Improving Machine Translation of Educational Content via Crowdsourcing

Maximiliana Behnke[1], Antonio Valerio Miceli Barone[1], Rico Sennrich[1], Vilelmini Sosoni[2], Thanasis Naskos[3], Eirini Takoulidou[3], Maria Stasimioti[2], Menno van Zaanen[4], Sheila Castilho[5] Federico Gaspari[5], Panayota Georgakopoulou[6] Valia Kordoni[7], Markus Egg[7], Katia Lida Kermanidis[3]

[1]School of Informatics, The University of Edinburgh
[2]Department of Foreign Languages, Translation and Interpreting, Ionian University
[3]Department of Informatics, Ionian University
[4]Department of Communication and Information Sciences,Tilburg University
[5]ADAPT Centre, Dublin City University
[6]Deluxe Media Europe
[7]Humboldt-Universität zu Berlin, Germany
maximiliana.behnke@ed.ac.uk; amiceli@inf.ed.ac.uk; rico.sennrich@ed.ac.uk; vilelmini@hotmail.com; anaskos@ionio.gr; rinoulit@gmail.com; stasimioti.maria@gmail.com; mvzaanen@uvt.nl; sheila.castilho@adaptcentre.ie; federico.gaspari@adaptcentre.ie; yota.georgakopoulou@bydeluxe.com; kordonie@anglistik.hu-berlin.de; markus.egg@anglistik.hu-berlin.de; kerman@ionio.gr

Abstract

The limited availability of in-domain training data is a major issue in the training of application-specific neural machine translation models. Professional outsourcing of bilingual data collections is costly and often not feasible. In this paper we analyze the influence of using crowdsourcing as a scalable way to obtain translations of target in-domain data having in mind that the translations can be of a lower quality. We apply crowdsourcing with carefully designed quality controls to create parallel corpora for the educational domain by collecting translations of texts from MOOCs from English to eleven languages, which we then use to fine-tune neural machine translation models previously trained on general-domain data. The results from our research indicate that crowdsourced data collected with proper quality controls consistently yields performance gains over general-domain baseline systems, and systems fine-tuned with pre-existing in-domain corpora.

Keywords: MOOCs, neural machine translation, crowdsourcing

1. Introduction

The European Union Horizon 2020 **TraMOOC** project (Translation for Massive Open Online Courses) aims at enhancing multilingual access to online education by providing machine translation solutions for the educational content available in MOOCs, i.e. video lecture subtitles, slides, assignments, quiz text, and course discussion forum text (Kordoni et al., 2016). Educational content is translated from English into eleven European and BRIC languages (Bulgarian, Chinese, Croatian, Czech, Dutch, German, Greek, Italian, Polish, Portuguese, and Russian).

This specific domain, i.e. the content of online courses, imposes a set of challenging properties, such as extensive use of domain-specific terms and entities, frequent occurrence of unknown words, subtitle segmentation, spoken language characteristics and social web text properties. Achieving high-quality machine translation in these conditions therefore requires significant amounts of in-domain data for training and testing. Creating such data by hiring professional translators would be expensive, especially considering that it would have to be done for eleven target languages. Therefore we turn our attention to crowdsourcing as a cost-saving alternative.

The impact of non-expert input, i.e. translations provided by non-professional translators, on the development and the evaluation of machine translation engines has been investigated in previous research (Callison-Burch, 2009; Zaidan and Callison-Burch, 2011; Ambati, 2012). Crowdsourcing has been used to this end, and the main research concern has been whether input by the general crowd, that has no expertise in linguistics or translation studies, can improve the quality of large-scale machine translation systems in a manner that is cost- and time-effective.

In this work we use crowdsourcing with carefully designed quality controls to collect translations of MOOC material from English to the eleven project target languages. We combine this data with pre-existing in-domain educational data and we use it to build translation systems in a transfer learning approach: we first train neural machine translation systems on general-domain data and then fine-tune them with the collected in-domain data. We report large improvements of translation quality obtained by using the crowdsourced data over both the general-domain baselines and the models tuned only with the pre-existing in-domain data.

2. Crowdsourcing Translations of Online Educational Content on a Large Scale

In order to facilitate worker recruitment and task implementation, we selected an established commercial crowdsourcing platform which had a crowd of workers who were already members and categorized by their experience on the platform. Demographic information (country, language) pertaining to the crowd, that enables crowd selection, was taken into account. We selected Crowdflower[1], due to logistic (e.g. payment options) as well as technical reasons (configurability, reliability, size of crowd channels, quality control mechanisms, technical support).

