Web 2.0 and Localisation

Lucía Morado Vázquez
Centre for Next Generation Localisation
Localisation Research Centre, Department of Computer Science and Information Systems
Limerick, Ireland
Lucia.Morado@ul.ie

Dimitra Anastasiou
SFB/TR8 Spatial Cognition
Computer Science/Language Sciences
University of Bremen
00491637435527
anastasiou@uni-bremen.de

Chris Exton
Centre for Next Generation Localisation
Localisation Research Centre, Department of Computer Science and Information Systems
Limerick, Ireland
Chris.Exton@ul.ie

Ian O' Keeffe
Centre for Next Generation Localisation
Localisation Research Centre, Department of Computer Science and Information Systems
Limerick, Ireland
Ian.OKeeffe@ul.ie

ABSTRACT

The more web-based communities multiply in number, type and geographical distribution, the more varied and global their requirements will become. This paper focuses on a selection of Web 2.0 domains and considers how these communities provide both new challenges, and new opportunities, to the software localisation community. In addition to discussing how localisation requirements are affected by this new content type, we consider two possible solutions to the challenges that Web 2.0 represents for localisation: Machine Translation and Social Translation.

Categories and Subject Descriptors

J 4 [Social and Behavioral Sciences], I 2.7. [Natural Language Processing]: Machine Translation.

General Terms

Management, Reliability, Security, Human Factors, Standardization, Languages, Legal Aspects, Verification.

Keywords

Localisation, Web 2.0, Social Translation, Machine Translation.

1. INTRODUCTION

Web 2.0 refers to a new form of web-based user interaction that was initially facilitated when users were encouraged to generate and share their own content. It includes a number of interaction modes such as social-networking sites, video-sharing sites, wikis and blogs. The concept of 'Web 2.0' began with a conference brainstorming session between O'Reilly and MediaLive International [1]. For the purposes of this paper we consider Web 2.0 to be the combination of a new type of user interaction that is based on user generated content (UGC) with the new technologies that facilitate and empower this new type of user interaction.

Prior to Web 2.0 the majority of web content was generated for a specific target audience. The language of communication used was generally not an issue, as content was produced for the needs of a particular audience. In addition Web 1.0 content was relatively static and when information was required to be published in multiple languages, this was carried out using well established methods utilised by the software localisation community. With the advent of Web 2.0 there are a number of new issues which seem incompatible with the traditional model of software localisation. First and foremost, the number of contributors has increased to potentially billions. As a by-product of this explosion we are also witnessing a large expansion in the number, style and quantities of languages used. Not only does the demand for traditional forms of localisation based on a professional pool of translators and localisation experts become unsustainable due to the ever growing diversity and scale of content, but the requirements and nature of the type of translation has shifted.

Web 2.0 changes the nature of web content from being centrally produced and published by a corporate or governmental body to a more dynamic paradigm where content is produced by a large, and mostly unprofessional and widely distributed, user base. Although the basic requirements may be the same, user expectations and requirements are fundamentally different, for example users tend to accept a much lower level of professionalism and presentation whilst the temporal aspect of the information such as currency is of major importance.

2. RELATED WORK

Undoubtedly, Web 2.0 and the constant increase of UGC lead to a higher demand for translation, thus in this paper we examine two possible solutions to cope with the high volume and the high speed of content production: the older technology of Machine Translation (MT) and the newer trend of Crowdsourcing/social translation which came into play only the last few years. The term 'Crowdsourcing' was first coined by Jeff Howe[2] in his blog as follows:

> [C]*rowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.*

Crowdsourcing is related with different areas, like art, music, and photography. We focus on Crowdsourcing translation, i.e. when the crowd or a motivated part of it, the community, participates in an open call and thus translators volunteer to translate some content. Crowdsourcing platforms are implemented to translate both enterprise and/or personal/social content; in this paper we focus on the latter kind. Examples of social translation are the translation of *Facebook*, but also the *TED Open translation project* which allows social translation of their video content.

