Translation Crowdsourcing:
Creating a Multilingual Corpus of Online Educational Content

Vilelmini Sosoni[1], Katia Lida Kermanidis[2], Maria Stasimioti[1], Thanasis Naskos[2], Eirini Takoulidou[2], Menno van Zaanen[3], Sheila Castilho[4], Panayota Georgakopoulou[5], Valia Kordoni[6], Markus Egg[6]

[1]Department of Foreign Languages, Translation and Interpreting, Ionian University, Corfu, Greece
[2]Department of Informatics, Ionian University, Corfu, Greece
[3]Department of Communication and Information Sciences,Tilburg University, The Netherlands
[4]ADAPT Centre, Dublin City University, Dublin, Ireland
[5]Deluxe Media Europe, Deluxe Entertainment Services Group, London, UK
[6]Department of English Studies, Humboldt-Universität zu Berlin, Germany
vilelmini@hotmail.com; kerman@ionio.gr; stasimioti.maria@gmail.com; anaskos@ionio.gr; rinoulit@gmail.com; mvzaanen@uvt.nl; sheila.castilho@adaptcentre.ie; yota.georgakopoulou@bydeluxe.com; kordonie@anglistik.hu-berlin.de; markus.egg@anglistik.hu-berlin.de

Abstract

The present work describes a multilingual corpus of online content in the educational domain, i.e. Massive Open Online Course material, ranging from course forum text to subtitles of online video lectures, that has been developed via large-scale crowdsourcing. The English source text is manually translated into 11 European and BRIC languages using the CrowdFlower platform. During the process several challenges arose which mainly involved the in-domain text genre, the large text volume, the idiosyncrasies of each target language, the limitations of the crowdsourcing platform, as well as the quality assurance and workflow issues of the crowdsourcing process. The corpus constitutes a product of the EU-funded TraMOOC project and is utilised in the project in order to train, tune and test machine translation engines.

Keywords: parallel corpus, MOOCs, online educational text, crowdsourcing

1. Introduction

Massive Open Online Courses (MOOCs) have been growing rapidly in number of enrollees and participating universities, in impact, and in target groups. According to 2016[1] statistics, they involve around 60 million students worldwide. In their majority, MOOC users are non-native English language speakers, and the language barrier proves to be the main obstacle towards further expansion of the MOOC market, as most courses are offered in English only.

The EU-funded TraMOOC (Translation for Massive Open Online Courses) project aims at providing machine translation solutions for the educational content available in MOOCs, thus enhancing access to online education. The content genre varies from video lecture subtitles, slides, assignments, and quiz text, to course discussion forum text (Kordoni et al., 2016). Given that the source language is English, in-domain trained and tested translation engines are built for 11 European and BRIC target languages (Bulgarian, Chinese, Croatian, Czech, Dutch, German, Greek, Italian, Polish. Portuguese, and Russian).

In an attempt to achieve optimal translation output, the TraMOOC goal is to develop as much in-domain data as possible, especially for the language pairs that are not adequately supported by the required infrastructure. To this end, crowdsourcing is adopted (as an alternative to using professional translators) for collecting translations on a large scale for all language pairs involved.

The present work describes the in-domain source data collected, the crowdsourcing experiment and the resulting multilingual parallel dataset, taking into consideration the challenges imposed by the text genre, the number of language pairs involved, the large data volume aimed at, the quality assurance of the experimental process, and the related crowdsourcing workflow issues.

2. Related Work

The creation of parallel data on a large scale for new language-pairs requires intensive human effort and availability of experts. For most language-pairs, the small number of expert translators available or the lack of access to fluent bilingual speakers makes it difficult and expensive to create parallel corpora for training machine translation systems. Recent research has looked at obtaining translations via crowdsourcing, in particular for low resource languages (Ambati & Vogel, 2010; ; Zaidan & Callison-Burch, 2011; Post, Callison-Burch, Osborne, 2012). Crowdsourcing as an approach to activate or use the knowledge and skills of a large group of people in order to solve problems has existed for a long time (cf. Ellis, 2014). Nowadays, it leverages Web 2.0 tools (O'Reilly, 2007) in order to take a job normally performed by a designated person and having it done by a large, undefined, and dispersed number of participants (Howe, 2008). In the area of translation, crowdsourcing has actually been used widely in the past years for the translation of online content. As Jiménez-Crespo (2017) observes, Facebook has used it for the translation of its social networking site and its user interfaces, Amara and TED for audiovisual practices, Kiva and the platform Trommons from the Rosetta Foundation for non-profit initiatives.

