WikiBABEL: A Wiki-style Platform for Creation of Parallel Data

A Kumaran[†] **K Saravanan**[†] **Naren Datha**[*] **B Ashok**[*] **Vikram Dendi**[‡]

[†]Multilingual Systems Research
Microsoft Research India

[*]Advanced Development & Prototyping
Microsoft Research India

[‡]Machine Translation Incubation
Microsoft Research

Abstract

In this demo, we present a wiki-style platform – WikiBABEL – that enables easy collaborative creation of multilingual content in many non-English Wikipedias, by leveraging the relatively larger and more stable content in the English Wikipedia. The platform provides an intuitive user interface that maintains the user focus on the multilingual Wikipedia content creation, by engaging search tools for easy discoverability of related English source material, and a set of linguistic and collaborative tools to make the content translation simple. We present two different usage scenarios and discuss our experience in testing them with real users. Such integrated content creation platform in Wikipedia may yield as a by-product, parallel corpora that are critical for research in statistical machine translation systems in many languages of the world.

1 Introduction

Parallel corpora are critical for research in many natural language processing systems, especially, the Statistical Machine Translation (SMT) and Crosslingual Information Retrieval (CLIR) systems, as the state-of-the-art systems are based on statistical learning principles; a typical SMT system in a pair of language requires large parallel corpora, in the order of a few million parallel sentences. Parallel corpora are traditionally created by professionals (in most cases, for business or governmental needs) and are available only in a few languages of the world. The prohibitive cost associated with creating new parallel data implied that the SMT research was restricted to only a handful of languages of the world. To make such research possible widely, it is important that innovative and inexpensive ways of creating parallel corpora are found. Our research explores such an avenue: by involving the user community in creation of parallel data.

In this demo, we present a community collaboration platform – WikiBABEL – which enables the creation of multilingual content in Wikipedia. WikiBABEL leverages two significant facts with respect to Wikipedia data: First, there is a large skew between the content of English and non-English Wikipedias. Second, while the original content creation requires subject matter experts, subsequent translations may be effectively created by people who are fluent in English and the target language. In general, we do expect the large English Wikipedia to provide source material for multilingual Wikipedias; however on specific topics specific multilingual Wikipedia may provide the source material (*http://ja.wikipedia.org/wiki/俳句* may be better than *http://en.wikipedia.org/wiki/haiku*). We leverage these facts in the WikiBABEL framework, enabling a community of interested native speakers of a language, to create content in their respective language Wikipedias. We make such content creation easy by integrating linguistic tools and resources for translation, and collaborative mechanism for storing and sharing knowledge among the users. Such methodology is expected to generate comparable data (similar, but not the same content), from which parallel data may be mined subsequently (Munteanu et al, 2005) (Quirk et al, 2007).

We present here the WikiBABEL platform, and trace its evolution through two distinct usage versions: First, as a standalone deployment providing a community of users a translation platform on hosted Wikipedia data to generate parallel corpora, and second, as a transparent edit layer on top of Wikipedias to generate comparable corpora. Both paradigms were used for user testing, to gauge the usability of the tool and the viability of the approach for content creation in multilingual Wikipedias. We discuss the implementations and our experience with each of the above scenarios. Such experience may be very valuable in fine-tuning methodologies for community creation of various types of linguistic data. Community contributed efforts may perhaps be the only way to collect sufficient corpora effectively and economically, to enable research in many resource-poor languages of the world.

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 29–32,
Suntec, Singapore, 3 August 2009.

2 Architecture of WikiBABEL

The architecture of WikiBABEL is as illustrated in Figure 1: Central to the architecture is the *WikiBABEL* component that coordinates the interaction between its linguistic and collaboration components, and the users and the Wikipedia system. WikiBABEL architecture is designed to support a host of linguistic tools and resources that may be helpful in the content creation process: *Bilingual dictionaries* for providing for word-level translations, allowing user customization of domain-specific, or even, user-specific bilingual dictionaries. Also available are *machine translation and transliteration* systems for rough initial translation [or transliteration] of a source language string at sentential/phrasal levels [or names] to the intended target language. As the quality of automatic translations are rarely close to human quality translations, the user may need to correct any such automatically translated or transliterated content, and an intuitive edit framework provides tools for such corrections. A *collaborative translation memory* component stores all the user corrections (or, sometimes, their selection from a set of alternatives) of machine translations, and makes them available to the community as a translation help ('*tribe knowledge*'). Voting mechanisms are available that may prioritize more frequently chosen alternatives as preferred suggestions for subsequent users. The *user-management* tracks the user demographic information, and their contributions (its quality and quantity) for possible recognition. The user interface features are implemented as light-weight components, requiring minimal server-side interaction. Finally, the architecture is designed open, to integrate any user-developed tools and resources easily.

