Multiplying the Potential of Crowdsourcing with Machine Translation

Victor Muntés-Mulero and Patricia Paladini
CA Technologies
Barcelona, Spain
Victor.Muntes@ca.com
Patricia.PaladiniAdell@ca.com

Marc Solé and Jawad Manzoor
Universitat Politècnica de Catalunya
Barcelona, Spain
msole@ac.upc.edu
jawad@ac.upc.edu

Abstract

Machine Translation (MT) is said to be the next lingua franca. With the evolution of new technologies and the capacity to produce a humungous number of written digital documents, human translators will not be able to translate documentation fast enough. However, some applications require a level of quality that is still beyond that provided by MT. Thanks to the increased capacity of communication provided by new technologies, people can also interact and collaborate to work remotely. With this, crowd computing is becoming more common and it has been proposed as a feasible solution for translation. In this paper, we discuss about the relationship between crowdsourcing and MT, and the main challenges for the MT community to multiply the potential of the crowd.

1 Introduction

Dynamic information in the Internet, together with the vast amount of documentation generated by software vendors and the ICT industry in general is growing exponentially. This makes it necessary to resort to automatic translation techniques that allow processing massive amounts of text in very short periods of time. Although human intervention is still essential in several industrial contexts, large companies like Microsoft Corp., Google Inc. or CA Technologies need to heavily rely on machine translation (MT) for speeding up the translation process. However, the quality provided by MT does not meet the strict requirements of some domains. Because of the current limitations of MT and the failure of human translators to provide a fast and efficient solution to translate this information, a shift in industry towards new technologies and solutions is being ignited.

In this context, crowdsourcing translation will be the next big breakthrough for the translation industry. By leveraging the potential of the crowd to perform tasks such as translating or post-editing, we increase our capacity to deal with large amounts of information at a fast speed. Instead of being an alternative for MT, crowdsourcing and automatic translation will be coupled to offer a comprehensive solution where quality and fast and non-expensive translation coexist.

MT is already being used to reduce translation time and homogenizing quality of human outputs. This is essential, for instance, in many scenarios where preserving style coherence in large documents is important. However, crowdsourcing will introduce a significant change in the profile and motivations of translators and the way translation will be performed. The average experience of translators in the crowd will change, compared to the classical scenario where only professional translators participate, and this will create new needs regarding the tools used for translation and localization. Crowdsourcing will provide new opportunities for customers to participate in the quality reviewing process and new mechanisms and algorithms will be devised to guarantee quality in the crowd.

In this scenario, a fundamental question is how MT can multiply the potential of crowd computing. In this paper, we discuss the challenges that MT must face to welcome crowd computing.

This paper is organized as it follows. In Section 2, we present an overview on crowdsourcing in general and applied to translation. Section 3 describes the main motivation for using crowdsourcing in industry and how crowdsourcing is mixed with MT. In Section 4, we discuss the main challenges presented by crowdsourcing related to MT. Finally, Section 5 concludes this paper.

2 Related work

In 2006, Jeff Howe coined the term crowdsourcing to describe the increasing practice of outsourcing tasks on internet through a call open to a large variety of users (Howe, 2006). Since then, several general purpose crowdsourcing platforms have appeared such as Amazon Mechanical Turk (mturk.com), or more customized systems such as CrowdFlower (crowdflower.com) or ClickWorker (clickworker.com), that extend the former by using gold standard units, redundant reviews and dividing complex tasks into smaller units, allowing the distribution of these tasks among the crowd based on their profile.

There are several challenges that must be overcome to allow the generalized use of crowdsourcing solutions. One of the main challenges of crowdsourcing is ensuring quality. In general, the capacity to provide quality depends on the type of work to be crowdsourced. Previous work such as (Lease and Yilmaz, 2012) or (Yan et al., 2010) show us that the crowd may perform very well in terms of quality in some scenarios but they might not be comparable to trained in-house experts in others, depending on the nature of the task to be completed. The rewarding system is also linked to quality. For instance, Harris (Harris, 2011) shows that financial incentives actually encourage quality.

