wikiBABEL: Community Creation of Multilingual Data

A Kumaran
Microsoft Research India
Bangalore, INDIA 560080
+91.80.6658.6000
a.kumaran@microsoft.com

K Saravanan
Microsoft Research India
Bangalore, INDIA 560080
+91.80.6658.6000
v-sarak@microsoft.com

Sandor Maurice
Microsoft Research
Redmond, WA 98052, USA
+1.425.722.0836
samaur@microsoft.com

ABSTRACT

In this paper, we present a collaborative framework – wikiBABEL – for the efficient and effective creation of multilingual content by a community of users. The wikiBABEL framework leverages the availability of fairly stable content in a source language (typically, English) and a reasonable and not necessarily perfect machine translation system between the source language and a given target language, to create the rough initial content in the target language that is published in a collaborative platform. The platform provides an intuitive user interface and a set of linguistic tools for collaborative correction of the rough content by a community of users, aiding creation of clean content in the target language. We describe the architectural components implementing the wikiBABEL framework, namely, the systems for source and target language content management, mechanisms for coordination and collaboration and intuitive user interface for multilingual editing and review. Importantly, we discuss the integrated linguistic resources and tools, such as, bilingual dictionaries, machine translation and transliteration systems, etc., to help the users during the content correction and creation process. In addition, we analyze and present the prime factors – user-interface features or linguistic tools and resources – that significantly influence the user experiences in multilingual content creation.

In addition to the creation of multilingual content, another significant motivation for the wikiBABEL framework is the creation of parallel corpora as a by-product. Parallel linguistic corpora are very valuable resources for both Statistical Machine Translation (SMT) and Crosslingual Information Retrieval (CLIR) research, and may be mined effectively from multilingual data with significant content overlap, as may be created in the wikiBABEL framework. Creation of parallel corpora by professional translators is very expensive, and hence the SMT and CLIR research have been largely confined to a handful of languages. Our attempt to engage the large and diverse Internet user population may aid creation of such linguistic resources economically, and may make computational linguistics research possible and practical in many languages of the world.

WikiSym '08, September 8-10, 2008, Porto, Portugal.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Art and Humanities - *system for multilingual content and/or linguistic corpora creation*. H.5.2 [**User Interfaces**]: N*atural languages, machine translation, corpora creation, usability studies*.

General Terms

Design, Languages, Experimentation, Human Factors.

Keywords

Multilingual content creation, Multilingual wiki, Linguistic data creation, Human aided machine translation, User-centered design.

1. INTRODUCTION

In our research, we explore methodologies for community creation of linguistic data, in particular, parallel corpora to support Statistical Machine Translation (SMT) and Crosslingual Information Retrieval (CLIR) research. In this paper, we present a community collaborative framework – wikiBABEL – which enables the creation of multilingual content, from which valuable parallel data may be mined effectively. The wikiBABEL framework enables the creation of multilingual data, leveraging two common facts (with respect to the content and the user demographics): First, in many typical situations the content is created in one language and then is translated into other languages. For example, the user manuals of Microsoft software systems (other examples include, Caterpillar earth-moving equipment, Pfizer drug information, Harry Potter novels, etc.), are created in English and subsequently translated to other languages. The proceedings of the many parliaments around the world are created in one language (the language used by the speaker in the parliament) and get translated to other official languages as mandated by the legislative procedures. Secondly, while the original content may need to be created by subject matter experts or professionals, in many cases the translations could be provided by those who are primarily fluent in the pair of languages involved. While some translations may need to be done by professionals due to legal or business reasons (for example, the EuroParl parliament proceedings, the Pfizer drug information, respectively), others need not be: For example, a Swahili Wikipedia article on Savanna may be created by someone who is fluent in English and Swahili, by referring to available material on Savannah (including, perhaps, the English Wikipedia article on Savannah). We leverage these two facts in the wikiBABEL framework, for enabling even non-expert users to create content in a target language, by leveraging available content in one language and supporting a collaborative platform for translation.

In the wikiBABEL framework a community of users may access a pre-defined collection of existing articles in a source language, translate them at any granularity, and progressively transforming them to a collection of documents in a target language. The framework integrates a suite of tools, linguistic resources and mechanisms for community-wide collaboration to make the translation more efficient. Though the framework focuses the translation efforts to a sentence at a time, thus encouraging creation of parallel sentences, it allows users to rewrite or restructure the content in a typical wiki fashion. As we show later, in some domains sentences are naturally translated one at a time, and hence the multilingual data that get created may indeed be parallel. However, the framework allows free creation of data, and thus may yield only comparable content (i.e., sentences or text passages with unequal, but related semantic content). However, even such comparable data have been shown to be useful for research, as parallel data may be mined from them effectively and used for improving SMT quality significantly [18] [22]. It is highly likely that the content created in a platform implementing the wikiBABEL framework will fall in this category. The combination of size, variety and the social aspect of the Internet user-community suggests a vast potential for the effective creation of multilingual data, and parallel linguistic data as a byproduct.

We present here our experience in realizing the wikiBABEL framework in two distinct usage scenarios, for creating bilingual content aided by a host of tools (e.g., machine translation and transliteration systems) and resources (e.g., bilingual dictionaries, collaborative translation memories). The first is a deployment for a community-participatory translation of software support material in Microsoft, where the original documents were produced in English and translated to Portuguese language by a community of developers. The second is an under-development prototype system to aid multilingual content creation in the Wikipedia [30] leveraging the existing English content. Our experience in such developments and deployments may be invaluable in fine-tuning the methodologies for linguistic data creation by non-expert Internet users effectively and economically, thereby may make computational linguistics research possible in many languages of the world.

1.1 Motivation for Linguistic Parallel Data

Modern state-of-the-art Natural Language Processing (NLP) systems are based primarily on statistical and machine learning principles, and are shown to be highly effective and robust on a wide variety of NLP Tasks [14]. Equally importantly, they scale well with languages; that is, such approaches are implemented as language-independent generic systems that may be adapted for a specific language, by training on appropriate data in that language. Such research methodology makes NLP research possible in many of the world's resource poor languages. However, such statistical systems suffer from the critical need for large amount of appropriate linguistic corpora for training. For example, a language-independent Hidden Markov Model-based part-of-speech (POS) tagging framework requires as training corpus, a few thousand POS tagged sentences in a given language to be effective. Similarly, a Statistical Machine Translation (SMT) framework developed using readily available language independent statistical tools, requires as training corpus, a few million parallel sentences in a pair of languages. Given that the current NLP tools and technologies are statistical in nature, there is a critical need for large linguistic corpora to advance the state of art in a given languages. When available, such data had provided tremendous impetus to the NLP research [10].

