Device-Type Influence in Crowd-based Natural Language Translation Tasks

Michael Barz[1], Neslihan Büyükdemircioglu[2], Rikhu Prasad Surya[2],
Tim Polzehl[2], and Daniel Sonntag[1]

[1] German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, Saarbrücken, Germany
{michael.barz,daniel.sonntag}@dfki.de
[2] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
neslihan.bueyuekdemircioglu@tu-berlin.de,
{rikhu.p.s,tim.polzehl}@qu.tu-berlin.de

Abstract. The effect of users' interaction devices and their platform (mobile vs. desktop) should be taken into account when evaluating the performance of translation tasks in crowdsourcing contexts. We investigate the influence of the device type and platform in a crowd-based translation workflow. We implement a crowd translation workflow and use it for translating a subset of the IWSLT parallel corpus from English to Arabic. In addition, we consider machine translations from a state-of-the-art machine translation system which can be used as translation candidates in a human computation workflow. The results of our experiment suggest that users with a mobile device judge translations systematically lower than users with a desktop device, when assessing the quality of machine translations. The perceived quality of shorter sentences is generally higher than the perceived quality of longer sentences.

Keywords: Crowd-based Translation · Natural Language Translation · Machine Translation · Human Judgment · Crowdsourcing.

1 Introduction

Nowadays, crowdsourcing is used for a variety of tasks ranging from image tagging to text creation and translation [2, 10, 7]. Incorporating humans in complex workflows introduces several challenges including a large variety in their contribution quality [5]. Recent research investigates approaches in which humans are included, if a machine learning model is uncertain, for example, in the domain of natural language translation.

We consider crowd-enabled natural language translation, particularly workflows in which human translators compete against machine translation systems that are developed for low cost and high-speed [1]. Previous research has shown that crowdsourced translations are of higher quality than machine translations, but professional human translators still outperform the crowd [8, 6]. Hence, roll-outs of respective business applications fail due to a lack quality in automated translation and require a human quality assurance. Several concepts and workflows are proposed for ensuring high translation quality, e.g., Minder and Bernstein [8] investigate the suitability of iterative and parallel workflow patterns for generating translations of high quality. Zaidan et al. [11] propose a model for automatically selecting the best translation from multiple translation candidates and calibrate it using professional reference translations. In the domain of machine translation, common metrics for quality assessment include human judgements, but also automated measures that compare translation candidates against reference translations [3]. Gadiraju et al. [4] investigate the effect of the device type on the quality of different crowd tasks, but did not include translation.

In this work, we focus on the influence of the device type of the human assessor on its quality assessment in a crowd-based translation setting and for machine translation. We present our preliminary results of a corresponding experiment in which the crowd was asked to translate and rate a subset of the IWSLT parallel corpus[3]. In addition, we asked them to rate machine translations of the same sentences.

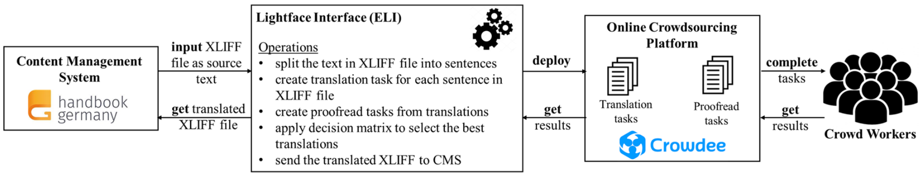
Content Management System
handbook germany
input XLIFF file as source text
get translated XLIFF file
Lightface Interface (ELI)
Operations
split the text in XLIFF file into sentences
create translation task for each sentence in XLIFF file
create proofread tasks from translations
apply decision matrix to select the best translations
send the translated XLIFF to CMS
deploy
get results
Online Crowdsourcing Platform
Translation tasks
Proofread tasks
Crowdee
complete tasks
get results
Crowd Workers

Fig. 1. Workflow diagram of the considered crowd-based translation system.

2 Crowd-Translation System

We implement a simple crowd-based workflow to investigate biases in the quality assessment of crowd-based translations. Our system is implemented using the crowdsourcing platform Crowdee [9], as it has shown to meet scientific requirements in the past, and seamlessly integrates into the enterprise-level content management system (CMS) Adobe Experience Manager, which can be used for administrating the content of multilingual websites, e.g., the refugee information portal handbookgermany.de (see figure 1). Our prototype consists of a combination of iterative and parallel processes including a translation and a proofreading/assessment task to obtain translations of adequate quality. In this setting, we investigate the behavior of human judgments depending on the device type used for the assessment task. As machine translations are commonly used for generating translation candidates, we investigate the same for translations of

[3] https://sites.google.com/site/iwsltevaluation2016/mt-track

the state-of-the-art machine translator Google translate[4]. For a given article, our workflow generates sentence-based translation tasks which can be processed in parallel. Resulting translations are used to create proofread tasks, which ask the crowd to rate and, if necessary, improve the candidate. We use these ratings for our evaluation. One aim of this work is to find suitable metrics based on human judgments and incorporating the inherent bias for (semi-)automatically selecting the best translations.