For the generation of parallel corpora, the most widely used crowdsourcing platform is MTurk[2] (Ambati & Vogel, 2010; Zaidan & Callison-Burch, 2011; Post, Callison-Burch, Osborne, 2012), although Negri &

[1] https://www.class-central.com/report/mooc-stats-2016/

[2] https://www.mturk.com/

Mehdad (2010) used CrowdFlower[3] for the creation of a bilingual Textual Entailment corpus for the English/Spanish language pair taking advantage of an already existing monolingual English RTE corpus.

Although the benefits from collecting corpora using crowdsourcing techniques are numerous, gathering data is cheap, quick to acquire, and varied in nature, it does not go without carrying risks such as quality control and workers that try to "game" the system.

Taking the pros and cons into account, the present work extends previous work by using the CrowdFlower platform for the translation of both formal and informal English sentences of educational content to both low- and high-resource languages (11 European and BRIC target languages) and by applying various quality measures and features.

3. Source Data

The source data are comprised of online educational course material in English. More specifically, they contain lecture subtitles and quiz assessment text (considered henceforth *formal* text), and course forum discussion text that students share among themselves and/or the instructor for posting questions, clarifications, opinions etc. (considered henceforth *informal* text). The topics of the courses varied from technical (e.g. Finance) to humanities (e.g. Future of Storytelling).

3.1 Data Sources

The English source text comes from several different channels. Henceforth, a 'segment' is a piece of text between two consecutive CR/LF characters.

Iversity.org. A large part of the dataset (~35,000 segments), formal and informal, originated from the MOOC provider Iversity.org.

Videolectures.NET. Videolectures.NET is a library of online educational video lectures. 800 segments of lecture subtitles (formal text) of the 'Complexity Science' course were included in our dataset.

Coursera. 27,000 segments originated from subtitles of online courses provided by Coursera (formal text). The number of courses exceeded 280, and varied between 'Web Applications', 'Public Policy', and 'Art History'.

QED. 28,000 segments were transcripts of video lectures (formal text) selected from the QCRI Educational Domain Corpus (QED)[4].

3.2 Data Description

Regarding the formal text, several properties rendered its processing quite challenging. On the one hand, it exhibited a high frequency of domain specific terms and expressions, named entities, scientific formulas, and words unknown to crowdworkers, as well as to any system for posterior processing. On the other hand, the subtitle genre contained spontaneous speech properties, truncated sentences, elliptical formations, disfluencies, repetitions, interjections and fillers.

Example 1: *What? What? He's going to score, he's in! Whoops. And I've got, ooh, here we go.*

Example 2: *Sharing Economy is the way of build resources and get to go ahead on the path, where we got new creation, trust, together and we feel better for the responsibilities.*

Example 3: *Hello, Imstuding fashion design and my aim it's to become a sustainable and ethical fashion designer.*

Regarding the informal text, it presented all the properties of social media text: slang, misspellings ('supa' instead of 'super'), lexical variants, abbreviations, acronyms, multilingual tokens, unorthodox syntax structures, disfluencies, awkward word choices, and repetitions.

Example 4: *Puuurrrfect!*

Example 5: *truthfully , i have no idea how much i reason per day day.*

Course title	No of segments	
	Formal	Informal
Business Analysis	378	5148
Contemporary Architecture	258	5219
Crystals and Symmetry	144	3338
Dark Matter	379	4443
Gamification Design	319	9121
Public Speaking	113	-
Web Design	270	1523
Critical Thinking	550	500
Social Innovation	550	500
Monte Carlo Methods in Finance	700	500
Modeling and Simulation using Matlab	150	440
Future of Storytelling	550	460
Total	4361	31192

Table 1: Segment size per course in the Iversity.org data.

3.3 Data Preparation

In order for the data to be appropriately formatted for further processing, it had to undergo a preparation phase, which included:

Conversion into plain text. The text is freed from all markup and meta information, and converted into plain UTF-8 text format, using Python and UNIX-based shell scripts. Special characters were removed, along with non-

[3] https://www.crowdflower.com/

[4] http://alt.qcri.org/resources/qedcorpus

content lines, and multiple or trailing whitespace characters.

Tokenization, and sentence segmentation. The plain text data was tokenized into words and punctuation. Correct segmentation involved the removal of incomplete segments, and/or of segments that contained multiple sentences.

Example 6: *and often*

Example 7: *say the ... in the case of the peacock's tail , say,*

Also, due to the subtitle transcription process, which was undertaken to a large extent by non-experts, and due to the presence of spoken language characteristics in the subtitle text (incomplete/unfinished sentences, intersentential change of topic), very often a segment includes the ending part of the previous sentence and the starting part of the next, as shown in the following example.

Example 8: *this is his name. Five years ago, whenever I took over as*

Scripts were built to deal with incorrect segmentation, although they were not completely error-free, because automatic deep understanding of the text was essential, but not feasible. The toolset is available online[5].