Creation of linguistic data is expensive - time-wise and resource-wise - and most available linguistic corpora are primarily due to the large consorted efforts of consortia involving Governments, Industry and Academia, over many years [2] [12]. Such efforts are undertaken only in a handful of languages – primarily Western European and East Asian, and unlikely to happen in many languages of the world. Creation of some specialized linguistic corpora (such as, parts-of-speech tagging, syntactic tree parsing, semantic annotations, etc.) requires knowledge of Linguistics principles and language-specific features, and hence may be created only by trained linguists. However, several other types of linguistic data may not require trained linguists, but may be contributed even by native speakers who are fluent in a given language; for example, summarization requires only ability to read and comprehend a passage in a language and the ability to produce a grammatically correct summary of the passage. In our research, we consider creation of a specific type of linguistic resource – the parallel corpora[1] – that is critically required for SMT research, and development of practical systems in a given pair of languages. The SMT systems typically acquire translation knowledge by statistical translation models built based on the parallel corpora provided as the training corpus. In general, the quality of the translations by an SMT system improves with size of the training corpus. Typically the SMT systems require a few million sentence pairs for developing a reasonably accurate and fluent translation system between a pair of languages; the required training corpora size may be an order of magnitude more for languages with richer linguistic features, such as, fine-grained morphology, agglutination, etc.

Hand-creation of such large parallel data is very expensive; hence, traditionally the SMT researchers relied on parallel data that had been created for some other purpose, such as, bilingual publication of parliament proceedings (e.g., EuroParl [8][11], Canadian Hansards [4], Rajya Sabha [24]), news translations (e.g., Xin Hua News Agency [35]), etc. Hence, SMT research was limited only to those languages where such data were available. Even when available, such data may still not be sufficient for general purpose SMT systems, as the word coverage and usage styles are restricted to narrow domains (e.g., political debates, narrative news, respectively). Hence, it is highly critical that we find new ways of creating parallel corpora, to make SMT research possible in many languages of the world.

1.2 Related Work

Collaborative creation of information by a community is a well proven methodology, as shown by the enormous success of Wikipedia [30]. Our work attempts the creation of linguistic parallel data by the community. While Wikipedia contains articles that discuss the same topic in different languages (such as, those between the English article on *London*, wiki-linked to the

[1] Parallel corpus is a set of parallel sentences in a pair of languages; a parallel sentence in a pair of languages refers to two sentences – one per language – that have the same semantic content, expressed in the respective languages.

Spanish article on *Londres*), usually such pairs are not parallel, due to the differences in size and the semantic content. However, such data still may be useful for other resources, such as, named entities, to aid SMT research. The other Wikipedia initiatives, such as, the Wiktionary [34] and Omegawiki [20] are excellent resources for SMT research, as they provide word-level lexical, semantic and ontological information, and their association between languages; however, SMT research requires, in addition, phrasal and sentential information for training. The Wikipedia translation [33] addresses translation efforts within Wikipedia, and provides a forum for the coordination of translation in the Wikipedia community, but has not yet addressed the creation of linguistic data.

Creation of linguistic corpora had been done successfully in the past, but such efforts had been largely confined to small specialized user-population. For example, the pyramid evaluation [19] involves semantic annotation of text, by linguists. LDC [11] has proven to be very successful in data creation, fueling the NLP research worldwide, but, to the best of our knowledge, still does not involve Internet population for content creation, which we attempt to do. On the other hand, gaming metaphors are effectively employed in collecting rich data involving the internet; for example, text annotation of images by the Internet users [1] [7], had proved itself to be very successful, and had helped to annotate nearly 10 million images in the web. Though similar methodologies have been employed for creating other linguistic content [27], none had been attempted for collecting parallel data. To the best of our knowledge, our work may be the first of its kind, in creating parallel data leveraging contributions by Internet user population.

Translation involving the Internet community has been proposed in [5], where the massive online collaboration for translation is proposed and analyzed; Also, the topic of massive online collaborations for translations is the main theme explored in the Wiki-Translation [29] community, including the resources, tools, technologies and processes needed to make such collaborations efficient and effective. Our wikiBABEL architecture and implementations may be possible realizations of such themes.

A systematic analysis of the processes and tools for collaborative creation and maintenance of related multilingual content by a set of content authors, was presented in [6], along with a partial realization of such a theme, in [13]. Our efforts overlaps some (not all) the issues addressed in [6], as we assume the original content is stable in one language, before it gets translated to the second language. We argue that such scenarios exists in specific domains as detailed in Section 3. Traduwiki [26] takes a parallel approach for involving the internet community for the translation of public documents. We take an additional step of attempting to create parallel linguistic data for SMT research. While the search engine portal Google [9] has integrated an edit mechanism for user-correction of their automatic translations, there are no known studies on their effectiveness; our study may provide an indication of effectiveness of such approaches.

In this paper, we restrict our attention to studying the enablers for collaborative creation of parallel data – specifically, compelling user-scenarios, usability of the tools and interface provided. As explained later, we focus the users on sentence-by-sentence translation, which may be the norm in some domains. In general domains, this approach is more likely to yield noisy parallel data, or even, comparable data. However, even such data may be very valuable, as significant quality improvements in SMT systems have been demonstrated using parallel data mined from noisy parallel data [18]. Even comparable data were successfully employed in SMT quality improvements, by the parallel data mined from comparable data [22].