Initially this paper will present a number of the subtle and seismic changes that Web 2.0 represents when compared to Web 1.0. It

Conference'10, Month 1–2, 2010, City, State, Country.

will then consider the challenges that these represent for the localisation community. As part of the discussion we will consider two possible approaches that may address some of these challenges: MT and Social Translation (ST). MT is a relatively well researched field when compared to ST; it however does have many drawbacks. The ST model addresses many of the failings of the MT model by leveraging the diversity and scale of the Internet to mobilise a people-power solution.

3. CHANGED WORLDS

Web 2.0 has attracted the attentions of many IT specialists and been responsible for the production of many studies, but comparatively little is known about its potential future and its consequences in our lives: "For better or for worse, Web 2.0 participatory media is reshaping our intellectual, political and commercial landscape" [3]. This quote predicts the changes that Web 2.0 may bring to us and to our way of dealing with data. If we consider this new phenomenon in relation to localisation, we discover a new scenario. For example, in many cases Web 2.0 users utilise a searchable and comprehensive historical archive of previous postings and discussions for the purpose of overcoming much of the deficit in terms of clarity and grammar which is often associated with social network content. Their mode of interaction tends to be quite constructivist or social constructivist in nature where an understanding is formed via a collage of differing sources and shared peer to peer exchanges. When we take the new types of user interactions that are facilitated by Web 2.0 environments into consideration, there are several challenges that we have to face, and a number of possible solutions that can be considered, when contemplating how Web 2.0 can be localised: time (immediacy), quantity, quality, and cost. Here we briefly discuss them separately:

a) **Immediacy**: Web 2.0 is a constantly changing environment where digital content is in a continual state of flux. UGC, like other content, has a range of time-sensitivity, but often much of the information has an extremely short use by date. Relevancy of information is closely associated with how up to date it is. In order for the information to be relevant to the consumer, it needs to be localised in the shortest possible time.

b) **Quantity:** As stated before, the information produced by users is constantly growing [4]. Web 2.0 applications are a relatively new and emerging phenomenon. It has still to be seen exactly what effect they will have on the broader social interaction of the Internet, as user interaction has begun a fundamental shift when compared to the more traditional web-based interactions. Nevertheless, it depends on the user how well they publicise, control, and manage their generated content. It seems very difficult to address such an amount of UGC with the traditional localisation processes anymore given its high volume and the high speed it is generated. UGC is constantly growing; this naturally implies an increase not only in good quality content, but also in the amounts of bad quality or malicious content. This makes it difficult to distinguish between data of high quality and poor data. However, in this paper we propose a way of facing the problem: selecting information by relevance and/or popularity. It should be noted that popularity is a very approximate measure of relevance. As most of the information present in Web 2.0 is user generated, we think that the best way to classify its relevance is by letting fellow users decide via a voting system. These voting systems are already present in Web 2.0, for example, consider the comments that users can leave in response to a *Youtube* video; other users can vote if they agree with them or if they think the comment is offensive. Similar voting systems are present in many Web 2.0 sites, e.g. *eBay*, *Amazon*, and *IMDB*. Our objective would be to localise only the information that had been selected as relevant by the majority of the users; the rest of the information would be ignored and would remain in the original language. This voting method is far from perfect, but it can represent a good starting point to sort information.

c) **Quality**: Although there may be professionals creating content in the Web 2.0 environment, most of the content to be found in Web 2.0 is not professionally created. There are thousands of poorly written blogs, misleading wikipedia articles, and insulting comments which show the potential pitfalls that await such an approach. [5] calls fake articles 'u-boots', because they can be detected only accidentally. As a result of this, localisers may well face extra difficulties in localising content and solving user generated errors, as the borders between reader and writer are blurred. However, [5] argues that also encyclopedias in printed form can have mistakes and these mistakes are transcribed to other dictionaries too.

d) **Cost:** A regular user (or his/her followers) may not be able to pay for a professional localisation process. Most of the content that is produced with Web 2.0 applications has little or no commercial value. The majority of the blogs or wikis are not backed up by the resources of large corporations, so it seems unrealistic that traditional and costly localisation processes could be applied to such UGC. However, UGC varies, and value depends on content. Tech blogs and *how-to* articles have high value in terms of cost saving.