Markup of special elements. Some textual elements, such as URLs and emoticons, are automatically replaced with special tags. This would ensure that translators would not try to translate these elements, whilst this abstraction is also considered beneficial for the MT systems, as the latter would be correctly trained to leave them untranslated.

Data setup for the translation crowdsourcing activity. A set of 5000 segments was selected to constitute the tuning and testing set for the upcoming machine translation experiments. These segments were taken from the pool of Iversity.org, and the entire set of Videolectures.NET segments, and were translated by at least two and at most three workers per target language for redundancy purposes. The rest of the segments were translated once per target language.

4. Crowdsourcing Experiments

A total of 2050 workers participated in the crowdsourcing experiments, which took place from March to June 2017. They were given clear and detailed instructions, both general and target-language specific, on how to translate the English segments presented to them. Instructions contained specific examples to help workers deal with typical linguistic (e.g. the translation of acronyms, proper names etc.) or formatting (e.g. dealing with punctuation issues) challenges appearing in the text. Segments were presented in a stand-alone, out-of-context manner, but workers were provided with the course title, when available, in order to get a better understanding of the context. No terminological lexicon was provided to them, as this was not feasible due to the vast number of diverse topics. In order to keep the crowdsourcing task as simple as possible, no further annotation was required by the workers, e.g. pertaining to the grammaticality of the input segment.

Quality assurance was supported by a test mode, where workers were asked to answer a set of test questions, i.e., choose the best translation for a source segment among three candidates. Thereby, the workers’ accuracy level was determined. The test mode occurred before the first translation questions, but also during the entire process task for continuous monitoring of the workers’ quality. Once the accuracy level of a worker dropped below a certain threshold, the workers could not continue working on the task. Additionally, the worker was blocked to continue working on other crowdsourcing experiments in the TraMOOC project.

Performing accuracy measurements this way is a trade-off. Using question types that are different from those in the task allows workers to easily identify the test questions. Once these are answered correctly, the actual translations can, for example, be generated randomly. The results showed that some workers used hill-climbing with different accounts to collect the correct answers. The entire process of filling in the data can then be automated, resulting, obviously, in unusable translations.

To make sure that only correct translations were collected, close and constant, albeit to a large extent automated, monitoring of the workers’ input proved to be the most important element for ensuring quality annotations, banning spammers and removing worthless input.

5. Multilingual Corpus of Online Educational Content

The outcome of the crowdsourcing activity we launched on CrowdFlower was the collection of a large dataset of parallel corpora in the 11 languages of the project, although the number of segments varied per language (Figure 1). Participants who followed the instructions and showed competent behavior (no blank inputs, no random translations, accuracy above 80% threshold) were labeled “trusted” and their translations were accepted as such. As shown in Figure 1 below, the Italian and Croatian languages collected over 90,000 ‘trusted’ translated segments. The Italian language collected the smallest amount of untrusted judgements (less than 2000 judgments) and the vast majority of Italian translated segments came from contributors located in Italy (~90% of the participants). Regarding the Croatian language, the large amount of collected translations is considered a significant pillar of linguistic infrastructure, taking into consideration that there are no sufficiently large parallel corpora available, and that it is the most weakly NLP-resource-supported language in the project. Due to malicious behavior of the contributors, the Croatian

[5] https://github.com/mixstef/tramooc

language collected over 10,000 untrusted judgments. Another interesting result is that half of the contributors were from Serbia followed by Bosnia and Herzegovina in second place. Datasets for Russian and Portuguese followed suit with more than 80,000 translated segments. These languages collected ~8000 and ~6000 untrusted judgements respectively. In the Portuguese crowdsourcing task only crowd workers who were located in Portugal (40%) and Brazil (60%) took part, while in the Russian task there were mostly contributors located in the Russian Federation (~75%). Greek and Polish crowdsourced translations reached over 70,000 segments, which is a remarkable amount for such low-resource languages. The amount of untrusted judgements was relatively small for both languages (less than 5000 and 7000 respectively). In the first case, the majority of the contributors were from Greece (90%) and in the second case from Poland (85%). Bulgarian and Chinese tasks gathered ~60,000 segments. Regarding the Bulgarian language, the amount of collected data is quite satisfactory, taking into consideration that it is a low-resource language. The percentage of untrusted judgements was 10% and the contributors were from Bulgaria (~70%), FYROM (~12%) and the Russian Federation (~8%). The untrusted judgements for the Chinese language were around 7000. Most of the contributors were located in Hong Kong, followed by China, Malaysia and Taiwan. German and Czech translated segments were above 50,000. The untrusted judgments were around 7000 and 5000 respectively. In the first case, the contributors were from Germany (~80%), followed by contributors from Austria (~10%). In the second case, the contributors were from the Russian Federation, Poland, Czech Republic and Venezuela. For the Dutch language, however, we managed to collect only ~40,000 translations (less than half of Italian task), with ~1000 untrusted judgements. The contributors were mainly from the Netherlands (~50%) and Belgium (~40%).