Most examples where crowdsourcing is used are characterized by using the crowd to solve simple and usually independent tasks, such as labeling an image. However, in (Kittur et al., 2011), the authors present a framework that makes it possible to use the crowd to solve more complex tasks by splitting them in subtasks and executing them in parallel.

There are many success stories that show that crowdsourcing can be used for commercial and industrial applications. These include NamingForce (namingforce.com), Threadless (www.threadless.com), InnoCentive (innocentive.com), TopCoder (topcoder.com) or uTest (www.utest.com), and many others.

Crowdsourcing Translation

One of the typical problems solved by crowdsourcing is translation. There are several reasons for this which we discuss later in this paper. Most crowdsourced translation proposals are based on the use of Amazon Mechanical Turk or similar platforms (Denkowski and Lavie, 2010; Gao and Vogel, 2010; Negri et al., 2011; Zaidan and Callison-Burch, 2011). The crowd is used for various purposes. For instance, Zaidan et al. (Zaidan and Callison-Burch, 2011) study the characteristics of workers that impact the quality of translation and use them to select the best translation among a set of candidates. In their experiments, they show that the quality of the work done by the crowd is close to that provided by professional translators. Crowd computing has also been used to evaluate the output of MT in (Bentivogli et al., 2011; Denkowski and Lavie, 2010), to perform word alignment (Gao and Vogel, 2010), and to create corpora to feed and enrich Statistical MT (Negri et al., 2011).

Crowdsourcing has also been proposed for post-editing tasks, such as text shortening and proofreading (Bernstein et al., 2010). In this case, authors propose the Find-Fix-Verify pattern to split the tasks into a series of phases that utilize independent agreement and voting to produce reliable results. At an industrial level, CA Technologies proposed the Action-Verification Unit (Muntés-Mulero et al., 2012b), a quality control mechanism that helps organizing the crowd to perform actions, as well as verification of the quality of the results during the process. Other platforms such as Gengo (gengo.com) also offer translation services based on crowdsourcing.

In general, previous experiments are based on the resolution of simple tasks at sentence level, using quality assurance methods based exclusively on the reduced amount of information contained in the segment. Besides, the use of automatic evaluation methods like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) to evaluate the quality of translation done by the crowd is quite common. Unfortunately, these methods are based on the existence of pre-computed golden translations which are usually not available for unpopular languages and cannot be used for new texts. This results in high cost and time consuming human intervention to check quality.

3 Crowdsourcing in Industrial Environments

The common limitations in the prevalent translation process that motivate companies to crowdsource translations are described in (Muntés-Mulero et al., 2012b). We can summarize them in the following four ideas: (i) translation is slow: since the localization process is time-consuming, products that need to be translated to other languages are usually released significantly later than the version in the original language, potentially causing loses to the companies; (ii) dynamic workloads require elastic solutions: the amount of text to be translated is heterogeneous in general, making it difficult for the inhouse team of translators to cope with peaks in the workload, hence forcing the companies to outsource the excess part of the work to external translation service providers; (iii) when exploring new markets, it is difficult as well as expensive for companies to find and hire translators for all languages; and (iv) the cost of software localization is very high in general, causing many products not to be considered for internationalization even for common languages such as Spanish, German or French.

As a consequence, many large companies are moving into crowd-based systems in order to perform translations. For instance, Microsoft Corp. crowdsources editing of MT of their knowledge base. Just as a second example, CA Technologies is building a semi-automatic management platform that allows the integration of MT and crowd computing for reducing the cost of post-editing phase in software localization. Crowdsourcing is becoming the answer for many companies to translate large amounts of text very fast.

Figure 1 depicts the process followed to localize a user guide of a software product. A similar figure could be used to depict the process followed to localize user interfaces. Once the source text has been obtained from the original document, MT is usually used for an initial translation. Usually, the original and the MTed texts are sent to human translators who post-edit the text written in the target language.

By using crowdsourcing techniques it is possible to achieve scalability, elasticity and reduction in the cost of the software localization process.

