Collaboratively Building Language Resources while Localising the Web

Asanka Wasala, Reinhard Schäler, Ruvan Weerasinghe* and Chris Exton
Centre for Next Generation Localisation/Localisation Research Centre
CSIS Department, University of Limerick, Limerick, Ireland
*University of Colombo School of Computing, 35, Reid Avenue, Colombo 00700, Sri Lanka
{Asanka.Wasala, Reinhard.Schaler, Chris.Exton}@ul.ie,
*arw@ucsc.cmb.ac.lk

Abstract

In this paper, we propose the collaborative construction of language resources (translation memories) using a novel browser extension-based client-server architecture that allows translation (or 'localisation') of web content capturing and aligning source and target content produced by the 'power of the crowd'. The architectural approach chosen enables collaborative, in-context, and real-time localisation of web content supported by the crowd and high-quality language resources. To the best of our knowledge, this is the only practical web content localisation methodology currently being proposed that incorporates the collaborative construction and use of TMs. The approach also supports the building of resources such as parallel corpora – resources that are still not available for many, and especially not for underserved languages.

1 Introduction

A vast amount of knowledge is available on the web, primarily in English. There are millions of people worldwide, who cannot assimilate this knowledge mainly due the language service barrier. Although English is still dominating the web, the situation is changing. Non-English content is growing rapidly (Large and Moukdad, 2000; Daniel Brandon, 2001; Wasala and Weerasinghe, 2008).

Localisation is the translation and adaptation of digital content. Localisation of a website involves "translating text, content and adjusting graphical and visual elements, content and examples to make them culturally appropriate" (Stengers et al., 2004). However, the scope of our research is limited to the translation of text, which is arguably the most crucial component of web content localisation.

The study of web content localisation is a relatively new field within academia (Jiménez-Crespo, 2011). The only reported approaches to website localisation are human (Daniel Brandon, 2001) and machine-based translation (Large and Moukdad, 2000; Daniel Brandon, 2001; Wasala and Weerasinghe, 2008), with only very basic collaborative (Horvat, 2012) or first in-context approaches (Boxma, 2012) attempted. Although researchers have reported on the use of Machine Translation (MT) in web content localisation (Gaspari, 2007), the low quality of the MT-based website translation solutions is known to have been a significant drawback (Large and Moukdad, 2000; Daniel Brandon, 2001). Moreover, the research and development of MT systems for less-resourced languages is still in its infancy (Wasala and Weerasinghe, 2008). Therefore, MT-based web content localisation solutions are clearly not viable for less-resourced languages.

Undoubtedly, Web 2.0 and the constant increase of User Generated Content (UGC) lead to a higher demand for translation. The trend of crowdsourcing/social translation came into play only in the last few years. In this paper, we focus on crowdsourcing translation, i.e. when the crowd or a motivated part of it, participates in an open call to translate some content, creating highly valuable language resources in the process.

Browser extensions enhance the functionality of web browsers. Various browser extensions already exist that are capable of utilising existing Machine Translation (MT) services to translate web content into different languages. We exploit the power of

Proceedings of the 3rd Workshop on the People's Web Meets NLP, ACL 2012, pages 15–19,
Jeju, Republic of Korea, 8-14 July 2012.

browser extensions to design a conceptual localization layer for the web. Our research is mainly inspired by the works of Exton et al. (2009) on real-time localisation of desktop software using the crowd, Wasala and Weerasngihe (2008) on browser based pop-up dictionary extension, and Schäler on information sharing across languages (2012a) as well as social localisation (2012b).

The proposed architecture enables in-context real-time localisation of web content by communities sharing not just their content but also their language skills. The ultimate aim of this work is the collaborative creation of TMs which will allow for the automatic translation of web content based on reviewed and quality-checked, human produced translations. To the best of the authors' knowledge, this is the first effort of its kind to utilise the power of browser extensions along with TMs to build a website independent conceptual localisation layer with the aid of crowdsourcing.

The rest of the paper is organized as follows: Section 2 describes the architecture of the proposed system in detail; the development of the prototype is discussed in section 3; section 4 discusses key outstanding challenges and constraints of the proposed architecture; and finally, this paper concludes with a summary and discussion of future research directions.

2 System Architecture

In this section, the main functionalities of the proposed system architecture are described in detail.

The proposed system architecture is based on earlier work by Exton et al. (2009). They proposed a client-server architecture known as Update-Log-Daemon (UpLoD) for the localisation of applications' User Interface (UI) by the crowd. However, in our architecture, clients (browsers) connect to the central server via a browser extension. The browser extension implements the UpLoD architecture, which acts as a proxy between the browser and the central server.

We also extend the functionality of the central server in this architecture by equipping it with a component to maintain TMs for different language pairs.