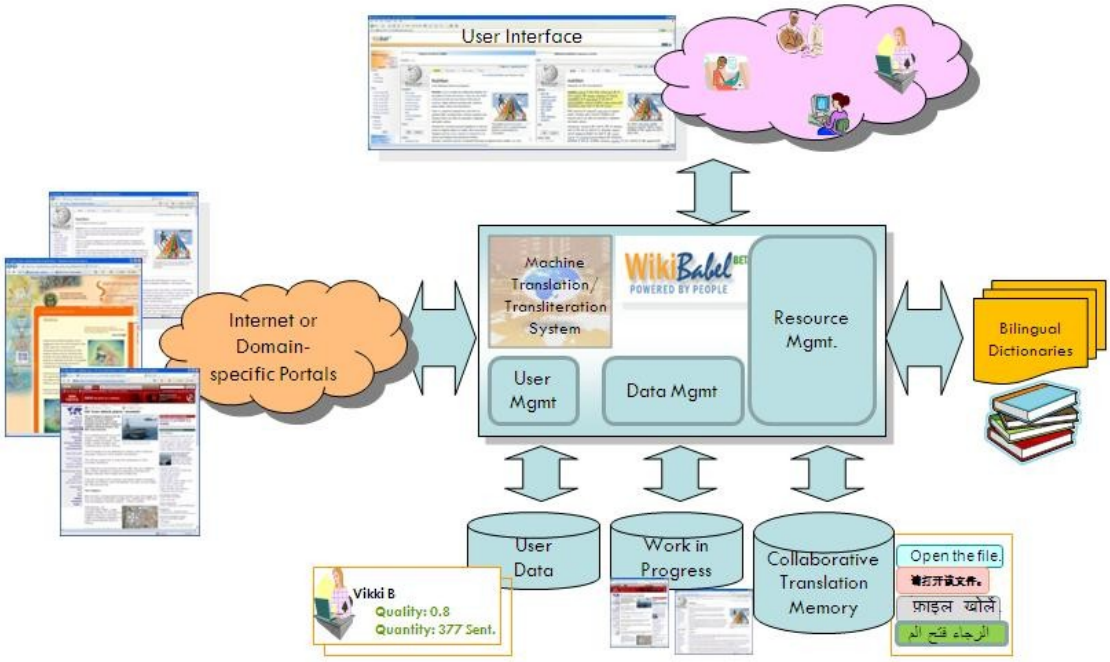

Figure 1: wikiBABEL Architecture

2. wikiBABEL ARCHITECTURE

In this section, we outline the architecture of the wikiBABEL system and detail its components. In a typical scenario, a set of tasks is created in the system, which are being worked on collaboratively by a community of users. In this paper the task involves translation of a set of documents from a source language to a target language, but the collaborative scenario may be applicable in many different tasks in other domains. The architecture is illustrated in Figure 1, and the components of the architecture are detailed below.

Central to the architecture is the **wikiBABEL** workflow engine that manages communication between all its components that manage users, content, tools and resources. The users are enrolled in wikiBABEL system, along with some relevant demographic information (e.g., fluent languages, interested domains). Users are managed by the **user-management** system that tracks not only the user credentials and demographic information, but also the quality and quantity of their contributions, for any possible rewards. The users may participate in any number of collaborative translation tasks, and a task may be worked on by multiple users. Some tasks may have task-specific demographic requirements (e.g., fluency in a languages, exposure to a subject domain, etc.). The users may access available tasks (e.g., list of documents that are being worked on) and may choose any of the tasks to contribute to.

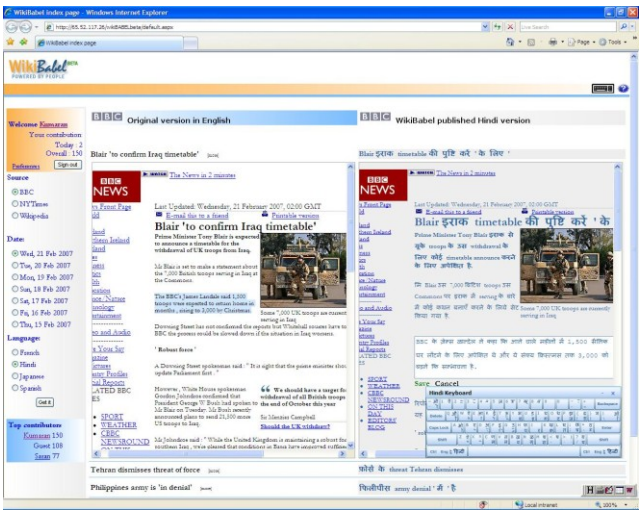

Figure 2. A Typical wikiBABEL User Interface

The **work-in-progress** data store stores all tasks (e.g., documents) at a granularity to make even small contributions possible and track-able. For example, a document that needs to be translated is stored at sentential (for normal text) or sub-sentential (for titles) levels so that a user may choose to translate as little as a single phrase or a sentence. Every user translation (sentential or sub-sentential level) is captured and stored in the data store, along with its association to the original fragment in the source language. This data store is a crucial component of wikiBABEL, as it tracks not only the task progression, but also provides a roll-back mechanism to handle any regressions in quality. More importantly, the parallel data at word, phrase or sentence levels is mined out of this store, as it tracks the association between the source language content and the target language content.

The **user interface** is, in general, task-specific, but we present here a typical 2-pane UI, as shown in Figure 2. For document translation, the original content in the source language is shown in the left panel, and the editable content in the target language, in the right panel, with the same look-and-feel as the original. Initially, a rough translation in the target language is produced using a machine translation system and presented in the right panel. Since the work-in-progress data store stores the entire document at a sentential granularity, the associations between the source content and the target content are readily available. On mouse over, corresponding text segments in the left panel (in the source language) and right panel (in the target language) are highlighted simultaneously, providing visual clue on the current state of the translation for that sentence. Users may choose to work on any text segment to translate or correct, and on mouse click, the right panel text segment (the current version), is copied into an edit box, and presented to the user. If selected text segment had any translation history (earlier edits by users), then all the versions are presented in the reverse chronological order, so an appropriate version may be selected for further edit by the user. The edit-box is produced in-line, in order to make visible the context of the sentence being edited, both in the source article and in the target article. The user may correct the text content fully or partially in the edit box, using several tools (detailed later) available. On save, the user edits are stored in the work-in-progress data store as the latest version of the text segment.

