WikiBABEL: A System for Multilingual Wikipedia Content

A. Kumaran[§], Naren Datha[§], B. Ashok[§], K. Saravanan[§], Anil Ande[§], Ashwani Sharma[§], Sridhar Vedantham[§], Vidya Natampally[§], Vikram Dendi[§§] & Sandor Maurice[§§]

[§]Multilingual Systems Research
Microsoft Research India
{a.kumaran,narend,bash,v-sarak,v-anila,
ashwanis,sriv,vidya}@microsoft.com

[§§]Machine Translation Incubation
Microsoft Research
{vikramde, samaur}
@microsoft.com

Abstract

This position paper outlines our project – WikiBABEL – which will be released as an open source project for the creation of multilingual Wikipedia content, and has potential to produce parallel data as a by-product for Machine Translation systems research. We discuss its architecture, functionality and the user-experience components, and briefly present an analysis that emphasizes the resonance that the WikiBABEL design and the planned involvement with Wikipedia has with the open source communities in general and Wikipedians in particular.

1 Introduction[*]

State-of-the-art Computational Linguistics research and practical Natural Language Processing (NLP) systems critically depend on availability of appropriate data. While specialized data require professional involvement, much of the NLP research and systems may leverage data created by speakers of a given language. For example, parallel corpora required for Machine Translation (MT) systems may be created even by those who are fluent in two languages and not necessarily linguistics or language experts, as the current MT systems are shown to be robust even in the presence of significant noise in the data [Quirk et al., 2007]. It is well known that the largest non-expert contributed knowledge-base, the Wikipedia, is widely used in academic research and in many practical systems, as evidenced by increasing number of publications based on the Wikipedia data [Wikipedia, 2010a, 2010b, 2010c]. In our research, we explore a community collaborative frameworks that may be adopted for creating language data through the Wikipedia eco-system. While specific Wikipedia communities create content for their Wikipedia leveraging the content in English Wikipedia, as a by-product, parallel data may be created for aiding Machine Translation research between English and the specific language. We devised such a co-creation methodology based on three important observations: (i) the large skew that exists among the content of Wikipedia across languages[1], (ii) the availability of imperfect but high-throughput machine translation systems, and (iii) the aspiration of many Wikipedia communities to improve their Wikipedia content.

In this position paper, we present a wiki-style framework – WikiBABEL [Kumaran et al., 2008] [Kumaran et al., 2009] that enables a community to source content from other Wikipedias, translate using an in-development MT system and correct and contribute to their Wikipedia. Our hypothesis is that such a methodology allows even non-expert native speakers of a language to participate in the Wikipedia content creation process, and as a by-product, produce multi-domain parallel corpora that may be used for re-training the MT system. Also, such a methodology may signal a powerful paradigm for creating MT systems in many resource-poor languages of the world.

[*] Presented in the special track on "*Collaborative Translation: Technology, Crowdsourcing and the translator perspective*" of the *American Machine Translation Association* (*AMTA*) Workshop, held in October 2010.

[1] English Wikipedia has about 3+ Million articles, and by far the largest of the 273 Wikipedias. While the English content is below the collective content in all Wikipedias, still there exists a huge skew in the content. For example, the 32 Wikipedias with 100K articles or more each have about 14M articles collectively, while the bottom 240+ Wikipedias with less than 100K articles each have about 2M articles collectively [Wikipedia, 2010d].

2 Architecture of WikiBABEL

WikiBABEL is architected as a browser based application that may be deployed on any wiki-site, and the current beta version is architected specifically for Wikipedia. Figure 1 outlines the functional architecture and the components.

Fig 1: Functional Architecture of WikiBABEL

The salient of WikiBABEL are:

1. Scenario independent WikiBABEL core that manages the users, content and the workflow; the current version is focused on Wikipedia.
2. Workflow that is focused on target content creation and not just on translation.
3. On-site design that stays on Wikipedia for the entire session, integrating services into the workflow. At the end of the session, any content created by the user gets posted on Wikipedia.
4. Vendor-neutral methodology for integrating linguistic services into the workflow.
5. Cloud-based collaborative mechanism that enables sharing of translation knowledge instantly.