2.1 Content Retrieval and Rendering Process

When the browser extension is installed and enabled, it allows a user to select the preferred locale. When a new URL is typed in, the browser will download the page. As soon as the content is downloaded, the browser extension will consult the central server for any TM matches in the user's preferred locale for the relevant URL. The TM matches will be retrieved with the contextual information. The next step is to replace the original content with the retrieved TM matches. With the aid of contextual hints that it received, the TM matches (i.e. target strings) will be replaced with the source strings. Finally, the content will be rendered in the browser. The contextual information may include: URL, last update date/time stamp, surrounding text with and without tags, XPath location of the segment, CSS properties among others as this information will helpful to precisely locate HTML elements in a web page (Selenium 2012). For replacing the original text with target strings, techniques such as Regular-expressions matching and XPath queries may be utilized.

2.2 Content Translation Process

The browser extension also facilitates the in-context translation of source content. Right clicking on a selected text will bring up a contextual menu where a "Translate" sub-menu can be found.

The extension allows in-context translation of the selected content segment in an editing environment similar to Wikipedia. Once the translation is completed, the extension sends the translated segment, original content and contextual information including URL to the central sever. Upon receiving translations from a client, the central server stores all the information that it retrieves in a TM.

The central server can be scheduled to periodically leverage translations as the TMs grow. Furthermore, later on, MT systems can be trained from the TM data and these trained MT systems can feed back into the system to speed up the translation process as well as to translate the content where TM matches are not found.

2.3 Translation Editing and Voting Process

As in the case of software localisation (Exton et al., 2009), a mechanism has to be built to choose the most appropriate translation of a given text segment. To assist in selecting the best translation for a given segment, a voting mechanism is proposed.

However, human intervention (mainly the opinions of experts) is essential to solve potential conflicts.

Right clicking on a translated segment brings up a context menu, where the current translation along with the top 3 alternative translations is displayed. The votes for each translation will also be displayed next to the translation. The users are given the opportunity to edit the current translation and/or to vote any of the alternative translations.

Furthermore, clicking on an alternative translation will take the user to a web page where the user can see all the alternative translations that are proposed for the selected segment. In that page users can vote for any of the alternate translations.

Considering the motivation factors related to crowdsourcing, a simple "thumbs up, thumbs down" voting is proposed over complex and confusing rating systems. If the user wishes to edit the existing translation, they can simply go to the in-context edit mode and edit the content. Once editing has been performed, the new translation is sent back to the central server. The central server compares the new changes with the existing translations and includes it as an alternative translation.

The central server needs to keep track of the votes as well as the voters. By keeping track of voters, users can be encouraged to vote for additional translations using ranking systems similar to those implemented in games.

3 Development of the Prototype

To test the above architecture, we developed a prototype with the aid of two open source Firefox Add-ons:

1. *Ingiya* – a pop-up dictionary Firefox add-on similar to the add-on described by Wasala and Weerasinghe (2008);

2. *FoxReplace* – a Firefox add-on that can automatically replace textual content with the aid of a predefined substitution list.

Ingiya, a non-intrusive add-on, shows Sinhala definitions of English terms when the mouse pointer is hovered on top of English words in a web site. It is also capable of temporarily replacing Sinhala definitions with English words (i.e. as soon as the page is refreshed, the translations disappear). Currently, the Ingiya add-on only supports individual words. The dictionary entries are stored within a local database.

The add-on was first modified to support phrases (selected text segments) in addition to individual words and to be able to collect translations for a selected phrase from the user. We submitted the selected text segment, user's translations and the URL of the active tab of the browser via Ingiya add-on to the central server as a RESTful call. We encoded the above data using the *Punycode* algorithm prior to submission.

We then implemented the central server using PHP. In this prototype, the server mainly performs three functions: 1) It accepts data sent via browser add-ons, decode the data and stores in it's local database 2) Upon a request from a client, it transforms and sends the data in its local database into a XML based format understood by FoxReplace add-don, 3) It can transform and sends data in it's local database into an XML Localisation Interchange File Format (XLIFF) file that can be used as a TM.

The FoxReplace add-on is capable of retrieving a regular expression-based source and target substitution list encoded in a specific XML format and replacing text in a web page. Different substitutions can be defined for different URLs. The FoxReplace add-on was configured to retrieve translations (i.e. substitution list) from the central server. When combined, these two add-ons along with the central server are able to implement and demonstrate the UpLoD architecture described in the pervious section. The exception is the voting mechanism which has not yet been implemented but is part of on-going work by the research group.

4 Discussion: Outstanding Challenges

While most of the issues and challenges emphasised in the UpLoD-based architecture (Exton et al., 2009) are common to the architecture proposed in this article, web content localisation also faces additional, unique technical challenges.

Web pages consist of not only text, but also non-textual content such as images, audio clips, videos and various embedded objects (e.g. Java, Flash, PDF or Silverlight content) (Daniel Brandon 2001; Stengers et al., 2004). Textual content represented in graphics such as banners is also very common in web sites. The current architecture however does not deal with localisation of non-textual content found in websites. Even with the

textual content, font and rendering problems may surface in the localised version.

Another issue that can occur in a crowdsourced localisation model as noted by Exton et al. (2009) is the primary focus on translation of the frequently used content by the crowd. This issue is likely to surface in the web content localisation scenario as well. It will result in untranslated content of infrequently visited sections of the web sites.