Due to the increasing quality of MT, the pool of motivated bilingual speakers may increase, making crowdsourcing models more popular. The role of professional translators will probably change from solely post-editors of MT to also post-editors of the work done by non-professional bilingual speakers. Besides, they will tend to move to terminological and quality control roles. As a consequence, the profile of the translator handling the output of MT will change, moving toward a more amateur profile.

As a result of this upcoming transformation in the way translation is performed and the profile of translators, several new challenges arise:

- ***The average experience of translators will decrease***. As a consequence, the time needed to translate a sentence may vary. It may increase due to the lack of experience of commited translators or decrease because of those unexperienced translators who do not take their work so seriously.

- ***Translation productivity will be strongly related to the clarity and usability of UIs***. Crowdsourcing implies multi-user web-based platforms. Translator productivity is related to how the user interacts with these platforms and other aspects linked to UIs.

- ***Quality control systems will have to be reinforced*** in order to meet the requirements of the new distributed and heterogeneous translation environment. In general, translators will be less agile in order to solve non-trivial situations, such as translations with a lack of context or to guarantee style coherence.

- ***Scalability will be essential***, because of the potentially large amount of translators working concurrently in the system. This brings new issues related to parallelism and concurrency management.

- ***Translation management will be done from the cloud***. Crowdsourcing will increase the

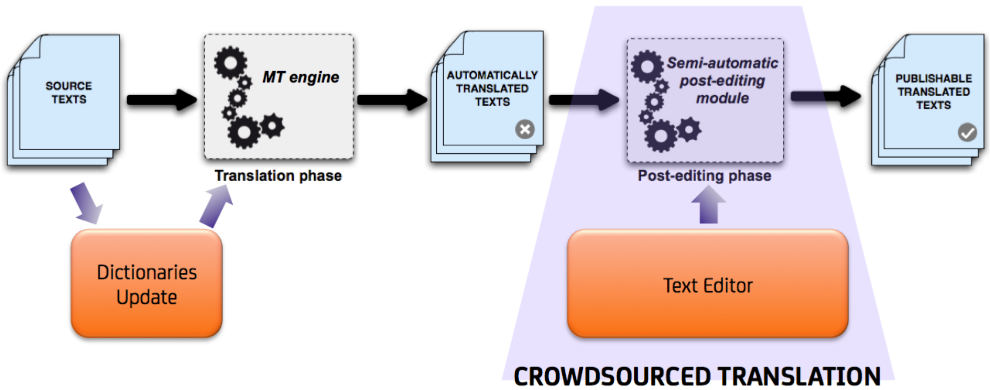

Figure 1: Use of crowdsourcing for software localization.

presence of occasional and non-expert users. Because of this, participating in a crowdsourced translation must not involve complex local software installations. Workers in the crowd will expect to use online applications through their web browsers, ignoring where processes are executed, platform constraints, performance aspects and resource requirements. Therefore, crowdsourcing demands for cloud-based management platforms where the translation workload can be managed elastically, using Software-as-a-Service models.

- ***The heterogeneity of devices used to perform post-edition will increase***. As the profile of translators gets broader, we cannot make assumptions about the characteristics of the equipments used by workers. Crowdsourcing may even increase the number of users working from mobile devices.

4 Research Challenges

Machine translation has a large effect on post-editing. Instead of being superseded by crowdsourcing, MT will be coupled with it and will multiply the potential of crowd computing for translation. In this sense, we forsee different areas where MT will be essential to guarantee the success of crowdsourcing.