2.1 Data Description

Texts from courses of MOOC providers Coursera[2] and Iversity[3] constituted the in-domain English data sources to

[1] https://www.crowdflower.com/
[2] https://www.coursera.org/

be translated via crowdsourcing, as well as the QCRI Educational Domain Corpus (QED)[4] and the Videolectures.net online educational video library[5]. The course contents varied from technical (e.g. Finance) to non-scientific (e.g. Future of Storytelling). From these datasets, texts in two types of text genres were identified, namely formal (lecture subtitles, slides, assignments, quizzes) and informal forum discussions. More details pertaining to the data sources and their processing can be found in Sosoni et al. (2018).

The formal text presented a high occurrence of domain-specific terms and expressions, as well as spontaneous speech characteristics, i.e. repetitions, interjections, truncated clauses, segments marked as inaudible. This is not surprising, as a significant part of this type of text consists of subtitles. Informal text has the idiosyncrasies of social web text, i.e use of slang, misspellings, lexical variants, abbreviations, etc. as forum posts are treated like social network interactions. Furthermore, to a large extent, MOOC students are non-native speakers of English and their language skills vary (DeBoer et al., 2013).

Bearing in mind that the crowdsourced translations were to be used for training, tuning and testing the translation engines for the eleven language pairs involved in TraMOOC, the project aimed at collecting a significant number of translated segments. Approximately 95,000 segments were chosen to be translated per language pair; Table 1 shows the data source distribution in more detail.

	Training	Testing and tuning
Iversity	30000	2500
Videolectures.net	-	2500
Coursera	27000	-
QED	23000	-

Table 1: Data source distribution in number of (English) segments.

The segments constituting the testing and tuning datasets were to be translated by at least two workers each, for redundancy purposes.

2.2 Running and Monitoring the Experiment

Workers were provided with detailed language-independent, as well as language-specific instructions that explained how to cope with the various linguistic phenomena present in the source datasets. Instructions were very explanatory and included specific examples and ways to cope with typical issues. Test questions (which were based on choosing the optimal translation among three options) were designed to help evaluate a worker's performance and ban spammers during the data collection process. The copy functionality was disabled in order to discourage workers from copy-pasting output from online MT systems. Pilot trials were subsequently run on a sample of the dataset and a subset of the languages to help configure the settings of the crowdsourcing task, such as the number of segments to be displayed per page, the minimum and maximum time limits that a worker would be allowed to work on one page, the acceptable worker accuracy threshold, etc.

After further parameterisation of the crowdsourcing platform following the pilot trials, the translation crowdsourcing task was run for all language pairs and the entire dataset. In order to control spammers, as well as provide a training opportunity to trustworthy workers (who needed to familiarize themselves with the text domain and genre), the main task consisted of two phases: a quiz mode, where workers were asked to pass an evaluation microtask, and a work mode, where the actual translations were submitted.

Settings were adjusted and adapted to meet the needs of every language pair, as worker flow varied significantly among languages. The number of test questions (90-300), their difficulty level, the timeframes for task completion, the worker fee (0.04-0.08/segment), the countries that the task was open to, the workers' experience level (according to contributors' categorization in experience levels 1/inexperienced and 2/experienced by CrowdFlower), were all contingent on the flow of each language. The main translation collection task was run from March until June 2017.

Close and constant monitoring of the workers' annotation process was crucial for identifying malicious behaviour among workers, keeping track of the workflow and ensuring quality translations. The monitoring process was automated to a large extent in order to optimally handle multiple language pairs. Automated software tools were developed in order to keep track of the time spent by each worker per page, the page submission time, worker accuracy level, their geographic location, and the difference in length between the source and the translated sentences (a difference of more than 60% was assumed to indicate malicious input).

2.3 Results

After banning spammers and filtering malicious translations, the number of trusted annotations varied significantly among languages. This difference occurred due to the difference in the workflow rate among the languages from the respective Crowdflower-supported crowd channels, as well as the availability of workers for specific languages in the crowd. Figure 1 shows the number of trusted worker judgements collected for every language for the training and testing, as well as the tuning datasets.

