Exploring Linguistic Points of View of Wikipedia

Paolo Massa
Bruno Kessler Foundation
Via Sommarive, 14
38123 Trento - Italy
massa@fbk.eu

Federico Scrinzi
Bruno Kessler Foundation
Via Sommarive, 14
38123 Trento - Italy
fscrinzi@fbk.eu

ABSTRACT

The 3 million articles of the English Wikipedia has been written since 2011 by more than 14 million volunteers. On each article, the community of editors strive to reach a neutral point of view, representing all significant views fairly, proportionately, and without bias. However, beside the English one, there are more than 270 Wikipedias in different languages and their relatively isolated communities of editors are not forced by the platform to discuss and negotiate their points of view. So the empirical question is: do communities on different languages editions of Wikipedia develop their own diverse Linguistic Points of View (LPOV)? To answer this question we created Manypedia, a web tool whose goal is to ease cross-cultural comparisons of Wikipedia language communities by analyzing their different representations of the same topic.

Keywords

Wikipedia, Cross-cultural, Linguistic Point of View, Multilingual

1. NEUTRAL POINT OF VIEW AND LANGUAGE EDITIONS OF WIKIPEDIA

Wikipedia, the online encyclopedia anyone can edit, is becoming the largely most accessed Web resource for information needs such that, for example, 53% of American Internet users look for information on it as of May 2010 [1].

The English Wikipedia has received more than 453 million edits by more than 14 million registered users since 2001. The community of editors who self-elect for editing Wikipedia pages strive to follow a Neutral Point of View (NPOV). This is one of the three core content policies and, as Roy Rosenzweig puts it in [2], the "founding myth" of Wikipedia. NPOV requires editors to write articles representing all significant views fairly, proportionately, and without bias. Of course, writing "without bias is difficult" since "all articles are edited by people" and "people are inherently biased" [2]. This is evidenced also by the frequent edit wars [3] occurring when two or more users disagreeing about the content of a page repeatedly override each other's contributions, rather than trying to resolve the disagreement by discussion in the associated talk page. In fact, Rosenzweig argues that the most frequent debate topic on talk pages is whether the article adheres to the NPOV and what are the major and minor points of view and how much relative prominence they should receive [2].

Some people are skeptical that neutrality can be reached. For example, one the two founders of Wikipedia, Larry Sanger, who quit Wikipedia, argues that "over the long term, the quality of a given Wikipedia article will do a random walk around the highest level of quality permitted by the most persistent and aggressive people who follow an article" [4]. Beside being optimistic or pessimistic about the fate of Wikipedia in the long run, we believe the simple act of discussing is central to democracy, an healthy global society and peaceful coexistence of different points of view. Rosenzweig considers that "those who create Wikipedia's articles and debate their contents are involved in an astonishingly intense and widespread process of democratic self-education" and reports that the classicist James O'Donnell has argued that the benefit of Wikipedia may be greater for its active participants than for its readers: "A community that finds a way to talk in this way is creating education and online discourse at a higher level" [2].

WikiSym '11, Oct 03-05 2011, Mountain View, CA, USA
ACM 978-1-4503-0909-7/11/10.

However, while the English Wikipedia was the first to be created, as of August 2011 there are more than 270 different language editions of Wikipedia, ranging from many with more than 700,000 articles such as the German, French, Polish, Italian, Japanese and Spanish ones, up to smaller ones in languages such as Wolof, Catalan, Latin, Esperanto, Tibetan, Haitian and more. The editors of each language edition of Wikipedia strive to reach a NPOV but the current Wikipedia socio-technical platform does not provide many opportunities for editors of different language Wikipedias to discuss and share points of view. Different language Wikipedias are quite isolated and connected mainly by interwiki links. In fact it is possible to link the article about, for example, "Palestine" in the Hebrew Wikipedia with its equivalent in Arabic Wikipedia simply by inserting an interwiki link of the form [[language code:Title]], for example [[ar:فلسطين العربية]]. The Wikipedia server interprets this interwiki syntax and offers links to the equivalent page in the other language Wikipedia on the left hand side of each Wikipedia page under a "Language" menu. These interwiki links must be inserted manually (or with the help of semi-automated programs called bots) by users who, at least in theory, know both the source and target language.