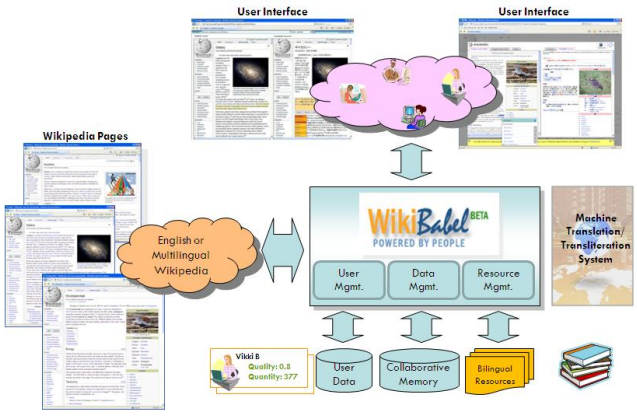

Figure 1: WikiBABEL Architecture

3 WikiBABEL on Wikipedia

IN this section we discuss Wikipedia content and user characteristics and outline our experience with the two versions on Wikipedia.

3.1 Wikipedia: User & Data Characteristics

Wikipedia content is acknowledged to be on par with the best of the professionally created resources (Giles, 2005) and is used regularly as academic reference (Rainie *et al*., 2007). However, there is a large disparity in content between English and other language Wikipedias. English Wikipedia - the largest - has about 3.5 Million topics, but with an exception of a dozen or so Western European and East Asian languages, most of the 250-odd languages have less than 1% of English Wikipedia content (Wikipedia, 2009). Such skew, despite the size of the respective user population, indicates a large room for growth in many multilingual Wikipedias. On the contribution side, Wikipedia has about 200,000 contributors (> 10 total contributions); but only about 4% of them are very active (> 100 contributions per month). The general perception that a few very active users contributed to the bulk of Wikipedia was disputed in a study (Swartz, 2006) that claims that large fraction of the content were created by those who made very few or occasional contributions that are primarily editorial in nature. It is our strategy to provide a platform for easy multilingual Wikipedia content creation that may be harvested for parallel data.

3.2 Version 1: A Hosted Portal

In our first version, a set of English Wikipedia topics (stable non-controversial articles, typically from Medicine, Healthcare, Science & Technology, Literature, etc.) were chosen and hosted in our WikiBABEL portal. Such set of articles is already available as *Featured Articles* in most Wikipedias. English Wikipedia has a set of ~1500 articles that are voted by the community as stable and well written, spanning many domains, such as, Literature, Philosophy, History, Science, Art, etc. The user can choose any of these Wikipedia topics to translate to the target language and correct the machine translation errors. Once a topic is chosen, a two-pane window is presented to the user, as shown in Figure 2, in which the original English Wikipedia article is shown in the left panel and a rough translation of the same article in the user-chosen target language is presented in the right panel. The right panel has the same look and feel as the original

