EasyChair Preprint

№ 333

Towards Hybrid Human-Machine Translation Services

Michael Barz, Tim Polzehl and Daniel Sonntag

July 9, 2018

Towards Hybrid Human-Machine Translation Services

MICHAEL BARZ, German Research Center for Artifial Intelligence (DFKI), Saarbrücken
TIM POLZEHL, Quality and Usability Lab, Technische Universität Berlin
DANIEL SONNTAG, German Research Center for Artifial Intelligence (DFKI), Saarbrücken

1. INTRODUCTION

Crowdsourcing is recently used to automate complex tasks when computational systems alone fail. The literature includes several contributions concerning natural language processing, e.g., language translation [Zaidan and Callison-Burch 2011; Minder and Bernstein 2012a; 2012b], also in combination with active learning [Green et al. 2015] and interactive model training [Zacharias et al. 2018].

In this work, we investigate (1) whether a (paid) crowd, that is acquired from a multilingual website's community, is capable of translating coherent content from English to their mother tongue (we consider Arabic native speakers); and (2) in which cases state-of-the-art machine translation models can compete with human translations for automation in order to reduce task completion times and costs. The envisioned goal is a hybrid machine translation service that incrementally adapts machine translation models to new domains by employing human computation to make machine translation more competitive (see Figure 1). Recently, approaches for domain adoption of neural machine translation systems including filtering of generic corpora based on sentence embeddings of in-domain samples [Wang et al. 2017] have been proposed, as well as the fine-tuning with mixed batches containing domain and out-of-domain samples [Chu et al. 2017] and with different regularization methods [Barone et al. 2017].

As a first step towards this goal, we conduct an experiment using a simple two-staged human computation algorithm for translating a subset of the IWSLT parallel corpus including English transcriptions of TED talks and reference translations in Arabic with a specifically acquired crowd. We compare the output with the state-of-the-art machine translation system Google Translate as a baseline.

2. EXPERIMENT

For this experiment, we apply our human computation algorithm (explained further down) to a subset of the parallel IWSLT evaluation corpus[1] that is commonly used as a gold standard for machine trans-

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

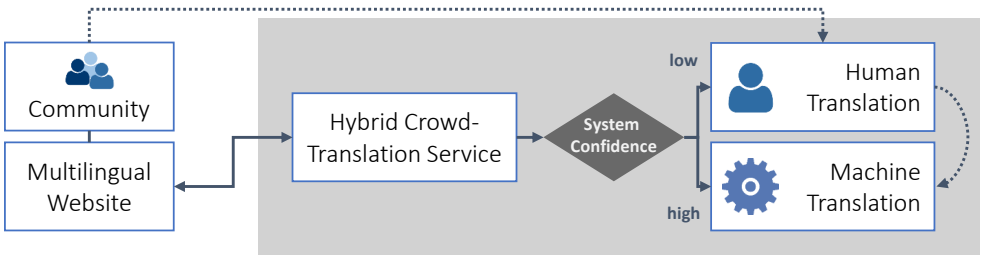

Fig. 1: Conceptual architecture of the hybrid translation system that obtains crowdworkers from the website's community and adapts the machine translation model to the new domain based on human-generated translations.

Table I. : Mean and standard deviation of performance metrics for workflow stages *S.1* and *S.2* and for the machine translation baseline *GT* including $n = 54$ complete translations.

Metric	**S.1**	**S.2**	**GT**
BLEU	$M = .387$ $SD = .189$	$M = .403$ $SD = .192$	$M = .37$ $SD = .144$
Human Judgement	$M = 3.97$ $SD = 1.18$	—	—

Table II. : Mean and standard deviation of performance metrics for *S.1* concerning the factor *platform*: Mobile (n=18) – Desktop (n=101).

Metric	**Mobile**	**Desktop**
BLEU	$M = .437$ $SD = .16$	$M = .443$ $SD = .21$
Task Completion Time	$M = 208.8s$ $SD = 148.6s$	$M = 196.9s$ $SD = 140.4s$
Human Judgement *(n=10 – n=44)*	$M = 4.72$ $SD = .395$	$M = 3.803$ $SD = 1.233$

lation systems: it contains English transcriptions of TED talks and corresponding reference translations in Arabic. We select 60 sentences from a random talk[2] as reference document that is semantically coherent, similar to inputs that we can expect from a website.

