Visualizing Cross-Lingual/Cross-Cultural Differences in Concerns in Multilingual Blogs

Hiroyuki Nakasaki†  Mariko Kawaba†  Sayuri Yamazaki†
Takehito Utsuro‡  Tomohiro Fukuhara‡

†University of Tsukuba, Tsukuba, 305-8573, JAPAN  ‡University of Tokyo, Kashiwa, 277-8568, JAPAN

Abstract
The goal of this paper is to cross-lingually analyze multilingual blogs collected with a topic keyword. The framework of collecting multilingual blogs with a topic keyword is designed as the blog feed retrieval procedure. Multilingual queries for retrieving blog feeds are created from Wikipedia entries. Finally, we present an interface for visualizing cross-lingual/cross-cultural differences in concerns and opinions that are closely related to a given topic. Preliminary evaluation results support the effectiveness of the proposed framework.

Introduction
Websites or blogs are considered to be one of personal journals, market or product commentaries. There are several previous works and services on blog analysis systems (e.g., (Fukuhara, Utsuro, and Nakagawa 2007)). With respect to blog analysis services on the Internet, there are several commercial and non-commercial services such as Technorati, BlogPulse, kizasi.jp, and blogWatcher. With respect to multilingual blog services, Globe of Blogs, Best Blogs in Asia Directory, and Blogwise can be listed.

The goal of this paper is to cross-lingually analyze multilingual blogs collected with a topic keyword. First, the framework of collecting multilingual blogs with a topic keyword is designed as the blog feed retrieval procedure recently studied in TREC 2007 Blog track as one of its task (Macdonald, Ounis, and Soboroff 2007). In this paper, we take an approach of collecting blog feeds rather than blog posts, mainly because we regard the former as a larger information unit in the blogosphere and prefer it as the information source for cross-lingual blog analysis. Second, multilingual queries for retrieving blog feeds are created from Wikipedia (English and Japanese versions1) entries, where interlanguage links are used for linking English and Japanese translated entries. Here, the underlying motivation of employing Wikipedia is in linking a knowledge base of well known facts and relatively neutral opinions with rather raw, user generated media like blogs, which include less well known facts and much more radical opinions. We regard Wikipedia as a large scale ontological knowledge base for conceptually indexing the blogosphere. Finally, we use such multilingual blog feed retrieval framework in higher level application of cross-lingual blog analysis. Here, multilingual blog analysis can be quite easily realized through an interface for visualizing cross-lingual/cross-cultural differences in concerns and opinions that are closely related to a given topic.

Overall Framework of Cross-lingual Blog Analysis
Overview of the proposed framework is shown in Figure 1. First, multilingual queries for retrieving blog feeds on a topic (in this case “whaling”) are created from Wikipedia entries. Next, from the collected blog feeds, terms that are characteristic only in one language or in both languages are automatically extracted. Here, we apply a statistical measure for mining cross-lingual differences between terms in two languages, as well as a monolingual measure for terms related to the given topic. Then, by counting occurrences of the topic name and the related terms extracted from the Wikipedia entry in blog posts in both languages, characteristic blog posts are ranked. Finally, through an interface for visualizing cross-lingual/cross-cultural differences in concerns and opinions that are closely related to a given topic, it becomes much easier to discover the most characteristic descriptions within top ranked blog posts, and then to efficiently discover cross-lingual differences in concerns and opinions of blog posts in two languages.

Sample Topics
We first selected about fifty topic keywords from Wikipedia entries, where each of them has both Japanese and English entries in Wikipedia, and sufficient number of Japanese and English blog feeds can be found. Then, we manually examine both Japanese and English blog posts for each of those topic keywords. For a preliminary evaluation of this paper, we selected four topic keywords in Table 1, where, for each topic, the table shows their short descriptions, and characteristic cross-lingual differences in facts / opinions included in the retrieved blogs. Those four topic keywords are closely related to political issues and cross-lingual differences are to some extent related to differences in opinions.

1http://{en, ja}.wikipedia.org/
Procedure of Cross-lingual Blog Analysis

Blog Feed Retrieval

For the purpose of cross-lingual blog analysis, in our framework, multilingual queries for retrieving blog feeds are created from Wikipedia entries. Next, in order to collect candidates of blog feeds for a given query, in this paper, we use existing Web search engine APIs, which return a ranked list of blog posts, given a topic keyword. We use the search engine “Yahoo!” API for English, and the Japanese search engine “Yahoo! Japan” API for Japanese. Blog hosts are limited to major ones, namely, for English, 12 and for Japanese, 11. We re-rank the list of blog feeds according to the number of hits of the topic keyword in each blog feed.