2.1 User Workflow & Interface

Figure 2 shows the basic window of WikiBABEL that captures the workflow as well. The user workflow consists of 3 steps. An intuitive user interface layout and tools were designed for each step focusing the users on (i) sourcing of appropriate material for contribution to target article, (ii) compiling the target article with sourced or new material, and, (iii) submitting the compiled article to the target multilingual Wikipedia. We outline the focus and workflow in each of the steps below:

Collect Step: This step focuses user on identifying appropriate content for enhancing target language Wikipedia article, translating the sourced content from English to the target language and providing appropriate mechanisms for correcting the translations. This step displays two side-by-side panes, in which the English Wikipedia article (that is inter-wiki-linked to the target language article of interest) and its machine translated version are presented. The user has the choice of correcting the translated content in the right pane, either by directly editing it or through a community-shared mechanism (refer to Section 2.2). The user may also add new material without any restrictions to conform to the original English content presented in the left pane. Given that parallel data may be mined efficiently from comparable corpora [Quirk et al., 2007], such latitude in creation of multilingual content is expected to satisfy the Wikipedia community, in addition to potentially yielding significant parallel data. User may also explore English and target language Wikipedia for other related content through a search mechanism with all the visited topics available persistently through the user-session.

Compose Step: In this step, the user focuses on composing the target language Wikipedia article, with the content sourced from potentially multiple articles in the collect step, and/or with new content. The UI of this step also sports 2-panes – the left-pane always displaying one of the visited articles, in its translated and corrected form (which is same as the right-side pane of the previous step), and the right-pane, the target language article being created or enhanced. The user may move content from left pane to the right pane, and further modify or add new content in order to shape the contribution to the target Wikipedia content, appropriately.

Submit Step: This third step puts the user directly in the edit page of the target language Wikipedia article that was being created on enhanced. The edit box is pre-populated with composed content of the right-pane of the second step. The Wikipedia editor could also be used for any additions or final edits on the article (such as, resolving any unresolved red-links or fixing references) before submission. The user can preview and submit the contribution to the target language Wikipedia, if he/she is satisfied, and the submissions are executed via the standard Wikipedia API's.

While the current user interface is tailored for Wikipedia users, it may be re-configured for the requirements of different wiki-communities, as well as for different user exposure to Wikipedia.

