Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour

Rafał Jaworski[1], Sanja Seljan[2] and Ivan Dunđer[2]

[1]Adam Mickiewicz University in Poznań, Poland
rjawor@amu.edu.pl

[2]University of Zagreb, Croatia
{idundjer}@ffzg.hr

Abstract

This paper presents an idea to bring crowdsourcing to a higher level, for the purpose of acquiring valuable machine translation and natural language processing resources. In the proposed scenario, students are being educated in order to improve the quality and effectiveness of their natural language processing (NLP) related work. Their motivation is ensured by introducing an element of gamification – a ranking is kept, where the best contributing users are decorated with medals. The ranking is available at all times to all users and is always up-to-date, hence the effects of the contributions are immediately visible to the users. This scenario was applied to a group of students enrolled in Natural Language Processing course, who were presented with a task of collecting parallel corpora for less-resourced language pairs, in this case Croatian-English and English-Croatian. The whole experiment was supervised with the help of a custom-made open-source system named *TMrepository*, developed and maintained by the authors of this paper.

Keywords: crowdsourcing, gamification, NLP, machine translation resources, parallel corpora, sentence alignment, less-resourced languages, Croatian, TMrepository

1. Introduction

Acquiring corpora is an essential task when it comes to machine translation or other aspects of natural language processing. However, due to the lack of available digitised textual corpora the preparation of crucial language resources becomes very challenging, especially for low-resourced languages, such as Croatian.

Globalisation and the increase in user-generated content have boosted the demand for translation. At the same time, the development of computer-aided translation tools has enabled virtual collaboration, which is evident in all translation scenarios and across the whole translation process – from authors to publishers, from translation agencies to translators (O'Brien, 2011). Collaborative translation essentially means that two or more agents, i.e. translators, cooperate in some way to produce a translation, with the possible inclusion of machine translation and computer-assisted translation.

Gamification can be defined as the application of game design principles to change behaviour in non-gaming context (Robson et al., 2016). If implemented and applied properly, gamification can increase engagement of all participants and the interactivity of elements in a specific environment.

Gamification primarily aims at increasing the positive motivations of users towards specific activities or the use of technology, and thereby, increasing the quantity and quality of the output or outcome of the given activities (Morschheuser et al., 2017).

However, applying gamification is not straightforward, as the source of idea, i.e. games, are complex and therefore difficult to incorporate in other environments. Gamification also involves understanding and deeper knowledge of psychology. Furthermore, gamification is commonly also to affect users' behaviour which increases the complexity of application and design of gamification (Morschheuser et al., 2017).

Crowdsourcing refers to recruiting an undefined, i.e. very often unknown, large group of people by making an open call, to handle a specific task which would otherwise be assigned to in-house employees (O'Brien, 2011).

The basic idea of crowdsourcing derives from the fact that labour of a talented crowd isn't always free or cheap, but it costs less than paying traditional in-house employees (Howe, 2006).

Crowdsourcing allows the power of the crowd to accomplish tasks that were once the province of just a specialised (Howe, 2008).

The paper presented by (Quinn and Bederson, 2011) points out several dimensions of crowdsourcing: motivation, quality, aggregation, human skills, participation time and cognitive load. The paper categorises crowdsourcing into seven genres: Game with a purpose, Mechanised labour with payment, Wisdom of crowds with hundreds of people participating and thinking independently, Crowdsourcing by the unpaid general public usually motivated by curiosity, Dual-purpose work in order to perform a task which can't be performed automatically (e.g. transcribing old scanned books which cannot be processed by OCR), Grand search with the task of finding the solution for a specific problem, and Knowledge collection from volunteer contributors based on the idea to create large databases of common facts.

2. Related work

The following subsections discuss related work associated with gamification and crowdsourcing in natural language processing

2.1. Gamification

Gamification has been successfully applied for the purpose of machine translation evaluation. For instance, in one research evaluators were required to provide a quality score, and in return they received feedback in form of stars that reflect how close their very score is to a gold-standard score (Abdelali et al., 2016). The intention of such a simple gamification strategy was to keep the evaluators engaged.

2.2. Crowdsourcing in NLP

The paper by (Sabou et al., 2014) presents a research on crowdsourcing used for the acquisition of annotated corpora, which are crucial for experimenting with natural language processing algorithms. In the paper, crowdsourcing is approached as a project-based task divided into specific phases.

Another paper shows that crowdsourcing is used for the collection of training data and to perform annotation on clustering and parsing, backed with a statistical natural language processing analysis (Li et al., 2015). However, the authors point out that crowdsourcing in science is still not a recognised research method.