The UI features are designed to be light, requiring minimal server-side interaction, to enable the system scale for large number of users. The edit box is capable of handling and editing multilingual text. For several languages we had integrated appropriate soft keyboards for input; along with the edit box, a keyboard layout in a given language appears on the screen. A contributor may use the mouse clicks on appropriate keys in the soft keyboard for multilingual input, or may over-ride the soft keyboard with an installed keyboard drivers in the client. A soft keyboard has an advantage of requiring no client side installation, and may be configured easily for several languages, and different keyboard layouts for a given language (typical, in Indic and East Asian languages). In addition, user-configurable layouts (side-by-side or top-and-bottom) for viewing the source and target content are also available. Finally, access to **Internet** is a necessary component for implementation of the usage-scenarios on this architecture.

wikiBABEL architecture also supports a few tools and resources that may be helpful in the translation process. First, a set of **bilingual dictionaries** were integrated and used for requesting word-level translations. Bilingual dictionaries are very useful for their coverage (all words of a language) and completeness (all possible translations for a word in a target language). In addition, domain-specific, or even, user-specific bilingual dictionaries may be loaded by the users, explicitly. Such customization may provide an efficient scratch-pad for the community of users. Second, an integrated **Machine Translation** system may be invoked for requesting a rough translation of a source language string, at phrasal or at sentential levels. In a typical scenario, an SMT system is used to generate rough translation in the target

language for every text segment of the original article, and the rough translations are used to populate the content of the right panel. Every subsequent user is shown the latest version of the target sentence. In addition, we had integrated a **Machine Transliteration** system for translating names, as name translations are not usually available in bilingual dictionaries, and not possible by machine translation systems for names that have not been encountered in the parallel data used during training.

A **collaborative translation memory** component stores all the user translations, which are made available to the community as a translation help ('*tribe knowledge*'). This component stores only user corrected translations, at every granularity, as the MT can provide automatic translations at phrasal or sentential level on request. When the user requests translation help for a source passage, the collaborative translation memory is searched, and any available user translations are presented to the user first. If this search fails, then the MT system is invoked for an automatic translation. Multiple translations (if available) are presented for users to choose from. Finally, the architecture is designed open, to integrate any user-developed tools and resources easily.

3. DEPLOYMENT SCENARIOS

In this section, we present implementations of wikiBABEL for two different usage scenarios: The first scenario is the collaborative creation of Microsoft user support manuals, and the second, the collaborative creation of multilingual Wikipedia articles; in both the cases the original content is created in English, and translated to another language based on the English content. In both the cases, the wikiBABEL architecture as presented in Section 2 is adhered to, but with scenario-specific user interface and workflow features. A homegrown wiki software used for development of these systems.

3.1 MSDNwiki: A Community for Technical Knowledge Sharing

Microsoft Software Developers Network (MSDN) [16] is an entity that manages Microsoft's relationship with the development community using Microsoft products; it disseminates technical information pertaining to software products, user manuals, support information and even code samples. While MSDN data is made available in several different languages, the data are created in English and hand-translated to other languages, incurring significant cost and time. A recent initiative – MSDNwiki – enables a user community to create the MSDN content collaboratively in a specific set of languages, based on the existing English content. In a pilot MSDNwiki deployment based on the wikiBABEL architecture components, the English content is made available in a common platform to be translated to Portuguese, by the Brazilian MSDN community.

3.1.1 Domain Characteristics

User Profiles In the MSDN community members are primarily product evangelists, software developers, technical writers or users of Microsoft products. The MSDN allows participants to download software and access support material, ranging from software administrative and user manuals, support and help pages to trouble shooting guides. In the user community there are privileged users (referred to as, Most Valuable Partner or MVP), who are experts in Microsoft products and technologies and are active in the user community.

Data Characteristics The content of MSDN is generated primarily by subject matter experts within the Microsoft in English, and the information gets disseminated through MSDN network. Typically, there are a few hundreds of thousands of articles in MSDN, covering wide variety of Microsoft products, tools and technologies, and as may be expected, the content is primarily technical in nature. Though the content is made available in many different languages of the world, there is a time-lag between the content availability in English to its corresponding availability in a foreign language, due to logistical reasons in translation. Also, due to business and logistical reasons, the content does not get translated to all the languages in which Microsoft products are available.

Machine Translation system It is important to note here that the MT system used within Microsoft is primarily trained on parallel corpora culled from technical documents and user manuals that have been translated and published by Microsoft over many years; hence, the quality of the MT system is significantly high in the technical domain, close to that of human translations.

3.1.2 Use-case Scenario

The MSDNwiki initiative supports collaborative translation of the MSDN technical data by the MSDN developer community, and is deployed for translation of MSDN documents from English and Portuguese with Brazilian MSDN community. In a workflow model similar to the one outlined in Section 2, and the deployment is based on components of wikiBABEL framework.

Any of the logged-in MSDN users may correct the rough initial content in Portuguese (the English MSDN content translated using the Microsoft MT system). Figure 3 shows a typical screen from the English-Portuguese MSDNwiki system. All the articles available for editing are shown on a side panel (3A), and the user may choose any of these articles to view and edit. A home grown state-of-the-art statistical machine translation system was used to pre-translate and publish in Portuguese, all MSDN English content. The selected article and its translation are shown in a 2-pane window; the English MSDN content is shown in the left panel as a reference, and its rough machine translated version in Portuguese is shown in the right panel. The correspondence between the text segments in English and Portuguese is visually indicated by parallel highlighting (3B) of corresponding segments, on mouse over in either of the panels. Users may choose to edit any right side text content, by clicking on the "*Editar*" button that is shown above any editable content; on click, a small edit box is popped up with the current Portuguese text segment (3C). Currently, the edit mechanism allows edits only on a sentence by sentence basis. While all users are allowed to edit the translations and suggest alternative translations, on save, the edited content is stored in a "pending approval" mode, to be reviewed by a set of authorized approvers. Typically, the approvers are MVPs of MSDN community in Brazil. Once approved, the suggested translations are posted, and become part of the page as the latest version available for all subsequent users. While the initial deployment allows only translation of existing English content into Portuguese, subsequent versions are planned to aid both editing and creation of new content in Portuguese.

3.1.3 *Qualitative Assessment of 2-Pane Interface*

To measure the usability of the interface, a 2-stage process was undertaken: In the first phase, the standard 2-pane interface without the edit capability was introduced to a set of users, in three different regions: Mexico, China and France. About 10-20 users in each demographics were introduced to the 2-pane interface, in which any arbitrary URL may be translated and shown; the original URL in the left pane and the translated URL (into a language of user's choice) in the right pane. The users were asked to familiarize themselves with the interface and the features, and were asked to perform a set of discovery and usability tasks (e.g., opening a URL, translating it to a target language, identifying a sentence translations, correcting the translation using a simple edit interface, etc.). At the end, they were interviewed to assess their experience *qualitatively*.