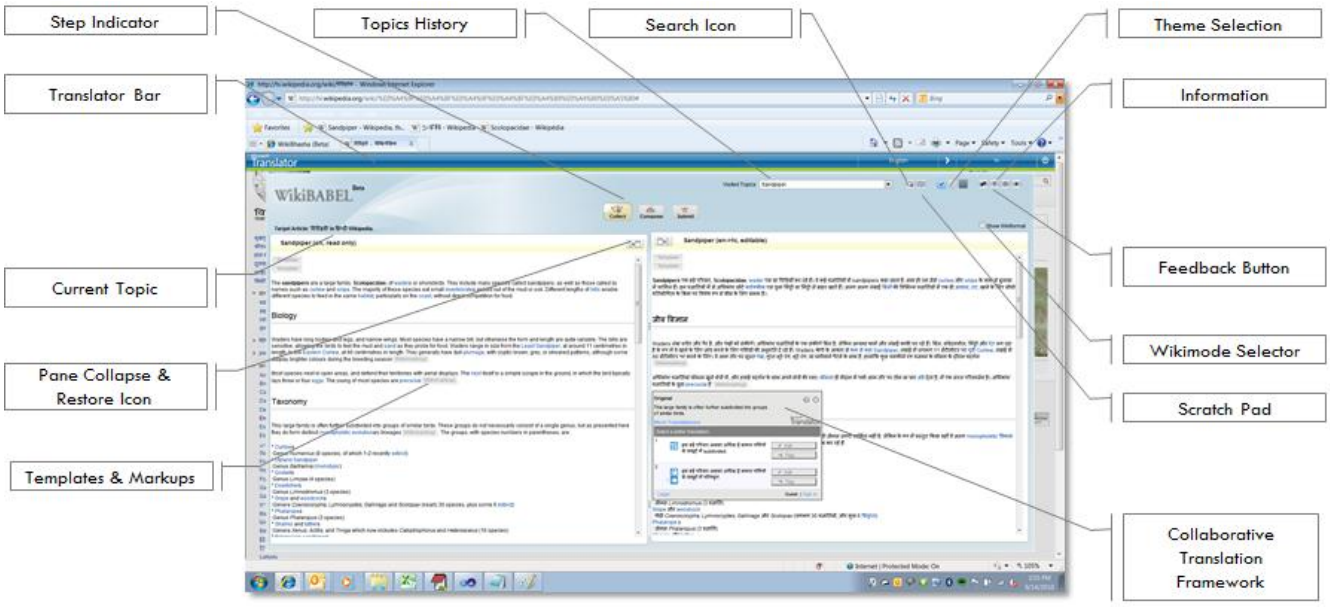

Fig 2: WikiBABEL User Interface

2.2 Linguistic and Collaborative Services

In WikiBABEL, the linguistic and collaborative services are integrated into the UI layer.

Linguistic Services: WikiBABEL integrates as default the Microsoft Translator [Microsoft Translator, 2010a] to provide translations in 31 languages. A development roadmap is provided to include other translation services as alternatives to the default translations, assuming the terms of use allow such inclusion of other translation services. While currently the MT service is the primary linguistic services integrated with the WikiBABEL core, the architecture allows integration of other linguistic services, such as dictionaries, thesauri, etc., to the content creation process.

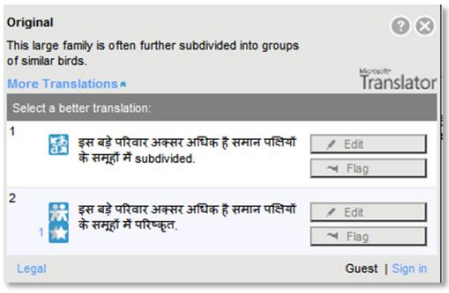

Fig 3: Collaborative Translation Framework

Collaborative Services: A cloud-based collaborative mechanism [Microsoft Collaborative Translations, 2010b] is integrated for capturing and sharing the translation knowledge instantly among the target language Wikipedia user community. Figure 3 shows the interface of collaborative translations window that displays the default translation of a given source sentence, and the history of modifications that the default translation has gone through (through user edits and corrections). The user may choose the default, or any of the alternatives to either submit to the underlying article, or modify further before submitting it, or to flag an inappropriate translation or content. Flagging inappropriate translations may be used in ranking high the good translations, which, over time may bubble up to the top as default translations. In specific sites, the administrators may have special privileges to mark specific translations as the preferred (or approved) ones, and hence presented as the default translation.

2.3 Wikipedia-specific Enhancements

Many features to help even non-Wikipedians to participate in the content creation process are included, as outlined below:

Wiki-markup: While Wikipedians are comfortable with wiki-marked up content, such could be a barrier to entry for casual users. To encourage even casual visitors to contribute content, we have included a simplified edit pane, which presents a simplified edit interface that hides away wiki-markups (but, preserving all the information in the background). The users can, optionally, switch between simplified or native modes.

Templates: Wikipedia content is enriched by the presence of templates, to present structured data as a part of an article. Most information in the templates needs to be preserved across languages, and we provide a template-mapping mechanism which can map a template and its content between two languages, using a configuration script. We hope to provide wizard-like interface for people to create such template mappings, working with the Wikipedia community.

3 WikiBABEL: An Analysis

Using crowd-sourcing as a methodology for creation/collection of parallel data is not a new concept. Previous attempts include WikiBABEL [Kumaran et al., 2008] [Kumaran et al., 2009], where a simpler system for translating and correcting translations was deployed. Controlled studies on this system indicated that such content creation methodology goes against the principles of Wikipedia communities (by having to adopt sentence-by-sentence translation). In addition, it revealed that while amateur translators were more productive with the tool, seasoned translators were slowed-down by the process, by as much as 30%.