Given this relative isolation of communities of editors of each language edition of Wikipedia, we are interested in investigating if and how much different communities will develop their own divergent representations for the same topic. This is what we name Linguistic Point of View (LPOV). For example, are the pages "Palestine" in the English, Arabic and Hebrew Wikipedias largely similar satisfying the global consensus hypothesis introduced in [5] or do the three representations crystallize strong cultural differences of the different communities of editors?

2. MANYPEDIA WEB MASHUP

There are some recent studies which compare different language Wikipedias from a cross-cultural perspective [6]. However in general knowledge of the involved languages is required. In order to make it easier to conduct cross-cultural studies, we have created Manypedia (accessible at http://www.manypedia.com) .

Manypedia allows to search for a page on every language Wikipedia and to compare it with the equivalent page on another language Wikipedia. It exploits the Google Translate API in order to automatically translate the second page in the language of the first page (see Figure 1 for an example). Currently 56 languages are supported in translation both as source and target language. On top of embedded Wikipedia pages, Manypedia shows

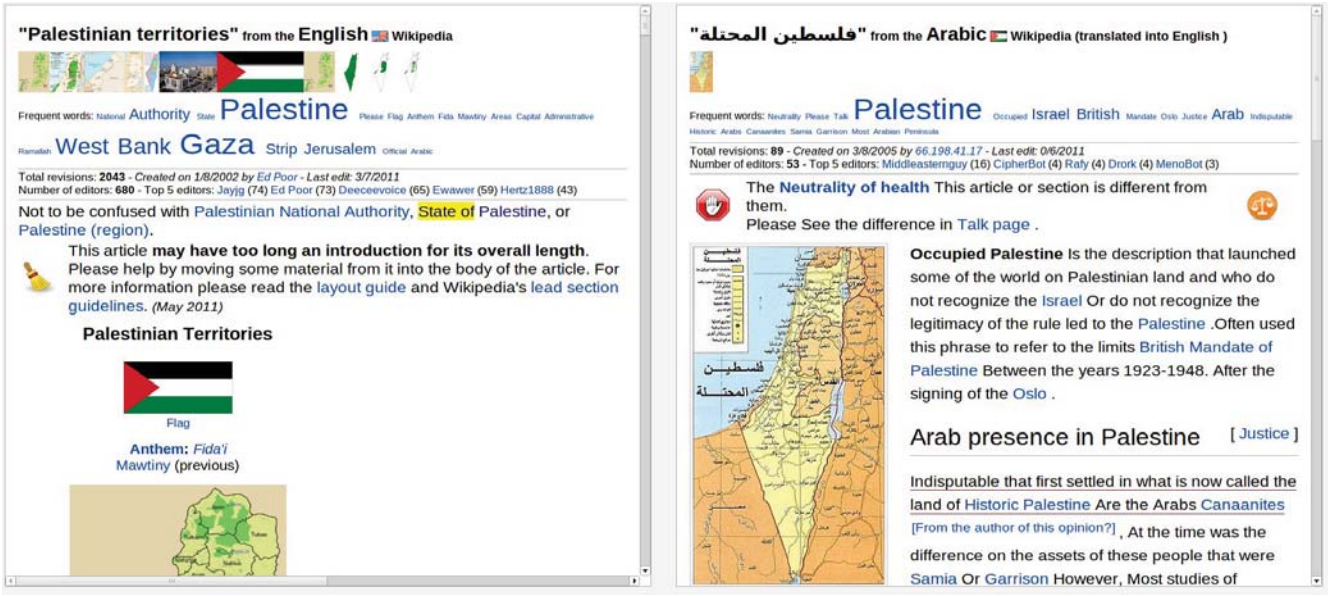

Figure 1: Screenshot of Manypedia comparing "Palestinian territories" page on English and Arabic Wikipedia.

information which can help in forming a first impression about the differences of the knowledge products created by the two language communities. Images included in the two articles allow to get a first visual understanding of the points of view represented. Most frequent words, number of total edits and editors, creation date, last edit date and top editors for both pages are included as well in order to summarize salient characteristics about the pages and the editing process behind them.