We find that we were able to create parallel data totalling around 1 million segments, or 10 million words, for a crowdsourcing budget of approximately 45,000 EUR. We estimate that the creation of the corpus would have costed well over 10 times as much as using professional translators (assuming a cost of 0.05-0.07 EUR per word), and well over 40 times as much as subcontracting the work to a Language Service Provider that did not use a machine translation post-editing workflow (assuming average pricing of 0.21 EUR/word). This calculation does not consider overhead costs, which exist in different forms in all scenarios. We argue that some of the disadvantages of using crowdsourcing for translation, such as lower quality expectations, and the lack of guarantee that the full

[3] https://iversity.org/
[4] http://alt.qcri.org/resources/qedcorpus
[5] http://videolectures.net/

text will be translated within a specific time frame, are acceptable for our use case of creating domain-specific training data for a machine translation system, and that crowdsourcing is a cost-effective method for this purpose.

The corpus will be made available through the EU (according to the H2020 Open Research Data Pilot) for research purposes after the end of the project, and taking into account copyright restrictions imposed by each source.

3. The Impact on Machine Translation Performance

Our main success criteria of the crowdsourcing effort is whether the collected data can effectively improve the translation quality of a machine translation system in the target domain. In this section, we describe the translation systems built within the TraMOOC project and experimentally validate the effectiveness of the crowdsourcing effort by measuring its effect on the translation quality of machine translation systems.

3.1 Methodology

Our translation systems are built in two steps: for each language pair, we first train a baseline system on a large "mixed-domain" dataset, using data from various sources. Then, we fine-tune this system on an "in-domain" dataset representative of MOOC materials. This domain adaptation step is performed either using only pre-existing in-domain data, or using both the pre-existing data and the crowdsourced data that we collected.

We report performance of three systems for each language pair:

- a mixed-domain baseline system;
- a system adapted towards pre-existing in-domain data;
- a system adapted towards both existing and crowdsourced in-domain data.

We measure and report the translation quality of these systems with the BLEU metric. Evaluation results are computed on held-out test sets that were also created via crowdsourced translations. For Chinese, we compute character-level BLEU. For all other language pairs, the evaluation was de-tokenized and case-sensitive.

3.2 Training Data

For training, we collected bilingual corpora for use in our baseline systems. The following data sets are considered mixed-domain training data:

- Europarl (Koehn, 2005);
- JRC-Acquis 3.0 (Steinberger et al., 2006);
- DGT's Translation Memory (Steinberger et al., 2012) as distributed in OPUS (Tiedemann, 2012);
- OPUS European Central Bank (ECB);
- OPUS European Medicines Agency (EMEA);
- OPUS EU Bookshop;
- OPUS OpenSubtitles 7;
- WMT News Commentary;
- WMT CommonCrawl;
- Chinese WMT training data;
- Wikipedia names and titles (English-Russian);
- SETimes (Tyers and Alperen, 2010);
- Yandex English-Russian Parallel Corpus[6];
- The United Nations Parallel Corpus v1.0 (English-Chinese) (Ziemski et al., 2016);
- CzEng v1.6pre 8;
- Croatian-English parallel corpus hrenWaC 2.0 (Ljubešic et al., 2016).

We consider the following data sets to be pre-existing in-domain data sets for the purpose of MOOC translation:

- TED from WIT3 (Cettolo et al., 2012);
- QCRI Educational Domain Corpus (QED) (Abdelali et al., 2014);
- Parallel data provided by Coursera;
- Web-crawled data collected in the TraMOOC project.

The amount of pre-existing in-domain data differs greatly between languages, ranging from tens of thousands to millions segment pairs, while the crowdsourced data that we collected is in the order of the tens of thousands (Table 2). The amount of mixed-domain training data ranges between 20 and 60 million segment pairs per language, 100-1000 times the amount of crowdsourced in-domain training data.

	Pre-existing	Crowdsourced
en-bg	63000	54000
en-cs	2177000	46000
en-de	258000	48000
en-el	124000	66000
en-hr	89000	80000
en-it	336000	86000
en-nl	226000	34000
en-pl	246000	59000
en-pt	575000	74000
en-ru	2301000	69000
en-zh	647000	18000

Table 2: Amount of in-domain parallel training data (segment pairs) per language pair for domain adaptation.

3.3 Machine Translation Systems

Our baseline translation systems are GRU attentive sequence-to-sequence neural machine translation models (Bahdanau, 2015).

