Social Translation: How Massive Online Collaboration Could Take Machine Translation to the Next Level

Position paper for the FLaReNet Forum 2010

Juan Antonio Pérez Ortiz

japerez@dlsi.ua.es
Transducens Research group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, Spain

Introduction

Internet users, mostly passive consumers in the first years of the web, have quickly become active *prosumers*[1] of information in the current era of the *web 2.0* and the *cloud*.[2] However, in spite of the vast amount of contents and collaboratively-created knowledge uploaded to the cloud during the last years,[3] linguistic barriers still pose a significant obstacle to universal collaboration as they lead to the creation of "islands" of content, only meaningful to speakers of a particular language. Until *fully-automatic high-quality machine translation* becomes a reality, massive online collaboration in translation may well be the only force capable of tearing down these barriers [3] and produce large-scale availability of multilingual information. Actually, this collaborative translation movement is happening nowadays, although still timidly, in applications such as Cucumis.org, OneHourTranslation.com or the Google Translator Toolkit.

The resulting scenario, which may be called *social translation*, will need efficient computer translation tools, such as reliable machine translation systems, friendly postediting interfaces, or shared translation memories. Machine translation technology provides draft translations which are then proofread by people (a process known as *postediting*), a task which may require less effort than doing the translation from scratch. Remarkably, collaboration around machine translation should not only concern the postediting of raw machine translations, but also the creation and management of the linguistic resources needed by the machine translation systems; if properly done, this can lead to a significant improvement in the translation engines.

Therefore, massive online collaboration could eliminate the language barriers on the web, but for this to happen as many hands as possible are necessary; and this includes involving speakers that, in principle, do not have the level of technical *know-how* required to improve machine translation systems or manage linguistic resources. Consequently, software that can make those tasks easier and elicit the knowledge of both experts and non-experts must be developed [1]. Note that people's contribution should not necessarily come in an intentional manner; for example, postedited texts are a rich source of information for learning algorithms.

The whole picture shows a world where people, intentionally or unintentionally, contribute to a *commons* of linguistic resources and translation engines, which, ideally, should be as open and widely accessible as possible to boost synergies and *network effects*.[4]

Paradigm shift for true social translation

In order to fully accomplish the goals of the scenario that has been drawn up for social translation, my thesis is that a new paradigm that takes into account the following issues should emerge. Note, however, that the real world is affected by a number of pulling and pushing forces as well as conflicts of interest which could set social translation at some distance far from the ideal point.

- **Data portability.** People should be the real owners of their data and be able to reuse them across fully interoperable translation applications (w*alled gardens* restricting data export or access from the outside are the other side of the coin).
- **Standard formats.** Reusable and portable data implies that standards or common practices are used for their encoding and representation.

[1] Users which are both producers and consumers.

[2] Two more neologisms which are commonly used as synonyms for the present-day internet.

[3] E.g., more than one million messages are sent every hour to the Twitter microblogging service [6].

[4] When network effects hold, the value of a product or service increases as more people use it.

- **Licenses.** In order to ease the global effort on social translation and prevent "wheel reinvention", appropriate open licenses should be embraced. Ideally, this kind of licenses should be recommended as default to every user.
- **Linked data.** "Data is relationships" Tim Berners-Lee said in a recent talk on Ted.com; in our case, this means that all the linguistic data should be annotated and properly interconnected.
- **Cloud computing.** Machine translation should be in the cloud instead of confined to the desktop of a single user. The global flow of translations, posteditings, linguistic data, etc. constitutes the main resource for social translation.
- **Scalability.** In order to fulfill the demands of global translation, related applications and *programming interfaces* (API) should be scalable [5] and provide low-latency delays.
- **Code availability.** As important as open data is the availability of free/open-source software that encourages research and development of applications in the social translation area.
- **Multiengine translation.** All the available machine translation systems should cooperate and offer public APIs; this could lead to hybrid systems choosing the most appropriate machine translator according to context.
- **Standard interfaces.** Interfaces for the different ways of interacting with machine translation systems (postediting, management of linguistic resources, etc.) should be made uniform and predictable (as it is the case, for example, with word processors).
- **Accessibility.** As many people as possible should be able to contribute to the social translation sphere.

The Tradubi Web Application for Social Translation

My research group at University of Alacant, Spain, is currently developing Tradubi [4], a free/open-source Ajax-based web application for social translation, whose aim is, firstly, to build a platform for collaboratively customizing and improving rule-based machine translation systems and, secondly, to offer an environment for postediting raw machine translations and subsequently sharing the corrected texts. Currently, Tradubi is in its early stages of development and built upon the free/open-source Apertium machine translation engine [2]. The application can be accessed at tradubi.com, or downloaded from tradubi.org and installed on a different server. With the help of Tradubi, users can create customized dictionaries for Apertium which focus on specific linguistic domains, or which correct translation errors made by the default system. We expect to augment the features of Tradubi so that it becomes a powerful free/open-source application for social translation.

Conclusions

The massive collaboration of internet users as well as the contribution of science in the development of efficient tools allowing the improvement of machine translation engines could be fundamental in the abolition of the language barriers that currently restrict universal access to the web. This large-scale collaboration of experts and non-experts implies a change of paradigm in the way linguistic resources are managed. Tradubi is an example, among others, of a web application that hopefully will help to normalize social translation.

Acknowledgments

This work has been partially funded by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01.

References

1. Font-Llitjós, A., J. Carbonell, and A. Lavie. A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. In *Proceedings of EAMT 10th Annual Conference*, 2005.
2. Forcada, M. L., F. M. Tyers, and G. Ramírez-Sánchez. The Apertium machine translation platform: five years on. *First International Workshop on Free/Open-Source Rule-Based Machine Translation 2009*, 3–10.
3. Garcia, I. Beyond Translation Memory: Computers and the Professional. *The Journal of Specialised Translation*, 12:199–214, 2009.
4. Sánchez-Cartagena, V. M. and J. A. Pérez-Ortiz. Tradubi: Open-Source Social Translation for the Apertium Machine Translation Platform. In *Open Source Tools for Machine Translation, MT Marathon 2010,* 47–56.
5. Sánchez-Cartagena, V. M. and J. A. Pérez-Ortiz. ScaleMT: A Free/Open-Source Framework for Building Scalable Machine Translation Web Services. In *Open Source Tools for Machine Translation, MT Marathon 2010,* 97–106.
6. Schonfeld, E. Pingdom Says People Are Tweeting 27 Million Times A Day. TechCrunch.com, November 12, 2009.