4. POSSIBLE SOLUTIONS

Taking into account the challenges that Web 2.0 presents, we propose two possible solutions: Machine Translation and Social Translation.

4.1 Machine Translation

There are several MT applications that are found for free online; some of them are even open source projects. Due to the complexity of the human language, machines find it difficult to deal with irony, ambiguity, or humour. As a result of that, the results we can get from MT are still far from those obtained with human translation. However, there is research carried out towards this direction and an example is a series of Workshops on Computational Approaches to Linguistic Creativity which address the issues of "creative language usage at different levels", from the lexicon to syntax to discourse and text. Also, there are cases where MT has proved very efficient without human interaction; for example, *Météo* System – developed by the TAUM group – was used by the government of Canada to translate weather forecasts from English into French and vice versa; this system used controlled language.

The issue we should address is whether MT can become a solution to the localisation of Web 2.0 content; this needs to be decided depending on the circumstances. It can be a solution if the quality requirements are not high and all we want is to have a first 'raw' translated version; for publishing purposes post-editing of the machine translated version is indispensable. It can be also a solution if the topic and the language used are restricted. Due to the nature of Web 2.0, this sounds unrealistic, as only in very specialised forums written by people aware of internationalisation requirements would this work with an acceptably high quality.

4.2 Social Translation

Recently, the Social Translation (ST) phenomenon has been attracting media attention. *TAUS* called it 'community translation', *Common Sense Advisory* called it CT3 (community translation+collaborative technology+crowdsourcing), while [6] proposed the term 'hive' translation, "since the unbounded nature of cyberspace associations clearly transcends old notions of community". The term Social Translation (ST) has existed for a while; in the round-table discussion on translation in the new millennium [7], Newmark mentioned that in 'social translation', social means social responsibility, while Sylfest Lomheim thought of categorising 'social translation' between literary and non-literary. We would prefer to correlate 'social' with the content and the lack of monetary compensation; thus we define ST as follows:

Social translation is the translation activity focused on social content carried out by one or more volunteer translators that do not receive any monetary compensation for their work.

Although it has been popularised by Web 2.0 applications like the social network *Facebook*, ST has always existed. Volunteer translators or simply advanced users have been localising software (often open source software) from the moment they were created. As [8] reflect in their book: "just because people don't get paid to participate in peering does not mean, however, that they do not benefit from their participation in other ways". Traditional ST requires a high knowledge of computing in order to be able to accomplish the work, as the interfaces and workflows tend to require specialist training or ability. However, the adoption of good Human Computer Interaction (HCI) design practices in the development of collaborative platforms can help to simplify the process and make it more accessible to regular users, thus allowing more and more people to join the social translation club. For example, the use of metaphor, where interaction with the system is made to look similar to tasks the user is already familiar with, can lead to greater user confidence and draw the user in.

The cost problem can also be addressed with ST as Web 2.0 technologies are often open source and free to use, and the translator's job is done for free [8]. The motivational aspect of participation in a ST project can be compared with other collaborative online projects such as *Wikipedia*. In such projects the motivation can be as diverse as simple altruism to reputation, freedom of movement or reciprocity (see [9]).