3 Experiment

For our experiment, we use a subset of the parallel IWSLT evaluation corpus including English transcriptions of TED talks and reference translations in Arabic. We selected an article focusing on climate change[5]. We recruit bilingual crowdworkers via social media channels targeting countries where most people speak English or Arabic. We ask these crowdworkers to participate in language proficiency tests for both languages designed by native speakers. Workers that reach a proficiency of 80% or higher in both tests are selected for participation. For the translation stage, we collect 3 translations for each sentence resulting in a total of 180 translation tasks. Subsequently, we publish 3 proofread tasks for each candidate yielding about 540 human judgments on 5-pt Likert scales. Please note, the actual number of analyzed judgements differs due to illegal or rejected crowd contributions and parallel execution of task repetitions. Overall, we limit the maximum number of translation tasks per crowdworker to 3, in order to include more workers. Similarly, we ask crowd workers to rate machine translations of the source sentences. The human judgment constitutes the dependent variable, the device type used for the assessment task is the independent variable in our experiment. Further, we observe the sentence length as a control variable, as we expect longer sentences to achieve lower quality judgments due to, e.g., lower translation quality or lower perceived quality. We consider two device types, mobile and desktop devices, and split the sentence length into a low and high group based on the median length: we split at 12.5. We apply Kruskal-Wallis tests for significant differences between groups on a 1% significance level, a standard procedure for an analysis of variance for non-parametric distributions and robust against unequally sized groups.

4 Results & Discussion

Concerning human judgments for crowd-translations ($n = 662$), we do not see a significant difference between quality assessments from mobile and desktop users. As a potentially influencing factor, we have only 83 samples from mobile users, yielding a very unbalanced dataset in contrast to our data concerning the machine translations. However, we do observe that human quality judgments are

[4] generated using https://cloud.google.com/ml-engine/

[5] TED talk with ID 535 from TED2009; segments 1 to 60.

significantly lower for long sentences ($Mdn = 4.16$) compared to short sentences ($Mdn = 4.33$). The overall human judgment is $Mdn = 4.3$ ($SD = .69$), which can be interpeted as good overall translation quality.

Concerning the human judgments for machine translations ($n = 163$), we observe that quality assessments from users with mobile devices ($Mdn = 3.55, n = 75$) are lower than those submitted with a desktop device ($Mdn = 3.93, n = 88$). For mobile users, this includes 41 assessments for short sentences and 33 assessments for longer ones. We observed a similar ratio for desktop users: 50 for short and 38 for long sentences. Further, we observe that long and short sentences have approximately the same frequency in the mobile and in the desktop group. This supports the implication that the differences in the quality assessments are induced by the device type and not by an unbalanced distribution of long and short sentences in each group. These findings are in line with the findings of Gadiraju et al. [4]: Using mobile devices negatively impacts the result of crowd-tasks. Here, a lower usability might be the cause for systematically lower quality assessments. However, additional factors originating from the workflow design might as well influence the quality assessments which are not taken into account in this paper.

5 Conclusion

We investigated the bias introduced by the device type used for assessing translation quality in crowd-based translation workflows. The results of our study suggest that we can confirm our hypothesis that users assessing translations with the mobile phone provide systematically lower results. This should be taken into account for, e.g., automated translation candidate selection based on human judgments. However, we reject generalizing this statement due to small amount of data included here. Future work should investigate this aspect on a more complete dataset; it should also include further factors that might add a bias to the quality assessment. Ongoing work includes language proficiency and user characteristics. In additon, we found a decline in translation quality for different length of sentences, which is subject to ongoing work on analysis whether this originates from actually lower translation performance on longer sentences or whether is is rather due to a higher task complexity.

6 Acknowledgments

We want to thank EIT Digital for supporting our research project ERICS and our collaborators in this project from T-Systems MMS, Aalto University and Crowdee.

References

1. Barz, M., Polzehl, T., Sonntag, D.: Towards hybrid human-machine translation services. EasyChair Preprint no. 333 (EasyChair, 2018). https://doi.org/10.29007/kw5h

2. Borromeo, R.M., Laurent, T., Toyama, M., Alsayasneh, M., Amer-Yahia, S., Leroy, V.: Deployment strategies for crowdsourcing text creation. Information Systems **71**, 103–110 (2017)
3. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Way, A.: Is Neural Machine Translation the New State of the Art? The Prague Bulletin of Mathematical Linguistics **108**(1), 109–120 (jan 2017). https://doi.org/10.1515/pralin-2017-0013, http://www.degruyter.com/view/j/pralin.2017.108.issue-1/pralin-2017-0013/pralin-2017-0013.xml
4. Gadiraju, U., Checco, A., Gupta, N., Demartini, G.: Modus operandi of crowd workers: The invisible role of microtask work environments. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(3), 49 (2017)
5. Goto, S., Ishida, T., Lin, D.: Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes. In: Fourth AAAI Conference on Human Computation and Crowdsourcing (2016)
6. Hu, C., Bederson, B.B., Resnik, P., Kronrod, Y.: MonoTrans2 : A New Human Computation System to Support Monolingual Translation. Chi '11 pp. 1133–1136 (2011). https://doi.org/10.1145/1978942.1979111
7. Malone, T.W., Rockart, J.F.: Computers, networks and the corporation. Scientific American **265**(3), 128–137 (1991)
8. Minder, P., Bernstein, A.: How to translate a book within an hour: towards general purpose programmable human computers with crowdlang. In: Proceedings of the 4th Annual ACM Web Science Conference. pp. 209–212. ACM (2012)
9. Naderi, B., Polzehl, T., Wechsung, I., Köster, F., Möller, S.: Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
10. Ross, J., Irani, L., Silberman, M., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: CHI'10 extended abstracts on Human factors in computing systems. pp. 2863–2872. ACM (2010)
11. Zaidan, O.F., Callison-Burch, C.: Crowdsourcing translation: Professional quality from non-professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 1220–1229. Association for Computational Linguistics (2011)