi: 10.1162/tacl_a_00330.

Kepler, Fábio, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics–System Demonstrations*, pages 117–122, Florence, Italy, 2019. Association for Computational Linguistics.

Niehues, Jan and Ngoc-Quan Pham. Modeling Confidence in Sequence-to-Sequence Models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages

575–583, Tokyo, Japan, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8671.

Popel, Martin, Dominik Machaček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. English-Czech Systems in WMT19: Document-Level Transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5337.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Zouhar, Vilém. Enabling Outbound Machine Translation. Bachelor thesis, Charles University, Faculty of Mathematics and Physics, 2020.

Zouhar, Vilém and Ondřej Bojar. Outbound Translation User Interface Ptakopět: A Pilot Study. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6969–6977, Marseille, France, 2020. European Language Resources Association.

Address for correspondence:
Vilém Zouhar
zouhar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic