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

Weblogs 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 versions[1]) 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 anlaysis 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.

[1] http://{en,ja}.wikipedia.org/

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.

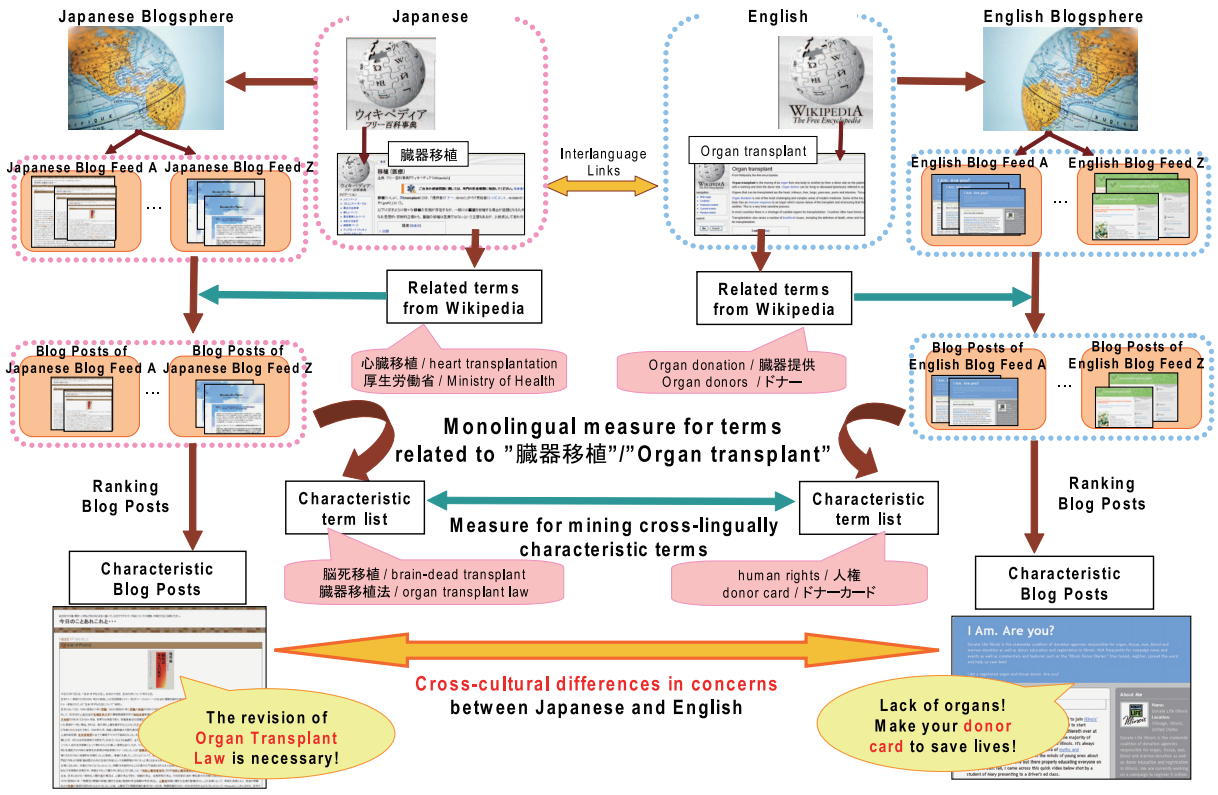


Figure 1: Overall Framework of Cross-Lingual Blog Analysis

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 [2] for English, and the Japanese search engine "Yahoo! Japan" API [3] for Japanese. Blog hosts are limited to major ones, namely, 12 for English[4] and 11 for Japanese[5]. 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.

For each topic, Table 2 shows the numbers of terms that are closely related to the topic and are extracted from each Wikipedia entry. Then, according to the above procedure, blog posts which contain the topic name or at least one of the extracted related terms are automatically selected. Table 2 also shows the numbers of the selected blog posts, as well as those of blog feeds for those posts and the total numbers of words/morphemes contained in those posts.