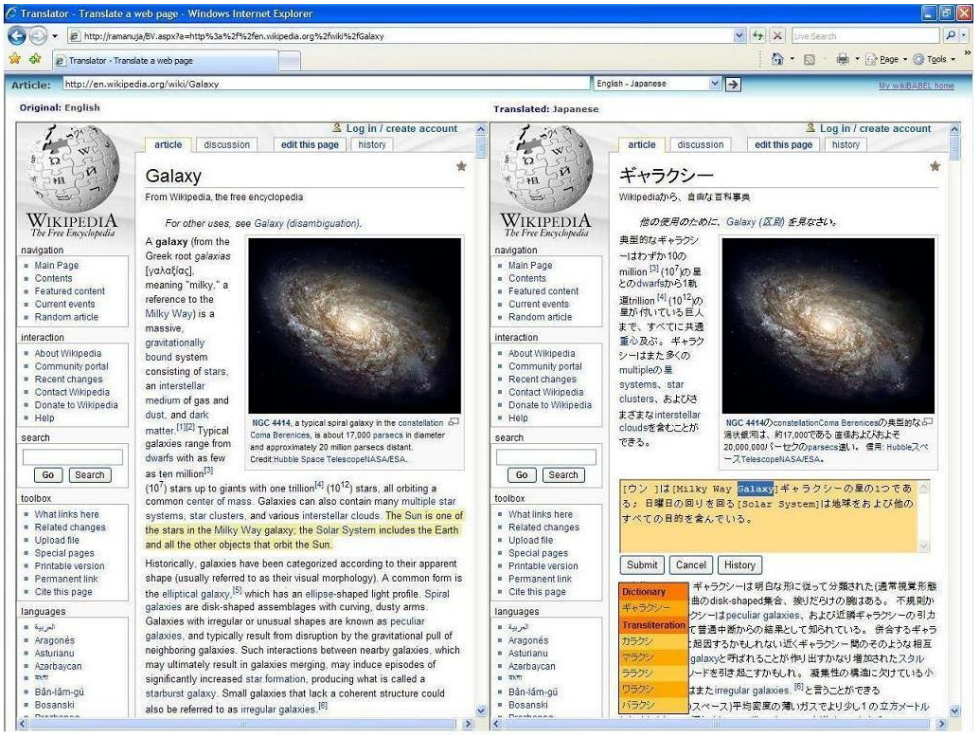

Figure 2: WikiBABEL Version 1

English Wikipedia article, and is editable, while the left panel is primarily intended for providing source material for reference and context, for the translation correction. On mouse-over the parallel sentences are highlighted, linking visually the related text on both panels. On a mouse-click, an edit-box is opened *in-place* in the right panel, and the current content may be edited. As mentioned earlier, integrated linguistic tools and resources may be invoked during edit process, to help the user. Once the article reaches sufficient quality as judged by the users, the content may be transferred to target language Wikipedia, effectively creating a new topic in the target language Wikipedia.

User Feedback: We field tested our first version with a set of Wikipedia users, and a host of amateur and professional translators. The primary feedback we got was that such efforts to create content in multilingual Wikipedia was well appreciated. The testing provided much quantitative (in terms of translation time, effort, etc.) and qualitative (user experience) measures and feedback. The details are available in (Kumaran et al., 2008), and here we provide highlights only:

- Integrated linguistic resources (e.g., bilingual dictionaries, transliteration systems, etc.) were appreciated by all users.
- Amateur users used the automatic translations (in direct correlation with its quality), and improved their throughput up to 40%.
- In contrast, those who were very fluent in both the languages were distracted by the quality of translations, and were slowed by 30%. In most cases, they preferred to redo the entire translations, rather than considering and correcting the rough translation.
- One qualitative feedback from the Wikipedia community is that the sentence-by-sentence translation enforced by the portal is not in tune with their philosophy of user-decided content for the target topic.

We used the feedback from the version 1, to redesign WikiBABEL in version 2.

3.3 Version 2: As a Transparent Edit Layer

In our second version, we implemented the significant feedback from Wikipedians, pertaining to source content selection and the user contribution. In this version, we delivered the WikiBABEL experience as an add-on to Wikipedia, as a semi-transparent overlay that augments the basic Wikipedia edit capabilities without taking the contributor away from the site. Capable of being launched with one click (via a bookmarklet, or a browser plug-in, or as a potential server side integration with Wikipedia), the new version offered a more seamless workflow and integrated linguistic and collaborative components. This add-on may be invoked on Wikipedia itself, providing all WikiBABEL functionalities. In a typical WikiBABEL usage scenario, a Wikipedia

Figure 3: WikiBABEL Version 2

content creator may be at an English Wikipedia article for which no corresponding article exists in the target language, or at target language Wikipedia article which has much less content compared to the corresponding English article.

The WikiBABEL user interface in this version is as shown in Figure 3. The source English Wikipedia article is shown in the left panel tabs, and may be toggled between English and the target language; also it may be viewed in HTML or in Wiki-markup. The right panel shows the target language Wikipedia article (if it exists), or a newly created stub (otherwise); either case, the right panel presents a *native target language Wikipedia edit page*, for the chosen topic. The left panel content is used as a reference for content creation in target language Wikipedia in the right panel. The user may compose the target language Wikipedia article, either by dragging-and-dropping translated content from the left to the right panel (into the target language Wikipedia editor), or add new content as a typical Wikipedia user would. To enable the user to stay within WikiBABEL for their content research, we have provided the capability to search through other Wikipedia articles in the left panel. All linguistic and collaborative features are available to the users in the right panel, as in the previous version. The default target language Wikipedia preview is at any time. While the user testing of this implementation is still in the preliminary stages, we wish to make the following observations on the methodology:

- There is a marked shift of focus from "*translation from English Wikipedia article*" to "*content creation in target Wikipedia*".
- The user is never taken away from Wikipedia site, requiring optionally only Wikipedia credentials. The content is created directly in the target Wikipedia.

The WikiBABEL Version 2 prototype will be made available externally in the future.

References

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

Munteanu, D. and Marcu, D. Improving the MT performance by exploiting non-parallel corpora. *Computational Linguistics*. 2005.

Giles, J. Internet encyclopaedias go head to head. Nature. 2005. *doi:10.1038/438900a*.

Quirk, C., Udupa, R. U. and Menezes, A. Generative models of noisy translations with app. to parallel fragment extraction. *MT Summit XI*, 2007.

Rainie, L. and Tancer, B. Pew Internet and American Life. *http://www.pewinternet.org/*.

Swartz, A. Raw thought: Who writes Wikipedia? 2006. *http://www.aaronsw.com/*.

Wikipedia Statistics, 2009.*http://stats.wikimedia.org/*.