For training, we used the same configuration as the Edinburgh's submission to the WMT-17 news translation task (Sennrich et al., 2017), which provides a strong baseline.

[6] https://translate.yandex.ru/corpus

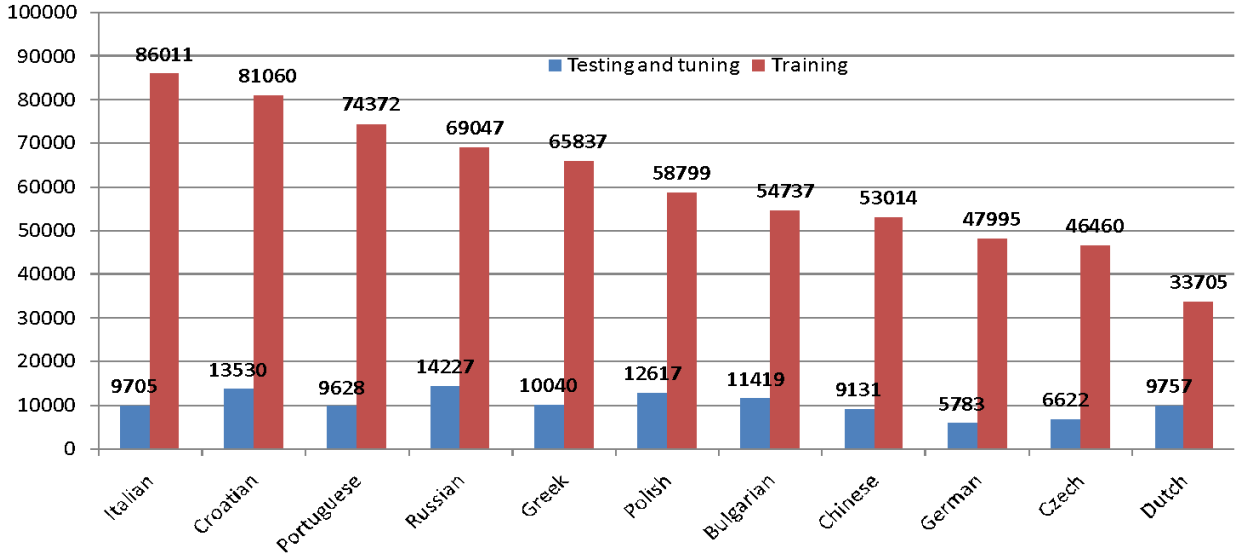

Figure 1: Number of trusted segments collected for each target language for training and testing/tuning.

We adapt the baseline systems to the in-domain MOOC data using continued training of the baseline system with MAP-L2 and dropout regularization (Miceli Barone et al. 2017).

3.4 Results

Evaluation results are shown in Table 3. We can see that domain adaptation via fine-tuning is effective for all language pairs.

	Baseline	+ Preexisting	+ Crowdsourced
en-bg	22.91	23.57	25.89
en-cs	29.86	31.06	32.06
en-de	29.29	32.14	33.69
en-el	35.54	38.01	40.76
en-hr	23.36	23.70	26.43
en-it	32.15	36.19	38.53
en-nl	35.59	38.04	40.07
en-pl	27.16	28.41	30.97
en-pt	39.44	47.68	48.71
en-ru	26.41	29.08	29.78
en-zh	27.93	28.51	29.77
avg	29.97	32.40	34.24

Table 3: Translation quality (BLEU) of baseline system, and systems adapted to domain with/without crowdsourced data.

Using only pre-existing in-domain data, we obtain improvements ranging from 0.34 BLEU (en-hr) to 8.24 BLEU (en-pt), and an average improvement of 2.43 BLEU.

As our main result we report the effect of the crowdsourced in-domain data on translation quality: we find consistent and strong improvements, ranging from 0.7 BLEU (en-ru) to 2.75 BLEU (en-el) with an average improvement of 1.84 BLEU over the systems with only pre-existing in-domain data, and an average 4.27 BLEU over the mixed-domain baselines.

There are various factors which affect the effectiveness of domain adaptation with crowdsourced in-domain data. Regarding the correlation between the amount of in-domain data for fine-tuning and translation quality, we note that Miceli Barone et al. (2017) found an approximately log-linear combination between the two, using random subsets of in-domain data of different size. In our experiments, this relationship is confounded by the fact that the baseline models are of varying quality, and that we have access to varying amounts of pre-existing data that we treat as "in-domain", with varying distance to our actual target domain.