Users from all demographics were very enthusiastic about the 2-pane interface that shows the original and the translations side-by-side, with visual clues on their correspondence. The user comments included, "*this is a fresh new look*" and "*this is perfect*". All the participants reported that such a 2-pane interface improved their understanding of content, even when the translation quality is not perfect (as indicated by their comments on the low quality of translations). Some users agreed to provide corrections to the machine translation output in order to improve the quality of the SMT system, while almost all users wanted to access what the other users might have provided for translations, primarily as an opportunity to learn a new language. A large fraction of users expressed concerns about that the time and effort to correct the data, and were skeptical about the impact of their data in improving the quality of the SMT system.

3.1.4 *Qualitative Assessment of MSDNwiki*

In addition, a prototype version of MSDNwiki, based on the 2-pane viewer, was demo-ed to a limited set of users from the Brazilian MSDN community. The community expressed strong enthusiasm for such a framework for providing or correcting translations, possibly due to the following three factors:

1. Availability of a robust machine translation system between English and Brazilian Portuguese within Microsoft, trained on a large parallel corpora culled from Microsoft technical documents published in English and Portuguese over the last several years. Hence the quality of automatic translation is very high for technical documents, and hence the translations required minimal changes by the users. In most cases, a stable version of the translation required minimal edits within a single edit session, leading to the user-perceived permanency of their contributions.

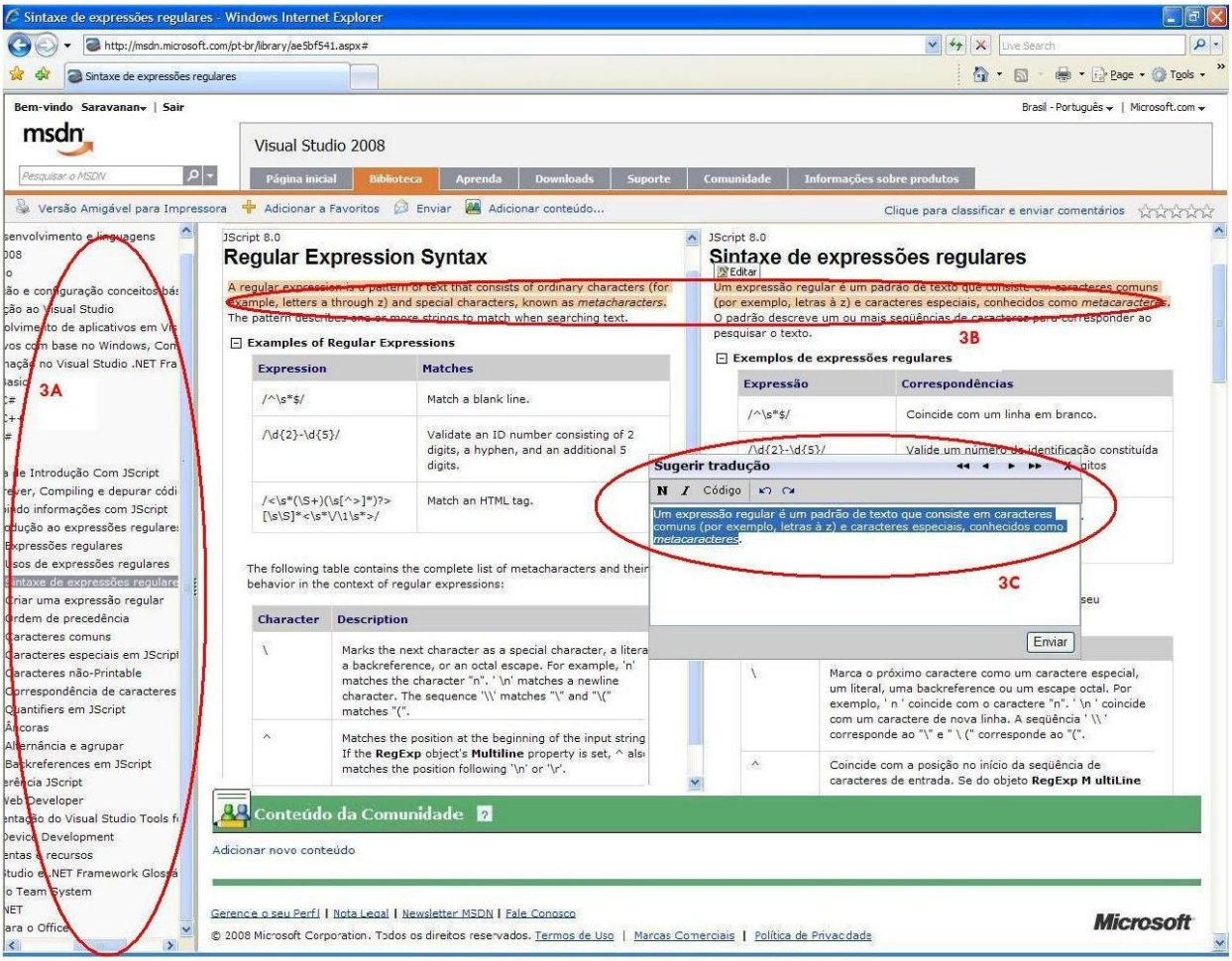

Figure 3: wikiBABEL on MSDNwiki Data

2. The content being translated is from technical domain; hence, there was a strong user preference to keep the presentation structure (including, intra-sentential structures) in order to avoid any change of the meaning in translation. Hence, the users did not consider the sentence-by-sentence mode of translation as too constraining. In addition, the technical content, by its very nature, does not lend itself to much creative rewrites.
3. The user community, primarily consisting of technical experts, were enthusiastic about contribution to gain recognition and respect from their peers. MSDN is a widely widely accessed and used forum for this community, resulting in strong desire to contribute.

In addition, nearly for an year before this deployment, the Microsoft support website content was available primarily as a machine translated content, with favorable reviews from the community about the quality. Hence, the community was not averse to the idea of using MSDNwiki was deployed first in Portuguese.

3.1.5 MSDNwiki Deployment

The above MSDNwiki system is deployed in April 2008 [17] to MSDN users in Brazil to create the MSDN content in Portuguese. The content being edited is monitored for quality and quantity, and captured in a data store. The sentences that are edited would be close to being truly parallel sentences (specifically, those that have been edited by MSDN users and approved subsequently by approvers) may be truly parallel, as MSDNwiki is a moderated community. Also, the workflow allows only sentence-by-sentence translations. We believe that this effort will yield parallel data in significant quantities, resulting in substantial improvement in the translation quality of our MT system. We intend to analyze the data that is being captured and publish our findings shortly.