Crowdsourcing can also be used for the collection of unstructured documents and reports in big open data (Munro et al., 2012). In the research, several methods have been applied, such as search-log-based detection, machine learning techniques and corrected crowdsourced annotation with the ultimate goal to detect seasonal epidemics on a global scale.

Crowdsourcing is nowadays being used in order to satisfy the translation demand, as collaborating on translation projects clearly has benefits (O'Brien, 2011).

Another paper also discusses the potential of collaborative work for natural language processing tasks (Sabou et al., 2012).

A collaborative three-phase workflow for language translation as a cost-effective crowdsourcing approach was also proposed (Ambati et al., 2012). The given workflow breaks an atomic task of sentence translation into separate phases: word translation, assisted sentence translation and synthesis of translations. Such sub-tasks are at all times consistent and verifiable, and enable better translations for a vast range of diverse non-expert translators.

Another research confirmed that crowdsourcing can play a very important role in building necessary language resources (Zaidan and Callison-Burch, 2011). By using Amazon's Mechanical Turk, it is indeed possible to obtain high-quality translations from non-professional translators while keeping the overall costs below the price of professional translations.

A similar approach shed some more light on the obstacles and challenges when formulating crowdsourcing translation tasks on Amazon's Mechanical Turk for the purpose of machine translation (Ambati and Vogel, 2010). One of the important conclusions was that when working with non-professional translators it is very important to address quality concerns alongside checking for any usage of automatic translation services.

One paper considers crowdsourcing translation to be the next big breakthrough in the translation industry (Muntés-Mulero et al., 2012). Interactive, elastic and scalable machine translation will be important to assist the crowd, whereas by harvesting the crowd potential to deal with tasks such as translating or post-editing, the translation environment becomes capable of increasing its capacity to quickly deal with large amounts of data and information.

Approaches to teaching computer-assisted translation in the 21st century are tightly connected with the rise of emerging translation standards and fast-growing technologies in the translation industry (Muegge, 2013). Nowadays, teaching computer-assisted translation heavily relies on collaborative translation, machine translation, translation management systems, and especially crowdsourcing. Taking advantage of a cloud-based and open-source learning management system in combination with the crowdsourcing principle, for teaching purposes is proposed. Such an online learning system should support students with any internet-enabled computer device, give them access to translation technology from anywhere and, at the same time, reduce teaching costs. It should also enable students to participate in real-time collaborative work and editing exercises (e.g. NLP tasks, post-editing of machine translations etc.), allow them to submit translation-specific assignments and to track their progress.

Distance and blended learning are also vital features of the training and teaching environment for NLP-related tasks (Canovas and Samson, 2011). Classroom facilities also dictate and limit teaching approaches, therefore, allowance must be made for hardware diversity. Other aspects such as application of different operating systems, available computer programs etc. influence what is perceived the most suitable for specific training objectives.

3. The *TMrepository* system as a crowdsourcing platform

TMrepository is a custom-made web application created for the purpose of this research. It is being constantly developed and maintained by the authors of this paper, and it is publicly available at: http://concordia.vm.wmi.amu.edu.pl/tmrepository/

Its main goal is to provide an easy-to-use repository of translation memories, aligned at the sentence level. In this particular research it was used to store and handle Croatian-English and English-Croatian translation memories. *TMrepository* also features gamification-inspired functionalities in order to boost the motivation of contributors.

TMrepository is intended to be used in the future by computer science and information science students that work on machine translations and other NLP-related tasks, or students of translation studies, who produce their own translation memories during their education process.

Registration to the system is free and open. Upon login, the user is presented with the list of their own contributions.

The main activity of the user in the system is uploading new resources. This is done by the means of the upload form. In the form, the user is asked to provide the title of the translation memory, its brief description and type of resource. Available types are the following:

- manual translation
- manual translation - automatically aligned

- corpus - automatically aligned
- renowned corpus

Translation memories of the type "manual translation" are produced by human translators during their work with a Computer-Aided Translation (CAT) tool. They translate documents in a sentence-by-sentence mode, therefore the resulting TM is already perfectly aligned at the sentence level and the quality of the translations is assumed to be high. Another option is "manual translation - automatically aligned". It is reserved for translation memories which were produced by human translators but were not aligned at the sentence level, e.g. a collection of MS Word documents, where source and target texts are in separate documents. Such documents must be split into sentences and aligned. While this can be done in external software, the *TMrepository* provides a feature of automatic splitting and aligning of a pair of MS Word documents. However, as this process is done automatically, it may not yield perfect alignments (extensive research on the impact of the automatic splitting and aligning quality is planned for future research). Therefore, the translation memory type "manual translation - automatically aligned" has slightly lower expected quality than the "manual translation".