Despite all confounding variables, if we consider that the amount of crowdsourced training data is 100-1,000 times smaller than the amount of out-of-domain training data used, and consistently smaller than the amount of pre-existing in-domain data, we conclude that the improvements that we observe from adding crowdsourced data cannot just be attributed to having more training data available. Based on the log-linear learning curves reported in related work (Koehn and Knowles, 2017; Miceli Barone et al. 2017), we would expect small or negligible improvements in translation quality if we added the same amount of out-of-domain training data, or pre-existing in-domain data, to the systems without crowdsourced data. This confirms the relevance of obtaining in-domain training data that is similar in terms of domain and genre

to the texts that are to be translated, and that the crowdsourced training data is of high value to the MT system.

4. Conclusion

We collected crowdsourced translations from English to eleven languages to create a parallel corpus for the educational domain.

We experimentally showed that using this data to train neural machine translation systems by means of domain adaptation provides large quality improvements, even on top of systems adapted using only existing in-domain translations. These results highlight the importance of in-domain training data for machine translation: even a small amount of crowdsourced translations, that may be noisy in nature, has a large positive impact on translation quality.

In conclusion, we show that crowdsourcing with proper quality controls is a viable and cost-effective way of creating valuable in-domain parallel resources for machine translation.

Acknowledgements

This work was done as part of the TraMOOC project (Translation for Massive Open Online Courses) funded by the European Commission under H2020-ICT-2014/H2020-ICT-2014-1 under grant agreement number 644333. This work was supported by grant EP/L01503X/1 for the University of Edinburgh School of Informatics Centre for Doctoral Training in Pervasive Parallelism from the UK Engineering and Physical Sciences Research Council (EPSRC).

References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Ambati, V. (2012). Active learning and crowd-sourcing for machine translation. PhD Thesis. Carnegie Mellon University. ISBN: 978-1-267-58215-7.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. Proceedings of the International Conference on Learning Representations (ICLR).

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Association for Computational Linguistics, pages 286-295.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT), pages 261-268, Trento, Italy.

DeBoer, J, Stump, G., Breslow, L, and Seaton, D. (2013). Diversity in MOOC Students' Backgrounds and Behaviors in Relationship to Performance in 6.002x. Proceedings of the sixth learning international networks consortium conference.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In MT Summit X, Phuket, Thailand.

Koehn, P., Knowles, R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, pages 28-39. Association for Computational Linguistics.

Kordoni, V., van den Bosch, A., Kermanidis K., Sosoni V., Cholakov, K., Hendrickx, I., Huck, M., and Way, A. (2016). Enhancing Access to Online Education: Quality Machine Translation of MOOC Content. Proceedings of the International Conference on Language Resources and Evaluation, pages 16-22, Portoroz, Slovenia.

Ljubešic, N., Esplà-Gomis, M., Ortiz Rojas, S., Klubíčka, F., and Toral, A. (2016). Croatian-English parallel corpus hrenWaC 2.0. Slovenian language resource repository CLARIN.SI.

Miceli Barone, A. V., Haddow, B., Germann, U., and Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark.

Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Barone, A. V. M., and Williams, P. (2017). The University of Edinburgh's Neural MT Systems for WMT17. Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Copenhagen, Denmark.

Sosoni, V., Kermanidis, K.L., Stasimioti, M., Naskos, T., Takoulidou, E., van Zaanen, M., Castilho, S., Georgakopoulou, P., Kordoni, V., and Egg, M. 2018. Translation Crowdsourcing: Creating a Multilingual Corpus of Online Educational Content. In Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), Miyazaki, Japan. (to appear).

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. In Proc. of the Int. Conf. on Language Resources and Evaluation (LREC) , pages 454-459, Istanbul, Turkey.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tus, D., and Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In Proc. of the Int. Conf. on Language Resources and Evaluation (LREC), pages 2142-2147, Genoa, Italy.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012).

Tyers, F. M. and Alperen, M. S. (2010). South-East European Times: A parallel corpus of Balkan languages. In Proc. of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages , pages 49-53, Malta.

Zaidan, O. F., and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1220-1229. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations Parallel Corpus v1. 0. In Language Resources and Evaluation (LREC 16). Portorož, Slovenia, May 2016.