Examples of trusted and untrusted segments are provided:

Trusted segments:
EN: *An interface is simply the point where two entities meet.*
EL: *Μια διεπαφή είναι απλώς το σημείο όπου δυο οντότητες συναντιούνται*

EN: *And I get C d theta dt.*
DE: *Und ich bekomme C d theta dt.*

Untrusted segments:
EN: *What Gamification, Game Thinking and Games are*
EL: *What Gamification, σκέψης παιχνίδι και τα παιχνίδια είναι*

EN: *Nope, I am good*
DE: *Amet, quidem sit accusamus et eveniet, repudiandae culpa, quam aut*

The provision of partly translated sentences (example of the EN-EL language pair), Latin text (example of the EN-DE language pair) and blank translations were common practices among "untrusted" contributors and such cases were spotted in all language pairs.

The language-specific workflow data serves as a means to assess the size of the crowd channels linked to the particular crowdsourcing platform. It turns out that, while certain languages are satisfactorily supported by crowd channels of workers that speak them, others are not. Furthermore, channel support is not aligned with Natural Language Processing (NLP) resource support. Languages satisfactorily equipped with NLP tools are not necessarily supported by large crowdsourcing channels (e.g. Dutch), and vice versa.

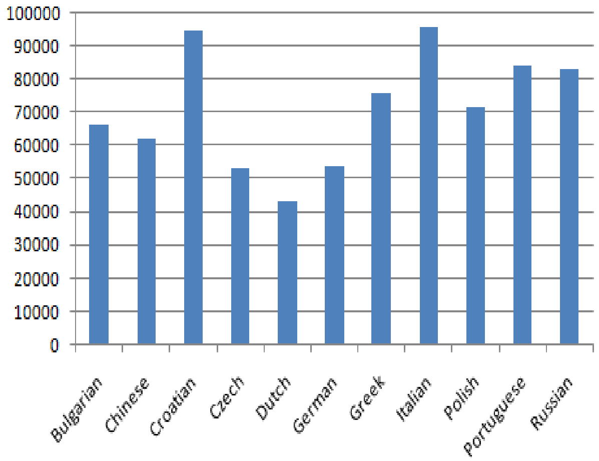

Figure 1: Number of trusted translated segments per target language.

As mentioned earlier, the primary use of the collected data is for developing translation engines for online course material. To this end, it was deemed particularly important to gather the data in the strongly supervised manner described earlier, so as to reach a satisfactory quality level that would prove beneficial for the MT systems (Behnke et al., 2018).

6. Conclusion

The multilingual corpus of online course material described in this article was developed manually in 11 languages via crowdsourcing. The challenges encountered due to the genre of the video lecture transcripts (i.e. spontaneous speech) and the social media forum text, as well as crowdsourcing workflow issues for some of the languages are presented. Language-specific workflow phenomena serve as an indication of the size of the crowd channels supported by the platform for every language.

Close monitoring of the crowdsourcing process proved to be key in addressing the aforementioned challenges, and ensuring the required quality threshold of the provided annotations. The process led to a satisfactory percentage of trusted judgements, resulting in a large-scale multilingual corpus of online course material. 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.

Acknowledgements

This work was performed 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.

References

Ambati, V., & Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems?. In Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk (pp. 62-65). Association for Computational Linguistics.

Behnke et al. (2018). Improving Machine Translation of Educational Content via Crowdsourcing. Proceedings of the International Conference on Language Resources and Evaluation, Miyazaki, Japan (to appear).

Ellis, S. (2014). A History of Collaboration, a Future in Crowdsourcing: Positive Impacts of Cooperation on British Librarianship. Libri 64 (1), pages 1-10.

Jiménez-Crespo, M. A. (2017) Crowdsourcing and Online Collaborative Translations: Expanding the limits of Translation Studies. Amsterdam/Philadelphia: John Benjamins.

Kordoni et al. (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.

Negri, M., & Mehdad, Y. (2010). Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (pp. 212-216). Association for Computational Linguistics.

O'Reilly, T. (2007) What is Web 2.0: Design patterns and business models for the next generation of software. International Journal of Digital Economics, 65, pages 17-37.

Post, M., Callison-Burch, C., Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 401-409). Association for Computational Linguistics.

Zaidan, O. F., & Callison-Burch, C. (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 (pp. 1220-1229). Association for Computational Linguistics.