The Google Translator Toolkit [Garcia and Stevenson, 2009] is another toolkit that attempts to create parallel data by providing a platform for translating English content (including Wikipedia articles) and correcting the translated content. Though [Ayyakkannu, R., 2010] criticized this methodology from philosophical, technical and operational viewpoints, they also emphasized two specific counterpoints: (i) that any toolkit is just that, and the responsibility of its proper use lies with respective Wikipedia communities; (ii) a grass-roots based movement is necessary to provide a voice to the community in evolving any strategy for increasing Wikipedia content.

WikiBABEL takes care of not only the criticisms of Google Translator Toolkit, but also addresses the specific recommendations – primarily in the philosophical and operational domains. WikiBABEL aids creation of content organically, using translation of English content only as one of the methods, without solely relying on it. It is also meant as a tool to be used by Wikipedians who are interested in content creation, and not by translators, and its workflow is primarily focused on target language content creation. WikiBABEL also provides mechanisms specific for Wikipedians – such as, handling wiki-markups gracefully, template translations, ability to preserve references and inter-wiki-links in translations, etc. We also hope to introduce and grow WikiBABEL organically in selected Wikipedia demographics to study its adoption, and to iron out any/all issues in its philosophy or implementation, working with the respective Wikipedia communities.

4 WikiBABEL: Status & Future Plans

WikiBABEL will be released as an open source project shortly. After open sourcing the code, we plan to engage actively with the Wikipedians over the next few months in studying its usability and usefulness in creation of multilingual content in many Wikipedias, and the adaptation patterns among different Wikipedia communities. A significant area of research is studying whether the content created through the tool satisfies the Wikipedia communities (for the accuracy, originality and expression). In parallel, we plan to collect all user correction data, as well as any new content that gets created through WikiBABEL, as potential parallel data and study its usefulness in Machine Translation research.

WikiBABEL currently supports all the languages supported by the Microsoft Translator, currently, numbering 31, in addition to English. Since Microsoft Translator is integrated as the default translation engine, WikiBABEL would automatically be able to support any additional languages supported by the engine. In addition, our development roadmap includes provisions to include other 3rd party translators, to provide alternative translations in a uniform way.

WikiBABEL has been tested on Internet Explorer (version 7.0 or above) on Windows XP, Windows Vista & Windows 7 platforms, and on Firefox (version 3.5 or above) on Windows and Linux Fedora platforms.

Acknowledgments

We wish to acknowledge significant interaction with Wikimedia Foundation and selected Wikipedia communities on WikiBABEL. We wish to thank all volunteers who participated in usability studies that resulted in the current design of WikiBABEL.

References

Ayyakkannu, R. A Review on Google Translation Project in Tamil Wikipedia: Role of voluntarism, free and organically evolved, in ensuring quality of Wikipedia. In proceedings of *WikiSYM* 2010, July 2010.

Garcia, I. and Stevenson, V. Google Translator Toolkit. Free web-based translation memory for the masses. *Multilingual*, Sept 2009.

Kumaran, A., Saravanan, K. and Maurice, S. 2008. WikiBABEL: Community Creation of Multilingual Data. In Proceedings of *WikiSYM* 2008, September 2008.

Kumaran, A., Datha, N., Saravanan, N., Dendi, V. and Maurice, S., 2009. WikiBABEL: a wiki-style platform for creation of parallel data. In Proceedings of ACL/IJCNLP-2009, August 2009.

MediaWiki. http://www.mediawiki.org/, 2010.

Microsoft. Microsoft Bing Translator. http://www.microsofttranslator.com, 2010.

Microsoft. Collaborative Translations: Announcing the next version of Microsoft Translator technology – V2 APIs and widget. http://blogs.msdn.com/b/translation/archive/2010/03/15/collaborative-translations-announcing-the-next-version-of-microsoft-translator-technology-v2-apis-and-widget.aspx. March 2010.

Quirk, C., Udupa, R. and Menezes, A. Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In Proceedings of MT Summit XI, 2007.

Wikipedia. Wikipedia as an academic source. http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_an_academic_source. Extracted in Sept 2010.

Wikipedia. Academic Studies of Wikipedia. http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia. Extracted in Sept 2010

Wikipedia. Researching Wikipedia
http://en.wikipedia.org/wiki/Wikipedia:Researching_Wikipedia. Extracted in Sept 2010.

Wikipedia. Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm. Extracted in Sept 2010.