3. DISCUSSION

The purpose of Manypedia is to make it easier to conduct cross-cultural investigations about the possible different points of views represented by different language communities of Wikipedia. On top of Manypedia interface there are links to featured comparisons which can constitute an useful starting point considering also that links in both embedded pages are transformed into links to comparisons. For instance, the Wikipedia page "List of controversial articles" is an interesting starting point since it contains links to pages which are considered problematic by each community. As an example, the Chinese page contains pages such as "Anti-Japanese War", "Nanjing Massacre", "Taiwan", "Human Rights in China", "Falun Gong", "Tiananmen Incident" while the Catalan page refers predominantly to issues about the term "country" and "region" and the concept of Catalan country itself.

Other examples of pages which might exhibit different Linguistic Points of View are those related to recent history, such as "Osama Bin Laden", and ongoing struggles such as the "Israeli–Palestinian conflict" or contested territories such as "Northern Cyprus" which might be treated differently, for example, in the Greek and Turkish Wikipedias. As Rosenzweig puts it, "like journalism, Wikipedia offers a first draft of history, but unlike journalism's draft, that history is subject to continuous revision. Wikipedia's ease of revision not only makes it more up-to-date than a traditional encyclopedia, it also gives it (like the Web itself) a self-healing quality since defects that are criticized can be quickly remedied and alternative perspectives can be instantly added" [2]. In fact research on the formation of collective memories of recent events exploits this feature of Wikipedia for which recent events tend to get created few minutes or hours after they happen and the community strives to fairly represent them as they unfold [7] but different communities might have different LPOVs especially during the first weeks after the event when the collective memory is still communicative [8].

Concluding, we agree with Hecht et al. when they state "researchers must be aware that sub-concept diversity does not simply represent information "inefficiencies" that need to be fixed" and suggests the "the potential for culturally-aware applications is enormous" [5]. We hope Manypedia can be a useful first step allowing everyone to assess the current situation in terms of Linguistic Points of View on Wikipedia and possibly offer opportunities of discussions and dialog between different language communities of Wikipedia.

4. REFERENCES

[1] Pew Research Center. 2011. Wikipedia, past and present. A snapshot of current Wikipedia users. Jan 13, 2011. Retrieved on April 4, 2011, from http://www.pewinternet.org/Reports/2011/Wikipedia.aspx

[2] Rosenzweig, R. 2006. Can History Be Open Source? Wikipedia and the Future of the Past. The Journal of American History 93(1), 117-146.

[3] Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. 2007. He says, she says: Conflict and coordination in Wikipedia. Proceedings of CHI 2007, 453-462.

[4] Sanger, L. M. 2009. The Fate of Expertise after Wikipedia. Episteme. 6, 52-73.

[5] Hecht, B. and Gergle, D. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and its Applications in a Multilingual Context. Proceedings of CHI 2010, 291-300.

[6] Massa, P. 2011. Cross-cultural Studies of Wikipedia. Retrieved on July 4, 2011, from http://www.gnuband.org/2011/04/08/cross-cultural_studies_of_wikipedia_/

[7] Ferron, M., and Massa, P. 2011. Collective memory building in Wikipedia: the case of North African uprisings. Proceedings of WikiSym 2011.

[8] Ferron, M., and Massa, P. 2011. Studying Collective Memories in Wikipedia. Proceedings of 3rd Digital Memories Conference. Prague.