Our human computation algorithm includes two stages: a *translation stage* **S.1** that asks workers to translate a single sentence from English to Arabic from scratch; and a *proof-reading stage* **S.2** in which workers are asked to rate and, if necessary, improve the Arabic translation given the source sentence. It is executed on the crowdsourcing platform Crowdee[3] with crowdworkers invited via `handbookgermany.de`, an information portal for refugees. Participants are asked to complete a paid language proficiency test for English and Arabic which is offered as a job on Crowdee. Only those who reach a score of at least 80% for both languages are considered for translation and proof-reading tasks. First, we deploy 180 translation jobs; three for each sentence. After all translation jobs are completed, three proof-reading jobs are uploaded per translation candidate yielding 540 approved translations. The results of both stages are compared to the commercial machine translation system Google Translate[4] (**GT**). Additionally, we investigate the impact of the *platform* type used by crowdworkers (mobile or desktop) for **S.1** and the influence of the *sentence length* for all stages and the baseline.

For measuring the translation performance, we consider the task completion time and the automated quality metric BLEU [Papineni et al. 2001] which reports translation quality in a range between 0 and 1. The automated metric is computed using the translations from the IWSLT dataset as reference and each system's result as hypothesis. In addition, we investigate the crowdworkers' *language proficiency* and *human judgements*. The *human judgement* is reported on a 5 point Likert scale for each translation of **S.1** in the beginning of **S.2** (it is only available for **S.1** and if, at least, one proof-read job gets completed).

2.1 Results

First, we process the output of **S.1** by merging the translation results with references from the IWSLT corpus and baseline data from **GT**, computing the BLEU scores and adding further measures such as the language proficiency. Due to problems with some mobile devices, we have to exclude 61 entries with empty or wrongly encoded translations. Then, we merge the results from **S.2**: we assign the average values of each metric, because there are up to three proof-reads per translation. After excluding invalid samples, e.g., due to incomplete proof-reading jobs, we have a set of $n = 54$ translations for **S.2**. On average, the language proficiency is 0.94 ($SD = 0.14$) for Arabic and 0.88 ($SD = 0.05$) for English.

[2]TED talk with ID 535 from TED2009; segments 1 to 60.
[3]https://www.crowdee.de/
[4]generated using https://cloud.google.com/ml-engine/ on 6th of December 2017.

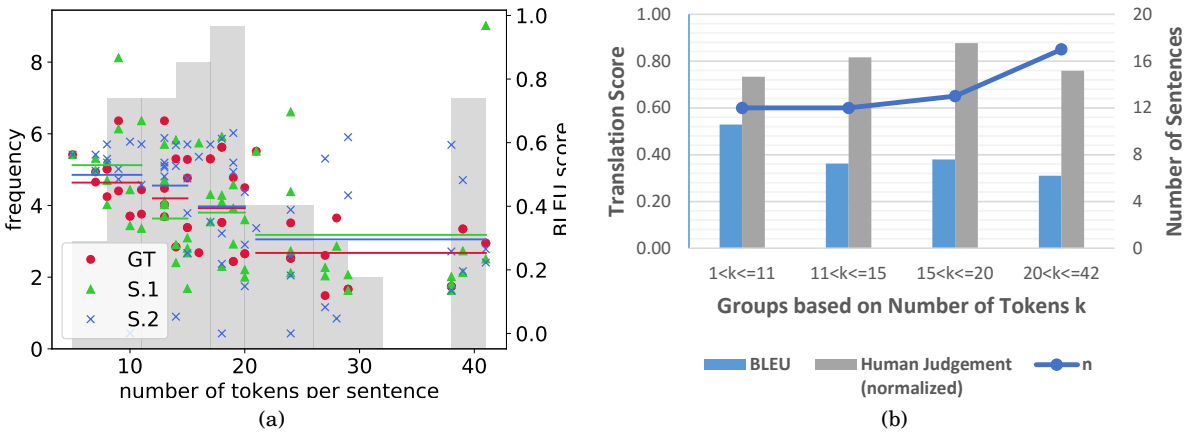

Fig. 2: (a) BLEU scores for each translation task and averaged for groups split by the number of tokens including all considered systems; (b) Mean BLEU score in comparison to normalized human judgements for different groups of **S.1**.