Another type of translation memories in the system is "corpus - automatically aligned". These translation memories are collected automatically from various sources, predominantly from multilingual websites. The collection process includes automatic scraping of source and target texts, splitting into sentences and aligning. As in this case the text scraping is another process done automatically, the quality of resulting translation memories is expected to be lower than "manual translation - automatically aligned".

The last option, "renowned corpus", refers to already existing and aligned corpora, whose quality is expected to be high and which do not need quality checking, e.g. Europarl (Koehn, 2005). As one of the very important goals of *TMrepository* is to concentrate in one system all available resources for a specific language pair, the authors decided to allow users to upload already well-known and precompiled high standard corpora.

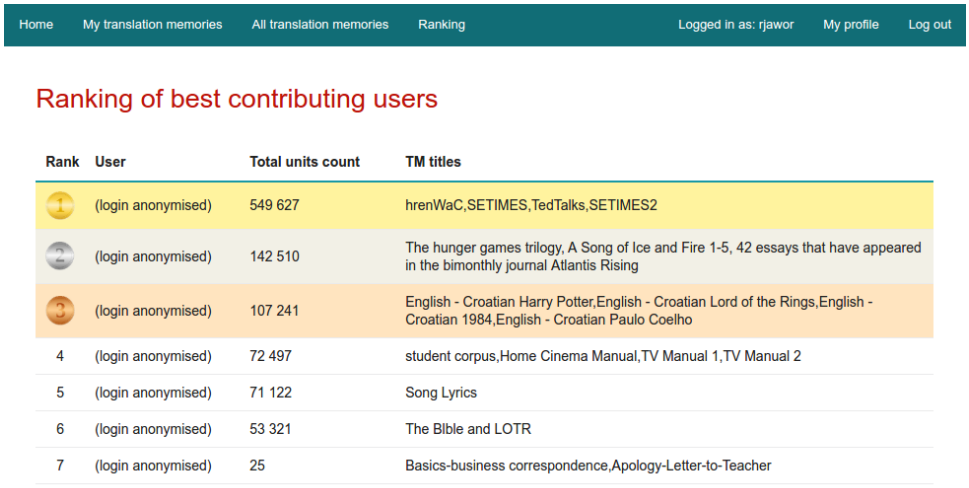
This study focuses on collecting translation memories of the type "corpus - automatically aligned", understood as resources automatically aligned by the contributing user, consisting of translations performed by other people.

Available import formats include: a pair of text files, a TMX file and a pair of Word documents. Import from TXT files assumes that two text files with UTF-8 encoding, having equal number of lines are provided. This form reduces various obstacles of text discrepancies, as presented in Seljan et al. (2007). One of the files contains sentences in L1, and the other in L2, accepting only 1:1 alignment. TMX files are processed in a natural manner – L1 and L2 are clearly marked in the TMX file. *TMrepository* imports units (*tu*) with a single variant (*tuv*) in L1, and a single variant in L2. In order to avoid problems with excessive memory consumption for large TMX files, a custom made stream TMX parser was implemented. The last option, a pair of Word documents in DOC or DOCX format, automatically aligns any two Word documents on sentence level. The segmentation is performed by an SRX-based splitter and the alignment is done by the *hunalign* software (Varga et al., 2005). There are no assumptions as to the format of documents.

Furthermore, an additional gamification element for evaluation purposes is implemented into the system. Namely, users fluent in languages of the parallel data are encouraged to engage in reviewing of uploaded aligned corpora in the system. A user can perform a review of 50 translation units (in one go) selected at random out of all units that do not come from renowned corpora and have not been checked before. Therefore, all registered users can participate in the evaluation process and are rewarded 10 points for each reviewed translation unit, which is equal to uploading a new translation memory of 10 units. Each translation unit can be accepted, which is the default and more frequently chosen option, or rejected.

For every translation memory a specific translation memory overview report is created on demand, which shows how many translation units of the current translation memory were checked and how many of those were accepted. It serves as an indicator of the translation memory quality.