4.1 Improving MT quality is essential for the success of crowdsourcing

Improving the quality of MT decreases the time of post-editing and improves the quality of the final output. This will be essential to compensate the lack of experience of translators in the crowd. In many situations translators face the problem of interpreting the context in order to produce the correct translation, and the lack of experience might be an important drawback. MT will have to become aware of context in different domains. In general, MT engines translate independent sentences, ignoring broader contextual information. Even at sentence level, Statistical Machine Translation (SMT) engines based on segments or phrases, such as those presented in (Koehn et al., 2003), only make a local use of the source lexical context, restricting the context to a limited number of words next to the phrase being translated. Other approaches try to extract semantic information from the text near the sentence to be translated. For instance, syntax-based SMT (Chiang, 2005) considers syntactic dependencies between long distance unconnected phrases within a sentence to compensate this lack of connection. Factored models (Koehn and Hoang, 2007) enrich phrase-based models with extra linguistic information associated to words. Other work focuses on word sense disambiguation to choose between possible translations of a word or a phrase. Usually, machine learning methods are used to select the correct words (Giménez and Màrquez, 2008). However, context might be related to other aspects beyond the words in the text. For instance, in (Muntés-Mulero et al., 2012a) authors suggest that MT should also be aware of context information in UIs to improve the quality of MT for software localization. Making MT more aware of context would improve the quality of the MT output, reducing the probability of mistakes during post-editing.

Nowadays, it is widely recognized that any of the MT approaches that exclusively use either rule-based or SMT (or other alternatives) for achieving good-quality automatic translation are not accurate enough, and that future models have to resort to Hybrid Machine Translation (HMT), to combine the best of the other methods. It is possible to achieve significant improvements by simply combining MT engines of different types. For instance, in (Hildebrand and Vogel, 2008) authors improve the quality of translation by choosing the best translation given by any of a set of MT systems working in parallel. Also, the work presented in (Li et al., 2009a) describes a method in which several translation system decoders collaborate by sharing partial translation results. Producing natural and correct text will also help in diminishing the mistakes in post-editing.

Finally, style coherence is jeopardized when translations are crowdsourced because of the partition of documents to be translated into relatively small tasks that are distributed among many different people. In this situation, it is interesting to remark that MT becomes essential in order to preserve the style coherence. Through previous training, MTed texts tend to homogenize the style in the text through different translators, making it even more important to use MT when crowdsourcing.

4.2 Interactive Machine Translation will be important to assist the crowd

Because crowdsourcing involves multi-user web-based platforms, designing the correct UIs and assisting the translator on-line will be essential. Interactive Machine Translation (IMT) (Langlais et al., 2004) was proposed to assist the human translators, adapting data driven MT techniques to be used in collaboration with human beings. For instance, in (Barrachina et al., 2009) authors propose a system in which SMT systems are used to produce target sentences hypotheses which can be accepted or amended by a human translator. Each corrected text segment is then used by the MT system as additional information to achieve improved suggestions. It will be important to optimize the relationship between dynamically applied machine translation and the translator's efficiency, tending to translator-centric approaches. A suggestion that changes on every key stroke may obtain accurate automatic results, but it may decrease the productivity of the translator because of the cognitive effort needed to process those changes.

4.3 Machine Translation must scale

A crowdsourcing platform may manage hundreds or even thousands of translators at the same time. As we said in the previous section, participating in the crowd should not involve installing and running software locally. Besides, MT systems usually require a lot of memory space once they are trained, making it difficult to handle them in local heterogeneous machines, ranging from local servers to mobile devices. Current proposals offer the possibility to be installed in a separate server to process remote translation requests. This poses an important challenge in order to manage online MT translations when IMT is used, for instance. Also, translation memory updates must be managed concurrently. Because of this potentially high concurrency and the large amounts of corpora to be processed, MT performance will become an important issue.

The traditional way to cope with large workloads is to establish farms of servers running independent instances of the MT system (Sánchez-Cartagena and Pérez-Ortiz, 2010). This scenario poses new questions regarding how different data repositories must be synchronized and how to guarantee the coherence of the different models generated in the system. Another interesting question is how to combine the relatively quick translation requests with the predictably much slower model update events. Since adaptive MT systems are not commonplace, up to now using server farms with replicated models has been able to cope with the increasing bandwidth demands, but in the foreseeable future it is likely that most of the models produced will no longer fit in one machine, but have to be distributed among a set of machines. In this scenario, a farm of servers processes the requests by running a huge shared single model (not a number of model copies) that is actually distributed in several machines (Li and Khudanpur, 2008). In this sense, MT systems will increasingly borrow functionalities from Database Management Systems (DBMS) which are the systems that have already faced similar scalability challenges. As an illustrative example, MT systems that have considered scalability issues (Talbot and Osborne, 2007; Li et al., 2009b) incorporate Bloom filters, a long-known and well-established feature to save space in the database world.