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 extracted 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

[2] http://www.yahoo.com/

[3] http://www.yahoo.co.jp/ (in Japanese)

[4] blogspot.com, msnblogs.net, spaces.live.com, livejournal.com, vox.com, multiply.com, typepad.com, aol.com, blogsome.com, wordpress.com, blog-king.net, blogster.com

[5] FC2.com, yahoo.co.jp, rakuten.ne.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa.net, jugem.jp, yaplog.jp, webry.info.jp, hatena.ne.jp

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

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

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

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

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:

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

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 Eijiro[6]. This measure is especially for mining cross-lingually characteristic terms for each language.

Ranking Blog Feeds/Posts

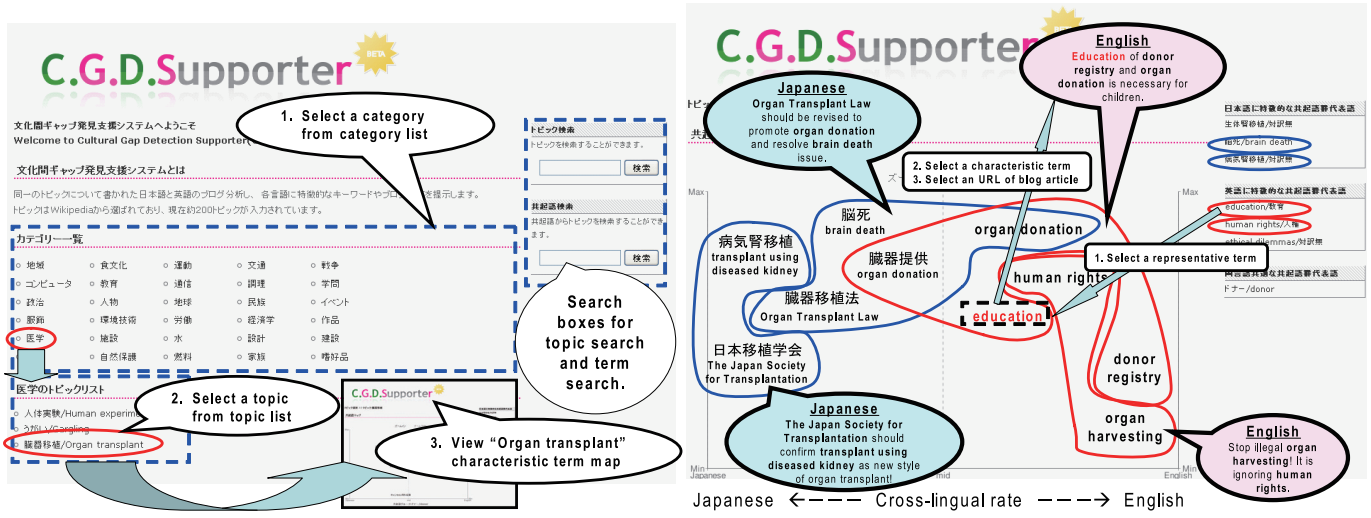
Finally, we rank the blog feeds/posts in terms of the topic name and the related terms extracted from the Wikipedia entry. Here, only the blog posts that are retrieved in the previous sections are ranked, and only the blog feeds that contain such blog posts are ranked. Ranking criteria are given below: Blog posts are ranked according to the score: $\sum_t weight(type(t)) \times freq(t)$, where $weight(type(t))$ 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.

[6] http://www.eijiro.jp/, Ver.79, with 1.6M translation pairs.

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 Wikipeida 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



(a) An Interface for Selecting a Topic

(b) A Map of Characteristic Terms for Visually Mining Cross-Lingual/Cross-Cultural Differences ("Organ transplant")

Figure 2: An Interface for Visualizing Cross-Lingual/Cross-Cultural Differences in Concerns

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 (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 in 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

Evans, D. K.; Ku, L.-W.; Seki, Y.; Chen, H.-H.; and Kando, N. 2007. Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task. In *Proc. 3rd Inter. Cross-Language Information Processing Workshop (CLIP2007)*, 456–463.

Fukuhara, T.; Utsuro, T.; and Nakagawa, H. 2007. Cross-Lingual Concern Analysis from Multilingual Weblog Articles. In *Proc. 6th Inter. Workshop on Social Intelligence Design*, 55–64.

Macdonald, C.; Ounis, I.; and Soboroff, I. 2007. Overview of the TREC-2007 Blog Track. In *Proc. TREC-2007 (Notebook)*, 31–43.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2-3):165–210.