3.2 Wikipedia: The Open Encyclopedia community

The Wikipedia is one of the most successful community contributed initiatives. A second wikiBABEL deployment is being built as a free collaborative portal that enables creation of multilingual Wikipedia content by the Internet users based on existing English Wikipedia articles.

3.2.1 Domain Characteristics

User Profiles Wikipedia is emerging as a de-facto source of information for large sections of Internet users: A recent study reported that the Wikipedia is consulted by a third of Americans, and the usage goes up with the academic levels of the users [31]. Overall, Wikipedia gets about 1 million visits a day resulting in 13 million page requests, clearly demonstrating the value of Wikipedia [32]. On the contribution side, Wikipedia has about 200 thousand contributors (> 10 total contributions), of which about 25% are active (> 5 contributions per month) and 4% 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 [25] that claims that large fraction of the content were created by those who made very few and/or very occasional contributions; further it is found that and majority of such contributions were editorial in nature, than new content creation[2]. Finally, the Wikipedia contributors are already tuned to the concept of content creation for the benefit of the community, and hence may be enthusiastic about data creation for the computational linguistics community as well.

Data Characteristics The quality of Wikipedia data has been reported in studies to be in par with that of Encyclopedia Britannica, and by far the most popular site for academic references [23], and is being cited increasingly in the academic publications [31]. However, content-wise, there is a large disparity among the different language Wikipedia's: Currently English Wikipedia, with about 2.3 Million articles is the largest among all languages [32]. Other major Wikipedia's (in major Western European and East Asian languages) have about 250-750 thousand articles each, but there is a long tail of more than 200 languages with less than 1% of English Wikipedia content [32]. Such a skew, despite the size of the user population or utility of the collection, indicates an enormous room for improvement in many non-English Wikipedias.

3.2.2 Use-case Scenario

In this scenario, we start with the premise that the large disparity in content between English and a target language Wikipedia may be leveraged for new content creation in other language Wikipedias. The characteristics of the users and their contributions reported in [25][24] suggest that the large user base of Wikipedia may be leveraged successfully for the creation of non-English Wikipedia content, even if we assume only small amount of corrections per visit, or even per contributor, primarily focused on correcting the rough content.

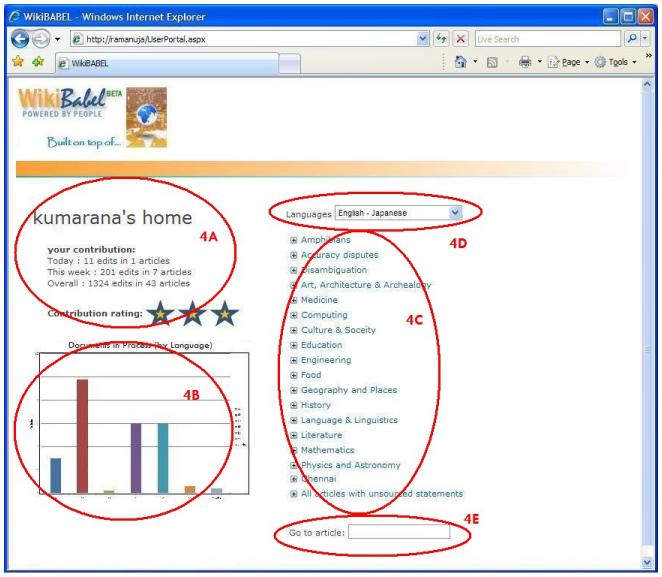

Figure 4. wikiBABEL for Wikipedia

The typical user-scenario in the wikiBABEL on Wikipedia deployment is that a community of collaborative users is

[2] The above study in English Wikipedia, is likely to be true for other language Wikipedias as well.

translating a set of Wikipedia articles from, say, English, a target language. The wikiBABEL portal is seeded with a set of Wikipedia articles to focus the translation activity, or the articles may be added on the fly, by the users. In either case, at any given time, there is a set of documents being worked on by the community. Figure 4 shows a typical user home screen, where the user credentials and their activity (quantity and quality) are shown (4A & 4B). In addition, the set of Wikipedia articles being edited by the community is also presented, organized in the Wikipedia-specified classification hierarchy (4C). A user may chose any of these existing articles for editing, or may specify a new Wikipedia topic to edit in a text box (4E); if a new topic is specified, then the corresponding articles form Wikipedia is presented to the user, and the user chosen article will be added to the above list automatically. The user may customize the presentation, with respect to the domain (e.g., History, Arts & Culture, etc.) or languages (e.g., Arabic, Japanese) of interest. The current deployment includes translations in a limited set of languages (4C), primarily from the Western European, East Asian and Indic languages. The deployment language choices are constrained mainly by to the availability of an MT systems to provide rough initial target language translations of the original content in English.

Once an article is chosen for editing, a 2-pane window is presented to the user, as shown in Figure 5, where the original English Wikipedia article (in the left panel) and a rough translation of the article in the user-chosen target language (in the right panel) are presented to the user. The translated article is presented with the same look and feel as the original Wikipedia article. The content in the right panel (in the target language) is editable, while the left panel is primarily intended for providing source material for reference and the context, for the translation correction. On mouse-over (on any sentence, in either left or the right panel), the parallel sentences are highlighted (5A), linking visually the text on both panels. On a mouse-click on a highlighted text segment, an edit-box is opened *in-place* in the right panel (5B), and the current translated content of the text segment is placed in it for editing. If the content is being edited for the first time, then the edit box contains the rough MT output. On subsequent edits, the latest user-provided version is placed in the edit box, and in such cases all previous versions of the translated sentence are made available through the "History" option provided below the edit box; the user may view and choose any of the earlier versions to edit.