5 Conclusions

The potential of crowdsourcing to adapt translation services to the ever growing amount of information is unquestionable. The volume of data and dynamism in the data workloads require systems that are elastic and scalable. Crowd computing is already being used to overcome these challenges. However, with the adoption of this new paradigm, new issues arise. MT has the potential to fill the new gaps created in terms of quality and performance. The translation industry is at the beginning of a new era in which MT has a central role.

Acknowledgments

Victor Muntés-Mulero would like to thank Subprograma INNCORPORA (línea INNCORPORA-Torres Quevedo) and the Ministerio de Economía y Competitividad for their support through grant PTQ-11-04852.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. 2009. Statistical approaches to computer-assisted translation. *Comput. Linguist.*, 35(1):3–28, March.

Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. In *MT Summit XIII*, pages 521–528.

Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Procs. of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, NY, USA.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270, June.

Michael Denkowski and Alon Lavie. 2010. Exploring normalization techniques for human judgments of machine translation adequacy collected using amazon mechanical turk. In *Procs of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 57–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2010. Consensus versus expertise: a case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 30–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Giménez and L. Màrquez, 2008. *Discriminative Phrase Selection for SMT*, pages 205–236. NIPS Workshop Series. MIT Press.

Christopher Harris. 2011. You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks. In Matthew Lease, Vitor Carvalho, and Emine Yilmaz, editors, *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 15–18, Hong Kong, China, February.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 254–261, Waikiki, Hawaii, October.

Jeff Howe. 2006. Wired 14.06: The Rise of Crowdsourcing.

Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. 2011. Crowdforge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 43–52, New York, NY, USA. ACM.

P. Koehn and H. Hoang. 2007. Factored Translation Models. In *Proceedings of the Conference on EMNLP*, pages 868–876.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL*, Edmonton, Canada, May 27-June 1.

Philippe Langlais, Guy Lapalme, and Marie Loranger. 2004. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation (Special Issue on Embedded Machine Translation Systems*, 17(17):77–98.

Matthew Lease and Emine Yilmaz. 2012. Crowdsourcing for information retrieval. *SIGIR Forum*, 45(2):66–75, January.

Zhifei Li and Sanjeev Khudanpur. 2008. A scalable decoder for parsing-based machine translation with equivalent language model state maintenance. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, SSST '08, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mu Li, Nan Duan, Dongdong Zhang, Chi-Ho Li, and Ming Zhou. 2009a. Collaborative decoding: partial hypothesis re-ranking using translation consensus between decoders. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 585–592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009b. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Victor Muntés-Mulero, Patricia Paladini, Cristina España-Bonet, and Lluís Màrquez. 2012a. Context-aware machine translation for software localization. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT12)*, pages 77–80, Trento, Italy, May.

Victor Muntés-Mulero, Patricia Paladini, Jawad Manzoor, Andrea Gritti, Josep-Ll. Larriba-Pey, and Frederik Mijnhardt. 2012b. Crowdsourcing for industrial problems. In *Proceedings of the 1st International Workshop on Citizen Sensor Networks (CitiSen2012)*, Montpellier, France. To be published.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 670–679, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA.

Víctor M. Sánchez-Cartagena and Juan Antonio Pérez-Ortiz. 2010. ScaleMT: a free/open-source framework for building scalable machine translation web services. *The Prague Bulletin of Mathematical Linguistics*, 93:97–106.

David Talbot and Miles Osborne. 2007. Smoothed bloom filter language models: Tera-scale lms on the cheap. In *EMNLP-CoNLL*, pages 468–476. ACL.

Tingxin Yan, Vikas Kumar, and Deepak Ganesan. 2010. CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. In *MobiSys '10: Procs of the 8th intl. conf. on Mobile systems, applications, and services*, pages 77–90, NY, USA, June. ACM.

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*, pages 1220–1229, Portland, Oregon, USA, June.