The ranking lists all the contributing users, sorted by the total number of sentence pairs they uploaded in a descending order. First three users are graphically exposed



Rank	User	Total units count	TM titles
1	(login anonymised)	549 627	hrenWaC,SETIMES,TedTalks,SETIMES2
2	(login anonymised)	142 510	The hunger games trilogy, A Song of Ice and Fire 1-5, 42 essays that have appeared in the bimonthly journal Atlantis Rising
3	(login anonymised)	107 241	English - Croatian Harry Potter,English - Croatian Lord of the Rings,English - Croatian 1984,English - Croatian Paulo Coelho
4	(login anonymised)	72 497	student corpus,Home Cinema Manual,TV Manual 1,TV Manual 2
5	(login anonymised)	71 122	Song Lyrics
6	(login anonymised)	53 321	The Blble and LOTR
7	(login anonymised)	25	Basics-business correspondence,Apology-Letter-to-Teacher

Fig. 1Ranking of contributing users

and awarded virtual medals. This element is used to introduce competitiveness among the users and thus increase their motivation. Importantly, the ranking also reveals the titles of translation memories contributed by the users. This is done in order to provide inspiration for other users, concerning potential sources of corpora.
For instance, noticing that people are uploading Lord of the Rings and other books might direct the search for corpora to some other book titles. Similarly, the fact of uploading TV manuals by one of the users might inspire others to acquire different translated user manuals and technical documentation. At all times, every user is able to see the current ranking (see Fig. 1).
The administrators of the *TMrepository* system have full access to all translation memories uploaded by the users via the administrative panel. The panel also shows the current translation unit (i.e. sentence pair) count in all the corpora.

4. Experimental scenario and results

This paper presents a study on applying crowdsourcing and gamification methods to a group of students with the use of the *TMrepository* system. The initial experiments were conducted in Poland.
A total of seven students who were taking part in the experiment were participating in an academic course of Natural Language Processing, as a part of their computer science studies. Before the experiment, they had completed between four to six semesters of their studies. Therefore, their background covered areas, such as: basic algorithms and C++ programming, intermediate programming, object-oriented programming (C\#, Java or PHP), basic web applications, mathematical analysis, algebra, logic and set theory. The background knowledge and technical skills were helpful in crawling the web and acquiring parallel data.
The students had no prior training in linguistics, machine translation or natural language processing. Furthermore, they were all native Polish speakers who did not speak Croatian, which was one of the reasons to include the reviewing feature in the system. Even though Polish and Croatian, as two Slavic languages, and are to some extent similar, they are not mutually understandable to speakers of only one of these languages. Nevertheless, the system interface is designed in English and is suitable for users who are not familiar or fluent in languages of the parallel data.
During the course, the students were first taught the basics of natural language processing using Linux bash scripts and the Python programming language. Exercises covered, among others: character and word frequency calculations, extracting and counting n-grams, sorting lines of input by different criteria. Afterwards, the students were given a lecture presenting the basics of web scraping, with the use of simple command-line tools such as *wget* and Python's *urllib* module. The lecture also mentioned the web crawling software framework *PyCrawler*.
After the lectures, the students were asked to start a project of their choice. They chose from a variety of suggested subjects, such as spell checkers, chatbots, lemmatisers, surname inflectors and many others. One of the suggested tasks was collecting Croatian-English or English-Croatian translation memories. The students who chose this task were instructed to acquire translation memories by "any means necessary", and to upload the harvested resources to the *TMrepository* system. The system was presented and explained to the students beforehand, as part of the training required in order to be able to use the system.
In total, 996343 translation units were collected during the study (see Fig. 1 which reports the actual numbers of units of participating users). Most common sources for parallel corpora included: commonly known Croatian-English parallel corpora, such as SETIMES or Ted Talks, but also books, technical documentation, song lyrics and business correspondence documentation.
The main characteristics of the collected parallel data are discrepancies in size of the translation memories, diversity of domains and domain independence, language register differences and style diversities. For instance, as the uploaded renowned corpora were very large in size, the distribution of collected units is very skewed across participants, i.e. one participant collected more than half of the total units, whereas another collected only 25 units.
The system also allows users to update collected resources. Taking into account that only seven participants took part in this pilot study, the authors consider the magnitude of the collected language resources a promising success. Furthermore, the resulting corpus contains interesting, original and previously unseen language resources, such as the song lyrics corpus.

5. Conclusions and future work

Data collection is traditionally very important in machine translation and other natural language processing tasks. Parallel data is frequently being used for training machine translation models, tuning machine translation systems, extracting data for developing monolingual statistical language models, building translation memories etc. Parallel corpora are also being utilised as an input for a vast range of natural language processing tasks.

Unfortunately, the acquisition of large amounts of parallel data is very time-consuming and tedious. Furthermore, parallel data is often collected biasedly, i.e. steered by various motivations. Crowdsourcing, on the other hand, can be initiated by an open call directed to the unknown public, or, as in this case, directed to a group of NLP students in order to harvest their labour.

As a result, the authors created a new tool, i.e. a custom-built crowdsourcing platform named *TMrepository* for which students managed to collect a total of nearly one million translation units. The collected data is ready to be exploited for the purpose of machine translation and in a variety of NLP tasks.