Blog Post Retrieval

We automatically select blog posts that are closely related to a topic, which is given as a title of an Wikipedia entry. To do this, we first automatically extract terms that are closely related to each Wikipedia entry. More specifically, from the body text of each Wikipedia entry, we extract bold-faced terms, anchor texts of hyperlinks, and the title of a redirect, which is a synonymous term of the title of the target page. Then, blog posts which contain the topic name or at least one of the extracted related terms are automatically selected.

Extracting Characteristic Terms

Next, this section gives the procedure of how to extract characteristic terms from the blog posts retrieved according to the procedure described in the previous section. First, candidate terms are automatically selected from the selected blog posts. Here, for Japanese, noun phrases are extracted as candidate terms, while for English, sequences of one word, two words, and three words are extracted as candidate terms. Then, those candidate terms are ranked according to the following two measures, so that terms that are characteristic only in one language or in both languages are selected: a) Total frequency of each term in the whole selected blog posts. This measure is used for filtering out low
frequency terms. b) Cross-lingual rates $R_J(X_J, X_E)$ and of $R_E(Y_E, Y_J)$ term probabilities below, where term probabilities $P_J$ and $P_E$ are measured against the whole selected blog posts:

$$R_J(X_J, X_E) = \frac{P_J(X_J)}{P_E(X_E)}, \quad R_E(Y_E, Y_J) = \frac{P_E(Y_E)}{P_J(Y_J)}$$

Here, the pairs $X_J$ and $X_E$, $Y_E$ and $Y_J$ are translation pairs found through interlanguage links of Wikipedia, or those found in an English-Japanese translation lexicon Ei-jiro\(^6\). This measure is especially for mining cross-lingually characteristic terms for each language.

### Ranking Blog Feeds/Posts

Finally, we rank the blog feeds/posts in terms of the topic name and the related terms extracted from the Wikipedia entry. Here, only the blog posts that are retrieved in the previous sections are ranked, and only the blog feeds that contain such blog posts are ranked. Ranking criteria are given below: Blog posts are ranked according to the score:

$$\sum \text{weight}(type(t)) \times \text{freq}(t),$$

where $\text{weight}(type(t))$ is defined as 3 when type(t) is the topic name or the title of a redirect, as 2 when type(t) is a bold-faced term, and as 0.5 when type(t) is an anchor text of a hyperlink to another entry in Wikipedia. Blog feeds are ranked according to the total frequencies for all the blog posts ranked above, where the total frequency for each blog post is calculated as above, in terms of the topic name and the related terms.

### Visualizing Cross-Lingual/Cross-Cultural Differences through a Map of Characteristic Terms

In order to visually mine cross-lingual/cross-cultural differences in concerns and opinions in blog posts, we design an interface for selecting a topic from a category in the category list in Figure 2 (a), and a map of characteristic terms extracted from Japanese/English blog posts in Figures 2 (b). In this category list, we list about 300 categories which are placed high in the Wikipedia category hierarchy. Each category contains Wikipedia entries, where each entry can be considered as the name of a topic. Then, by selecting a topic, the map of characteristic terms is shown.

In the map of characteristic terms, a Japanese term $X_J$ with the English translation $X_E$ is plotted at the coordinate $(-R_J(X_J, X_E), P_J(X_J))$ or

### Table 2: Statistics of # (Japanese/English) of terms used for collecting blog feeds/posts, blog feeds/posts, words/morphemes

<table>
<thead>
<tr>
<th>Topic</th>
<th># of topic-related terms from Wikipedia</th>
<th># of blog feeds</th>
<th># of blog posts</th>
<th># of total words/morphemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whaling</td>
<td>162 / 174</td>
<td>121 / 239</td>
<td>2232 / 6532</td>
<td>5024966 / 2611942</td>
</tr>
<tr>
<td>Organ transplant</td>
<td>100 / 231</td>
<td>89 / 206</td>
<td>696 / 1301</td>
<td>995927 / 781476</td>
</tr>
<tr>
<td>Tobacco smoking</td>
<td>399 / 276</td>
<td>86 / 252</td>
<td>1481 / 400</td>
<td>1323767 / 492727</td>
</tr>
<tr>
<td>Subprime lending</td>
<td>39 / 68</td>
<td>134 / 205</td>
<td>1088 / 1216</td>
<td>980552 / 883450</td>
</tr>
</tbody>
</table>

\(^6\)http://www.eijiro.jp/, Ver.79, with 1.6M translation pairs.