4.2.1 Drawbacks of Social Translation

ST should not be regarded as the panacea for all the problems that Web 2.0 generates, as it could also demonstrate bad side-effects: low quality, lack of CAT (Computer-Aided Translation) tools, 'trolls', and vandalism.

a) **Quality**: As well as the content being created by non-professional authors, most of the translations are done by non-professional translators. This means that they are not always aware of the problems that are involved in creating a translation. There have been some strong criticisms about the quality of the content in Web 2.0: "But the more self-created content that gets dumped onto the Internet, the harder it becomes to distinguish the good from the bad" [3]. Most of the translation applications for Web 2.0 have control systems. However, it is dependant on the regulators themselves as to whether this is effective or not. For example, in the translation of the interface of *Facebook*, translations have to be submitted and approved by the same people. It is a simple democratic voting system, but history has shown us that majority decisions can also be wrong. [3] also addresses this issue:

   In theory, Web 2.0 gives amateurs a voice. But in reality it's often those with the loudest, most convincing message, and the most money to spread it, who are being heard.

   Other applications have a stronger control system. For example, "Launchpad", the translation platform of Ubuntu, lets the users decide how their product is going to be translated or controlled.

   The number of volunteer translators is also a factor to be taken into account. Some languages, especially minority languages, do not have access to a large group of volunteers who are available to work suitable hours, thus leading to much smaller quantities of time being exercised on translation and proofreading. As a result, the quality could be negatively affected in such small speech communities. It also does not seem economically viable to hire a professional proofreader to do the final job and assure good quality in these circumstances. Only projects with a good budget would be in a position to do this.

   The quality problem can also be generated in the original language, as not all the users are professional writers or have the time to proofread their production. An exacerbating factor is the large number of second language speakers who do not always produce perfect speech. A possible solution to improving the quality of the original work, and to help users to improve their writing skills, is by providing them with writing aid tools like spell and grammar checkers or dictionaries. These tools, that are so prevalent in text editors, are not available in most of the Web 2.0 content generators. Another aspect that we should take into account about the quality of ST is that users are not as exacting as somebody who has paid for a software utility. ST users know that the products are far from perfect and they accept that situation.

b) **Lack of CAT translation tools**. The collaborative platforms that are available to implement ST projects, like *Launchpad*, lack the use of basic aid translation tools like Translation Memories (TM) or Terminology Data Bases (TDB). These two kinds of tools are especially helpful for recycling old translations, and in software localisation that is something to take into account as new versions might not differ significantly from the old ones. The TDBs and TMs are also very helpful for maintaining terminology consistency during the project. This is especially important in ST as many heads might be involved in the same project.
   Another possible solution would be to link the content to related or similar content in other languages. We are living in a multilingual society and users might know other languages that are not in their initial search of content.

c) **"Trolls", vandalism and bias.** There have always been people who enjoy destroying others' work. If users let everybody modify, translate, and localise their content, they also let people with destructive mentalities do the same. Trolls' job is upsetting people, i.e. they enter blogs and forums and say something to create an inflammatory response; sometimes even insult people. Most of the blogs and forums are now aware of this practice and warn their users not to 'feed the trolls' and ignore them. Acts of cyber vandalism destroy others' work; we see this everyday in Wikipedia [8]. The bias phenomenon is in our own nature as human beings, as we think and act from our own perspectives and beliefs. Bias can even be seen in the work of professional translators, because almost every word in a language has a connotation; this means that even unconsciously they might be expressing their bias through their work. We still have to try to avoid this phenomenon in our work and production as much as possible. A good way of doing so is by letting others assess our work and control it. As mentioned before, we think one good way of checking the work of others is by a voting system, or even discussion pages like the ones present in *Wikipedia*.