Issues related to translation voting, especially the 'thrashing' scenarios as described by Exton et al. (2009) need to be addressed in this scenario too. The optimum human translation rating mechanisms, as well as motivations for rating these, have to be explored further.

Another important factor is the design of a methodology for coping with constant updates of websites. We would expect that a large TM might help to alleviate the above problem to a certain degree.

One of the advantages of the above methodology is that, once the entire web page is completely translated, the translated page can be cached in the central server for improved performance. On the other hand, the localisation layer is only accessible via the browser extension. Therefore, users are not able to interact with the website using their native language, nor would these pages be indexed by search engines (i.e. the localised version).

In addition to various technical issues discussed above, legal issues could potentially be encountered which need to be thoroughly examined, identified and addressed prior to the deployment of the proposed solution. The first question that needs to be answered is if people have a right to localise websites without the consent of the web site owners. Moreover, the TMs (for each language pair) will keep on growing once the crowd starts using this framework. Legal implications around the TMs have to be thoroughly considered. For example, questions such as who owns the TMs needs to be addressed.

The accuracy of the translations is one of the crucial aspects that need to be considered. It is essential to investigate necessary steps to prevent possible misuse. Misuse of the service can be alleviated to a certain extent by developing a log-on mechanism where users have to be authenticated by the central server to access the localisation service. Furthermore, individuals who contribute translations as well as individuals who vote for translations can be tracked and rewarded. Thus, these individuals can be further motivated with the use of public announcements and ranking (or medal offering) systems as in games.

Website localisation is not just the translation of text in a website. Various ethical, cultural and regional issues have to be taken into account when localising a website. Therefore, a reviewing mechanism such as observed in the Wikipedia community has to be built in to this model.

5 Conclusions and Future Work

In this paper, we have discussed the development of a browser extension-based website independent client-server architecture that facilitates the collaborative creation of TMs used for the localisation of web content. As this approach uses TMs constructed with the aid of the crowd and reviewed by experts where necessary, rather than an MT system, better quality translations can be expected. The development of the prototype has proven the viability of the proposed approach. Future research will focus mainly on addressing the issues related to central server services discussed above. Moreover, the development of a (single) Firefox add-on encompassing all the functionalities described in section 3 has already shown good results.

To the best of our knowledge, this is the only practical web content localisation approach proposed which is based on the collaborative construction of TMs utilising the power of browser extensions combined with micro-crowdsourcing. The current architecture will be especially useful in the case of less-resourced languages where MT systems are not (yet) viable. The proposed system focuses on the building of language resources, such as translation memories but also parallel corpora, which could be used for the development of MT systems in the future.

Acknowledgments

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. The prototype was implemented based on Ingiya and FoxReplace add-ons. The authors would like to thank the authors of and the contributors to the above add-ons.

References

Boxma, H. (2012). RIGI Localization Solutions Retrieved April 01, 2012, from https://sites.google.com/a/rigi-ls.com/www/home

Daniel Brandon, J. (2001). Localization of web content. *J. Comput. Small Coll., 17*(2), 345-358.

Exton, C., Wasala, A., Buckley, J., & Schäler, R. (2009). Micro Crowdsourcing: A new Model for Software Localisation. *Localisation Focus, 8*(1), 81-89.

Gaspari, F. (2007). *The Role of Online MT in Webpage Translation.* Doctor of Philosophy, University of Manchester, Manchester, Retrieved June 28, 2011, from http://www.localisation.ie/resources/Awards/Theses/F_Gaspari_Thesis.pdf

Horvat, M. (2012). *Live Website Localization*. W3C Workshop: The Multilingual Web – The Way Ahead, Luxembourg, Retrieved April 01, 2012, from http://mozeg.com/pontoon-mlw.html

Jiménez-Crespo, M. A. (2011). To adapt or not to adapt in web localization: a contrastive genre-based study of original and localised legal sections in corporate websites. *JoSTrans (The Journal of Special Translation)*(15).

Large, A., & Moukdad, H. (2000). Multilingual access to web resources: an overview. *Program: Electronic Library and Information Systems, 34*(1), 43 - 58. doi: 10.1108/EUM0000000006938

Schäler, R. (2012a). Information Sharing Across Languages Computer-Mediated Communication across Cultures: International Interactions in Online Environments (pp. 215-234): IGI Global.

Schäler, R. (2012b). Introducing *Social Localisation*. Workshop. Localization World, Silicon Valley,. Retrieved April 02, 2012 from http://www.slideshare.net/TheRosettaFound/social-localisation

Selenium Project. (2012). Selenium-IDE - Locating Elements. Retreived May 14, 2012 from: http://seleniumhq.org/docs/02_selenium_ide.html#locating-elements

Stengers, H., Troyer, O. D., Baetens, M., Boers, F., & Mushtaha, A. N. (2004). *Localization of Web Sites: Is there still a need for it?* Paper presented at the International Workshop on Web Engineering (held in conjunction with the ACM HyperText 2004 Conference), Santa Cruz, USA.

Wasala, A., & Weerasinghe, R. (2008). *EnSiTip: A Tool to Unlock the English Web*. Paper presented at the 11th International Conference on Humans and Computers, Nagaoka University of Technology, Japan.