---

**Table 1: Sample Topics used in the Evaluation and their Descriptions**

<table>
<thead>
<tr>
<th>Topic — Short Description</th>
<th>Differences in Facts/Opinions</th>
<th>(English Blogs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whaling — There are arguments for and against whaling.</td>
<td>Most blogs are against whaling, especially, whaling in Japan. Some are blogs for whale watching.</td>
<td></td>
</tr>
<tr>
<td>Most blogs are for whaling. Some of them are nationalistic.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organ transplant — A medical operation for the purpose of replacing damaged organ with a working one from the donor’s body.</td>
<td>Many blogs strongly recommend donor registration because of shortage of organs for patients. Some blogs are criticizing illegal transplant.</td>
<td></td>
</tr>
<tr>
<td>Many blogs point out that Organ Transplant Law of Japan should be revised. Some blogs are picking up the news about transplant by the Japanese doctor using diseased kidney.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tobacco smoking — The fact that smoking does harm to health is mostly argued.</td>
<td>Most bloggers are against smoking because it may cause lung cancer.</td>
<td></td>
</tr>
<tr>
<td>Although most bloggers are against smoking, one or two blogger(s) are for smoking.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subprime lending — Beginning in late 2006, the U.S. subprime mortgage industry entered what many observers have begun to refer to as a meltdown.</td>
<td>Financial analysts argue issues of subprime problems, housing bubble, and the resulting financial crisis.</td>
<td></td>
</tr>
<tr>
<td>Most bloggers argue influences of the U.S. subprime problem on Japanese economy.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Visualizing Cross-Lingual/Cross-Cultural Differences through a Map of Characteristic Terms**

In order to visually mine cross-lingual/cross-cultural differences in concerns and opinions in blog posts, we design an interface for selecting a topic from a category in the category list in Figure 2 (a), and a map of characteristic terms extracted from Japanese/English blog posts in Figures 2 (b). In this category list, we list about 300 categories which are placed high in the Wikipedia category hierarchy. Each category contains Wikipedia entries, where each entry can be considered as the name of a topic. Then, by selecting a topic, the map of characteristic terms is shown.

In the map of characteristic terms, a Japanese term $X_J$ with the English translation $X_E$ is plotted at the coordinate $(-R_J(X_J, X_E), P_J(X_J))$ or
at \((-\text{maximum rate in the map}), P_J(X_J)\) (when \(P(X_E) = 0\)). Similarly, an English term \(X_E\) with the Japanese translation \(X_J\) is plotted at the coordinate \((R_E(X_E, X_J), P_E(X_E))\) or at \((\text{maximum rate in the map}, P_E(X_E))\) (when \(P(X_J) = 0\)). In the map, several terms which have relatively high point-wise mutual information are grouped together along with an excerpt from typical posts including those terms. Terms that are characteristic only in one language tend to be plotted apart from Y-axis, together with excerpts typical only in one language. On the other hand, terms that are characteristic in both languages tend to be plotted close to Y-axis, together with excerpts typical in both languages.

The followings roughly summarize the findings for some of the four topics. For the topic “Whaling”, many terms which are characteristic in English blogs represent against-whaling opinion. On the other hand, many terms which are characteristic in Japanese blogs are those for expressing criticism against anti-whaling activities in Australia. For the topic “Organ transplant”, many terms which are characteristic in English blogs represent opinions against illegal organ transplant. On the other hand, many terms which are characteristic in Japanese blogs are closely related to kidney transplant using diseased kidney.

Based on those observation results, we can argue that major contribution of this paper is that we successfully invent a framework of visualizing cross-lingual differences of cultural concerns in blogs of two languages. It can be obviously seen from the results shown in this section that one of most important near future works is to incorporate multilingual sentiment analysis techniques such as those previously studied in (Evans et al. 2007; Wiebe, Wilson, and Cardie 2005). Then, it will become for us to easily classify those top ranked blog posts and feeds into for, neutral, and against with respect to the issue of the given topic. For example, for the topic “Whaling”, both languages share a certain set of characteristic terms, where opinions discovered with those shared characteristic terms from the two languages are quite opposite. Such differences in opinions should be typically discovered automatically by incorporating multilingual sentiment analysis techniques.

**Conclusion**

This paper proposed how to cross-lingually analyze multilingual blogs collected with a topic keyword. In addition to proposing an interface for visualizing cross-lingual/cross-cultural differences in concerns and opinions in blogs of two languages, this paper showed the effectiveness of the proposed framework with detailed examples of efficiently mining and comparing cross-lingual differences in concerns and opinions. Future works for cross-lingual blog analysis on facts and opinions include incorporating multilingual sentiment analysis techniques.

**References**