In the edit box, in addition to the standard features (e.g., multilingual input mechanism, editors, etc.), several linguistic

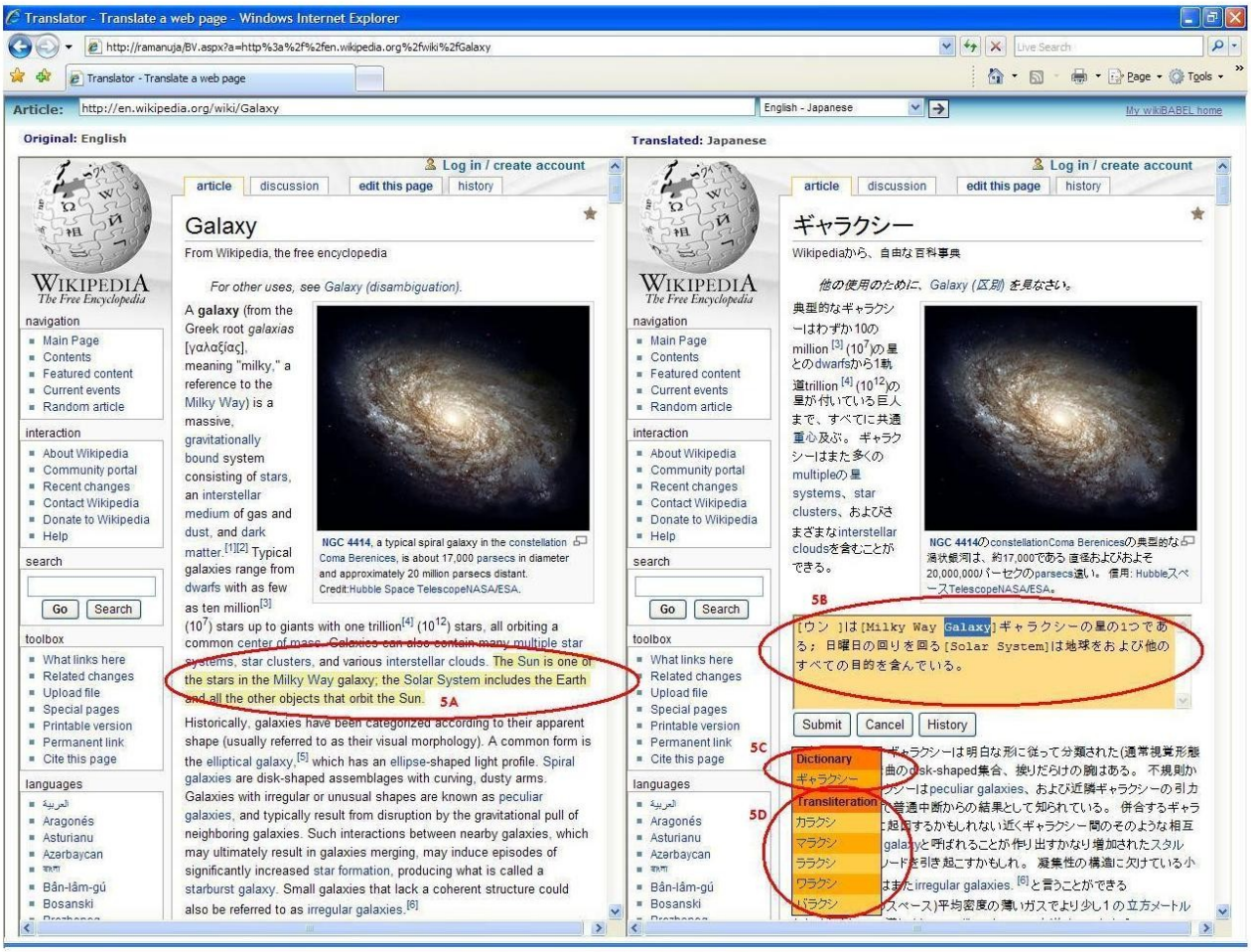

Figure 5: wikiBABEL on Wikipedia

help are provided. First, the user may highlight any sub-sentential fragment in English and choices from the collaborative translation memory and the Machine Translation system are provided. If a single word is highlighted, then a bilingual dictionary lookup (5C) and a set of suggestions from a machine transliteration system (5D) are provided. The number of choices presented to the users and their preferred order are customizable by the user. The user may pick one of the choices provided, replacing the highlighted text segment in the edit box. On "save", the edited content becomes the current version of the text segment being edited, and pushed into Wikipedia, and relying on users to keep the content correct.

3.2.3 Usability Testing of the System

We have usability tested the wikiBABEL interface translating content between English and Hindi, in about 20 sessions, each session spanning 4-5 hours, totaling about 100 hours of usability testing. The profiles of the usability testers included both professionals (translators or content creators in the respective languages) and amateurs (students, casual users of Wikipedia, etc). Each session involved a subject translating two articles; the articles consisted of approximately 20 sentences each, with similar word count, word and phrasal difficulty and sentence-complexity, and were hand-picked from the same domain. In each of the phases, they rated the difficulty of each of the sentences at word and sentential levels, quantitatively, on a scale of 1 (easy) to 5 (hard), before the translation. Table 1 provides the difficulty of the passages as rated by the testers at word and sentential levels.

Table 1. Perceived Difficulty of Articles Translated

Tester Profile	# of Session	# of Sent.	Baseline		wikiBABEL	
			Sent	Word	Sent	Word
Amateurs	14	361	2.535	2.348	2.463	2.418
Prof.'s	12	391	2.018	1.913	2.274	2.150

The amateurs ratings were in general more than the professionals, reflecting their linguistic and subject matter exposure. Very importantly, there is hardly any difference in the perceived hardness of the passages translated (words or sentences), and hence should have resulted in equivalent translation times. The tester was asked to translate the first article in the traditional way using a multilingual tools and editor (referred to as, *baseline*), and the second article using wikiBABEL interface (referred to as, *wikiBABEL*), with all the in-built access to multilingual resources. The user performances were measured quantitatively (translation times) and qualitatively (post-testing interviews).

Here we present our findings briefly: All users provided very positive qualitative feedback on the usability of side-by-side presentation; they reported that the visual layout made the translation process intuitive and easy. In addition, the amateur translators exhibited significant improvements in translation completion and efficiency (~40% faster translation times, from 6.71 to 4.82 minutes per sentence translation). Also, they reported substantial value derived from the integrated resources, such as, the rough translations and bilingual dictionaries. However, the professional translators reported that they were distracted by the rough translations provided, corroborated by a significant drop in their efficiency (~30% slower translation times, 2.89 to 3.78 minutes). However, they completed translating every one of the sentences in the test set. In several instances, we observed that the professional translators give up after trying to correct the MT output, delete it and start all over again, resulting in wasted effort. they reported minimal use of the integrated linguistic resources, due to their fluency and expertise in the language .