The authors' main focus for future work lies on building machine translation systems that are fed with corpora collected by the proposed method. Furthermore, the authors plan to combine the *TMrepository* system with other developed NLP software, such as Concordia – a tool which encompasses capabilities of standard concordance searchers and translation memory systems (Jaworski et al.,

2017). The process of collecting resources with the help of students will be continued, resulting in a constantly expanding repository of machine translation resources, i.e. parallel corpora. As for now, the system does not check for duplicate translation units, and therefore, it is possible to upload the same translation units. However, an intelligent deduplication mechanism is planned in the near future.

Considering the variety, uniqueness and magnitude of the corpora, the authors expect the performance of the machine translation pipeline and possibly other natural language processing tools trained with the acquired data to be competitive among other solutions for a less-resourced language, such as Croatian.

References

Abdelali, A., Durrani, N. and Guzmán, F. (2016) iAppraise: A Manual Machine Translation Evaluation Environment Supporting Eye-tracking. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA*, June 12-17, 2016, pages 17–21. The Association for Computational Linguistics, 2016.

Ambati, Vamshi and Stephan Vogel. (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, CSLDAMT '10*, pages 62–65, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Ambati, Vamshi, Stephan Vogel, and Jaime Carbonell. (2012) Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 1191–1194, New York, NY, USA, 2012.

Canovas, Marcos and Richard Samson (2011). Open source software in translator training. *Tradumática*, 9, 2011.

Howe, Jeff. (2006) The Rise of Crowdsourcing. *Wired Magazine, 14(6)*, 2006.

Howe, Jeff. (2008) Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. *Crown Publishing Group, New York, NY, USA, 1 edition*, 2008.

Jaworski, Rafał, Ivan Dunđer, and Sanja Seljan (2017). Usability Analysis of the Concordia Tool Applying Novel Concordance Searching. In *Proceedings of the 10th International Conference on Natural Language Processing (HrTAL2016)*. In print, Springer Verlag – Lecture Notes in Computer Science (LNCS/LNAI). Croatian Language Technologies Society (Zagreb), Institute of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb (Zagreb), 2017.

Koehn, P. (2005) Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005*.

Li, Hanchuan, Haichen Shen, Shengliang Xu, and Congle Zhang. (2015). Visualizing NLP annotations for Crowdsourcing. *CoRR, abs/1508.06044*, 2015.

Morschheuser, Benedikt, Karl Werder, Juho Hamari, and Julian Abe. (2017). How to gamify? Development of a method for gamification. In *Proceedings of the 50th Annual Hawaii International Conference on System Sciences (HICSS)*, Hawaii, USA, 2017.

Muegge, Uwe. (2013). Teaching computer-assisted translation in the 21st century. *Frank & Timme*, 2013.

Munro, Robert, Lucky Gunasekara, Stephanie Nevins, Lalith Polepeddi, and Evan Rosen. (2012). Tracking Epidemics with Natural Language Processing and Crowdsourcing. In *Wisdom of the Crowd, Papers from the 2012 AAAI Spring Symposium, volume SS-12-06 of AAAI Technical Report. AAAI*, 2012.

Muntés-Mulero, Victor, Patricia Paladini, Marc Solé, and Jawad Manzoor. (2012). Multiplying the Potential of Crowdsourcing with Machine Translation. In *Proceedings, Tenth Biennial Conference AMTA*, San Diego, 2012.

Packham, Sean and Hussein Suleman. (2015). Crowdsourcing a Text Corpus is not a Game, pages 225–234. Springer International Publishing, Cham, 2015.

Quinn, Alexander J. and Benjamin B. Bederson. (2011) Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 1403–1412, New York, NY, USA, 2011. ACM.

Robson, Karen, Kirk Plangger, Jan H. Kietzmann, Ian McCarthy, and Leyland Pitt. (2016). Game on: Engaging customers and employees through gamification. *Business Horizons, 59(1)*:29–36, 2016.

Sabou, Marta, Kalina Bontcheva, and Arno Scharl. (2012). Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW '12*, pages 17:1–17:8,

Seljan, S., Gašpar, A. and Pavuna, D. Sentence Alignment as the Basis For Translation Memory Database. In: *The Future of Information Sciences: INFuture 2007 - Digital Information and Heritage.* Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, Zagreb, 2007. 299-311.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón, V. (2005) Parallel corpora for medium density languages. In: *Proceedings of the RANLP 2005*, pages 590–596.

Zaidan, Omar F. 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, HLT '11*, pages 1220–1229