4.2.2 Importance of ST for Web 2.0 creators

Web 2.0 creators and developers want users to participate in ST and there are several reasons for this eagerness:

a) **Reduction of cost:** Localisation processes are one of the major costs that an international company has to assume. If they have only had to worry about creating a platform to enable users to translate their content, they will be saving large amounts of money and spare the cost intensive traditional translation workflow.

b) **Conquering more markets**. There are small markets that may not be commercially viable if the localisation process is too expensive. However, if this process becomes free for companies (or for less cost), they might have thousands of new customers waiting for their products.

c) **Personalising translation**: If the vendors speak the same language as their customer, then the latter will look positively at the former's company or product. Consequently, the trust level as well as the benefits will arise.

d) **Loyalty**: If the users are allowed to change the content and adjust it according to their own preferences, they might 'love' the product and everything related to it. This creates a special brotherhood between the company and the users, it arouses though fanaticism. We can see this phenomenon with Linux users (they would defend Linux philosophy, would create Linux contents, translate them, etc.). In fact, nowadays if a company creates fans within their users, they are also creating loyal clients that would buy whatever they produce.

5. PROPOSED MODEL

With this proposed model we combine the technology of MT and the trend of ST. Our proposed model is to combine both MT and ST in an effective way in order to balance cost savings and ensure good quality. First of all, in our opinion, not the whole UGC should be translated. As the translation workflow includes many time-consuming and cost intensive complicated steps, it should be clear for which reason UGC has to be translated. Given that UGC is to be translated, our proposed model is to send it to an MT system and then have it proofread by volunteer proofreaders. In other words, instead of volunteer social translators, we have volunteer social proofreaders. These volunteer proofreaders should certainly be carefully selected, managed, and controlled.

On one hand, MT has potential, its quality increases over the years, and it starts to be adopted by more and more enterprises. The existence of free and open-source MT systems today proves that there is high competition and not many MT monopolies. On the other hand, the social translation, being human translation, is of high quality, and thus many enterprises implement social translation platforms. However, there are some challenges included (described in 4.2.1) which have to be overcome. According to our opinion, MT and ST, being both successful productivity models, both having advantages and disadvantages, should collaborate in order to have the desired result. To mention only a few characteristics of this collaboration, MT gives the speed, while ST the quality; MT covers the terminology consistency, while ST the multi-language support, and so on. The design of our proposed model is in the pipeline and more detailed information will follow in other papers.

6. SUMMARY

To sum up, localising Web 2.0 content will require a fundamental change in how we view the whole localisation process. New ideas will be needed and new processes will have to be implemented. Giving power to the users will lead to a more heterogeneous body of digital content with new technical problems to be solved. Social and Machine Translation can be the response to the new scenario, however they also bring with them shortcomings which should be taken into account. Thus the collaboration of them, i.e. sending the UGC to MT and having it proofread by social proofreaders will balance the shortcomings.

7. ACKNOWLEDGMENT

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at the University of Limerick.

We gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center SFB/TR 8 Spatial Cognition - Subproject I3-SharC.

8. REFERENCES

[1] O´Reilly, T. 2005. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *SSRN eLibrary.*

[2] Howe, J. 2006. Crowdsourcing: A definition. [*Weblog*], available: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html (consulted 01.03.10).

[3] Keen, A. 2008. *The Cult of the Amateur: How blogs, MySpace, YouTube, and the rest of today's user-generated media are destroying our economy, our culture, and our values*. 1st ed., Doubleday Business.

[4] Netcraft: http://news.netcraft.com/archives/2010/01/07/january_2010_web_server_survey.html (consulted 15.03.11).

[5] Hennig, B. 2009. Fehler, Fakes und Fälschungen. Kann man Nachschlagewerken alles glauben? *Texte für Technik*, Ausgabe Herbst 2009, 18-19.

[6] Garcia, I. 2009. Beyond Translation Memory: Computers and the Professional Translator. *The Journal of Specialised Translation*, Issue 12-July 2009, 199-214.

[7] Anderman, G. and Rogers, M. (Eds.) 2003. *Translation Today, Trends and Perspectives*. Multilingual Matters Ltd.

[8] Tapscott, D. and Williams, A.D. 2006. *Wikinomics: How Mass Collaboration Changes Everything*. The Cromwell Press.

[9] Kuznetsov, S. 2006. Motivations of contributors to Wikipedia. *ACM SIGCAS Computers and Society*, 36.