3.2.4 Effect of Rough MT Quality

Not surprisingly, we observed that the quality of the machine translation system played a significant role in the perception of the usability of the framework. When a high quality MT system was used for providing the initial translation (with BLEU score [20] of ~34) all testers provided positive feedback about the system, though the professionals less enthusiastic than the amateurs. All reported having used, fully or partially, the translation suggestions provided by the MT system in their final translation (~60% for the amateurs and ~35% for professionals). Translations provided by a low quality MT system (with BLEU score of ~15 or less) were usually deleted by the professionals, and the entire translation redone; amateurs still used the words, but rearranged them to form the final translation A significant adverse effect of this workflow was that it slowed the translation efforts of professionals much, thus taking longer time than if they had started with a blank edit box. We concluded that providing a low-quality MT system is worse than providing no suggestions at all.

3.2.5 Our Deployment Plans

The Wikipedia system is currently under development, and is being tested for usability and robustness. We plan to deploy the system by the end of Summer of 2008, and publish our experience with the deployment.

4. CONCLUSION & FUTURE WORK

In this paper, we present a community collaborative framework – wikiBABEL – that enables creation of multilingual content by a community of users, from which a rich linguistic resource – the parallel corpora – may be mined. The wikiBABEL framework provides a platform for user-communities to collaboratively create multilingual content, by editing, correcting and/or verifying the rough initial content provided by an MT system. We detailed our framework, the processing components and the integrated linguistic tools and resources to implement the framework. We outlined two specific usage scenarios – MSDNwiki and Wikipedia – targeting two distinct user communities. The scenarios are quite diverse in terms of data need and user profiles: MSDNwiki is a moderated user community for creation of Microsoft technical documents, and the Wikipedia initiative is meant for the creation of multilingual Wikipedia articles by any Internet user. We present our current experiences in each of these implementations, along with the results from controlled user experiments. Over next few months, we intend to study the effectiveness of these deployments in collecting parallel data, the quality improvements to the SMT system resulting from the parallel data mined from the user-edits. Such research requires significant user edits (in the order of few million corrected sentences) and may need significantly long deployments to gather sufficient data. We plan to publish our findings, regarding the quality and quantity of the collected linguistic parallel data, and its effect on SMT quality, subsequently. We plan to continually improve the platform and

the methodology, to make this framework a robust platform for collecting parallel linguistic data.

Finally, as indicated earlier, large parallel corpora are critical for the statistical machine translation systems research, and hand creation of such data is prohibitively expensive. However, active involvement of Internet population may signal a vast potential for creating such parallel data effectively, making SMT systems research possible in many resource-poor languages of the world. Our experience in such deployments may be invaluable in fine-tuning the methodologies for data creation by the Internet population, thereby make SMT systems practical, more likely possible, in many languages of the world.

5. REFERENCES

[1] Ahn, L. V. and Dabbish, L. Labeling images with a computer game. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319-326, 2004. http://doi.acm.org/10.1145/161468.161471.

[2] British National Corpus. http://www.natcorp.ox.ac.uk/.

[3] Everything Development Company. http://www.Everything2.com.

[4] Canadian Hansard Parallel Corpus. http://www.isi.edu/natural-language/download/hansard/.

[5] Désilets, A. Translation Wikified: How will massive online collaboration impact the world of translation?. Keynote in *Translating and the Computer*. 2007. http://www.aslib.com/conferences/TranslationWikified.pdf.

[6] Désilets, A., Gonzalez, S., Paquet, S. and Stojanovic, M. Translation the Wiki Way. *Proceedings of the 2006 international symposium on Wikis*. pp. 19-32. 2006.

[7] ESPGame. http://www.espgame.org/.

[8] EuroParl Parallel Corpus V3. 2007. http://www.statmt.org/europarl/.

[9] Google. http://www.google.com/.

[10] Harmon, D. Meeting of the MINDS: Future directions for human language technology. *Report of the MINDS workshop*. Retrieved on December 12, 2007. http://www.itl.nist.gov/iad/894.02/MINDS/FINAL/exec.summary.pdf.

[11] Koehn, P. EuroParl: A parallel corpus for statistical machine translation. *Proceedings of the MT Summit IX*. pp. 79-86. 2005.

[12] Linguistic Data Consortium. http://www.ldc.upenn.edu.

[13] Lizzywiki. http://lizzy.iit.nrc.ca.

[14] Manning, C. and Schutze, H. Foundations of statistical natural language processing. 1999. MIT Press.

[15] MoulinWiki. http://www.moulinwiki.org.

[16] MSDN. http://msdn.microsoft.com.

[17] MSDNwiki. http://msdnwiki.microsoft.com.

[18] Munteanu, D. and Marcu, D. Improving the machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*. 31(4):477-504. 2005.

[19] Nenkova, A., Passonneau, R. and McKeown, K. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*. 4(2). 2007.

[20] Omegawiki. http://www.omegawiki.org.

[21] Papineni, S, Roukos, S., Ward, T. and Zhu, W. J. BLEU: A method for automatic evaluation of machine translation. *TR-RC22176, IBM Research*, 2001.

[22] Quirk, C., Udupa, R. U. and Menezes, A. 2007. Generative models of noisy translations with applications to parallel fragment extraction. *Proceedings of MT Summit XI*, pp. 337-384, 2007.

[23] Rainie, L. and Tancer, B. Pew/Internet: Pew Internet and American Life Project. 2007. (Retrieved on May 1, 2008) http://www.pewinternet.org/pdfs/PIP_Wikipedia07.pdf.

[24] Rajya Sabha of the Parliament of India. http://rajyasabha.nic.in/.

[25] Swartz, A. Raw thought: Who writes Wikipedia?. 2006. (Retrieved on May 2, 2008) http://www.aaronsw.com/weblog/whowriteswikipedia.

[26] TraduWiki. http://www.traduwiki.org.

[27] Verbosity. http://www.gwap.com/gwap/gamesPreview/Verbosity.

[28] Wales, J. Internet encyclopedias go head to head. *Nature* **438**. pp. 900-901. 2005.

[29] Wiki Translation. http://www.wiki-translation.com.

[30] Wikipedia. http://www.wikipedia.org.

[31] Wikipedia Reliability. http://en.wikipedia.org/wiki/Reliability_of_Wikipedia.

[32] Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm.

[33] Wikipedia translation. http://meta.wikimedia.org/wiki/Translation.

[34] Wiktionary. http://www.wiktionary.org.

[35] XinHua News Chinese-English Parallel Corpus. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007T09.

[36] WordNet. http://www.cogsci.princeton.edu/~wn.