Including all complete translations ($n = 54$), the pairwise system comparisons **S.1–S.2** and **S.2–GT** yield no significant differences in means for BLEU (Wilcoxon signed-rank test); *human judgements* are available for **S.1** only and average to $M = 3.97$ (see Table I). Concerning **S.1**, we analyse 119 translation results, of which 18 are completed with a mobile device and 101 with a desktop-like device. *Mobile* and *desktop* users achieve similar BLEU scores and task completion times (see Table II). The Mann-Whitney U test confirmed that differences are not statistically significant. Concerning *human judgements*, we observed significantly better results for *mobile* users compared to *desktop* users as depicted in Table II ($U = 109.5, p = .005$). In addition, we split results in five groups with nearly equal sample size based on the *sentence length* according to percentiles. Group borders are $[0, 11, 15, 20, 42]$ tokens per sentence: the first group includes all sentences that contain k tokens with $0 < k \leq 11$. Figure 2a shows BLEU scores averaged per group and for each translation tasks including complete translations for all considered systems: the scores decrease with increasing *sentence length*. In Figure 2b, we compared the BLEU scores of **S.1** with the corresponding normalized *human judgements*. In contrast to BLEU scores, they increase with increasing *sentence length*.

3. DISCUSSION & CONCLUSION

Overall, the quality of crowd-based translations is not better than the automatic translation baseline (according to BLEU scores). However, as the automated BLEU score does not correlate well with human judgements (see Figure 2b), we will investigate more into crowd-based solutions and evaluations of subjective judgements. A future experiment should consider the fluency and adequacy of translations as metrics for text coherence. The comparably good results of the automatic machine translation suggest that these can be used as candidates for proof-reading to reduce the crowd workload as done in more sophisticated workflows [Minder and Bernstein 2012b]. Machine translations could also be used to identify and filter low-performing crowdworkers, which we identified in our scenario (see outliers of **S.2** in Figure 2a). Interestingly, the *mobile* users achieved better scores from humans, which might be casued by auto-correction features of modern smartphone keyboards. Our next experiments will focus on incremental model improvement of the hybrid translation service and its integration into existing dialogue frameworks [Sonntag et al. 2009; Sonntag et al. 2010; Sonntag 2010; Prange et al. 2017].

REFERENCES

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1489–1494. https://www.aclweb.org/anthology/D17-1156

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Association for Computational Linguistics, 385–391. DOI:http://dx.doi.org/10.18653/v1/P17-2061

Spence Green, Jeffrey Heer, and Christopher D. Manning. 2015. Natural Language Translation at the Intersection of AI and HCI. *Queue* 13, 6 (2015), 30. DOI:http://dx.doi.org/10.1145/2791301.2798086

Patrick Minder and Abraham Bernstein. 2012a. *Crowdlang: Programming Human Computation Systems - Interweaving Human and Machine Intelligence in a Complex Translation Task*. Technical Report. University of Zurich. 13 pages.

Patrick Minder and Abraham Bernstein. 2012b. How to translate a book within an hour. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*. ACM Press, New York, New York, USA, 209–212. DOI:http://dx.doi.org/10.1145/2380718.2380745

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics, Morristown, NJ, USA, 311. DOI:http://dx.doi.org/10.3115/1073083.1073135

Alexander Prange, Margarita Chikobava, Peter Poller, Michael Barz, and Daniel Sonntag. 2017. A Multimodal Dialogue System for Medical Decision Support inside Virtual Reality. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Saarbrücken, Germany, 23–26.

Daniel Sonntag. 2010. *Ontologies and Adaptivity in Dialogue for Question Answering*. AKA and IOS Press, Heidelberg.

Daniel Sonntag, Robert Nesselrath, Gerhard Sonnenberg, and Gerd Herzog. 2009. Supporting a rapid dialogue system engineering process. *Proceedings of the 1st IWSDS* (2009).

Daniel Sonntag, Norbert Reithinger, Gerd Herzog, and Tilman Becker. 2010. A Discourse and Dialogue Infrastructure for Industrial Dissemination. In *Spoken Dialogue Systems for Ambient Environments*, Gary Geunbae Lee, Joseph Mariani, Wolfgang Minker, and Satoshi Nakamura (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 132–143.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence Embedding for Neural Machine Translation Domain Adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*. Association for Computational Linguistics, 560–566. DOI:http://dx.doi.org/10.18653/v1/P17-2089

Jan Zacharias, Michael Barz, and Daniel Sonntag. 2018. A Survey on Deep Learning Toolkits and Libraries for Intelligent User Interfaces. (mar 2018). http://arxiv.org/abs/1803.04818

Omar F. Zaidan and Chris Callison-Burch. 2011. 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*. Association for Computational Linguistics, 1220–1229.