Cross-Lingual Blog Analysis by Cross-Lingual Comparison of Characteristic Terms and Blog Posts

Hiroyuki Nakasaki* Mariko Kawaba* Takehito Utsuro*
Tomohiro Fukuhara‡ Hiroshi Nakagawa‡ Noriko Kando♯

*University of Tsukuba, Tsukuba, 305-8573, JAPAN ‡University of Tokyo, Kashiwa/Tokyo, 277-8568/113-0033, JAPAN

♯National Institute of Informatics, Tokyo, 101-8430, 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 cross-lingually and cross-culturally compare less well known facts and opinions that are closely related to a given topic. Preliminary evaluation results support the effectiveness of the proposed framework.

1 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., [3]). 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 [4]. 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, we cross-lingually and cross-culturally compare less well known facts and much more radical opinions that are closely related to a given topic.

In addition to proposing the overall framework of cross-lingual and cross-cultural comparison of concerns and opinions in blogs in two languages, this paper shows the effectiveness of the proposed framework with detailed examples of efficiently mining and comparing cross-lingual differences in concerns and opinions.

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

2 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 those characteristic terms in blog posts in both languages, characteristic blog posts are ranked. Finally, by manually analyzing top ranked blog posts, we can efficiently discover cross-lingual differences in concerns and opinions of blog posts in two languages.

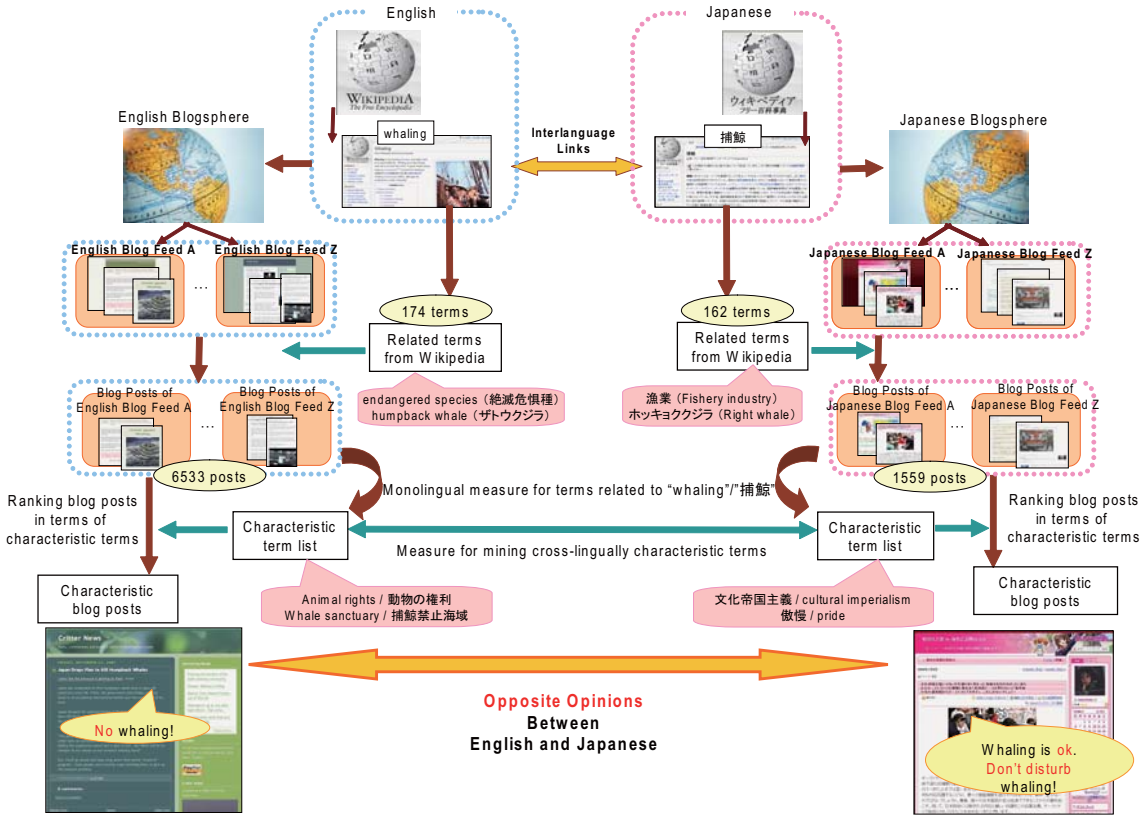


Figure 1. Overall Framework of Cross-Lingual Blog Analysis

3 Related Works

3.1 Sentiment and Concern Analysis in Multilingual News

There exist several works on studying cross-lingual analysis of sentiment and concerns in multilingual news [7, 5, 8, 1]. [7] studied how to combine reports on epidemic threats from over 1,000 portals in 32 languages. [5] studied how to combine name mentions in news articles of 32 languages. [8] also studied mining comparative differences of concerns in news streams from multiple sources. [1] studied how to analyze sentiment distribution in news articles across 9 languages. Those previous works mainly focus on news streams and documents other than blogs.

3.2 Blog Distillation Task in TREC 2007 Blog Track

The Blog distillation task [4] can be summarized as *Find me a blog with a principle, recurring interest in X*. For a given target X, systems should suggest feeds that are principally devoted to X over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about X (i.e. a user may be interested in adding it to their RSS reader). As reported in [4], for most participants, best performance is achieved by creating queries only from the title of a retrieval topic. Based on this result, in the preliminary evaluation of this paper, we simply use the titles of Wikipedia entries in each language as retrieval queries of multilingual blog distillation.

4 Procedure of Cross-lingual Blog Analysis

4.1 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. This section briefly describes how to retrieve blog feeds given a query for each language (in this paper, English and Japanese).

First, 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"

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

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

English Topic (Japanese Topic)	Short Description	
	Differences in Facts/Opinions	
	(English Blogs)	(Japanese Blogs)
Whaling (捕鯨)	There are arguments *for* and *against* whaling.	
	Most blogs are *against* whaling, especially, whaling in Japan. Some are blogs for whale watching.	Most blogs are *for* whaling. Some of them are nationalistic.
Organ transplant (臓器移植)	A medical operation for the purpose of replacing damaged organ with a working one from the donor's body.	
	Many blogs strongly recommend donor registration because of shortage of organs for patients. Some blogs are criticizing Chinese illegal transplant.	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.
Ethanol fuel (バイオエタノール)	There is a discussion whether ethanol fuel can be used as new fuel instead of gasoline or not.	
	There are arguments for and against ethanol fuel.	Most blogs are against ethanol fuel. Some blogs point out that mass production of ethanol fuel will cause dramatic increase of food price.
Genetically modified organism (遺伝子組換え食品)	An organism which its genetic information has been modified by genetic engineering. Also known as GMO. GM food is made from it.	
	There are arguments for and against GMO. Some blogs point out that GMO production may cause environment disruption.	There are arguments for and against GMO. Some blogs point out that some people think GMO is dangerous without any evidence.
Abortion (中絶)	The removal of a fetus from a parent body. There are arguments *for* and *against* it.	
	Most blogs are against abortion. Some blogs discuss about Partial-birth Abortion Ban Act.	Some blogs talk about blogger's abortion experience. There are few arguments about abortion.

API [3] for Japanese. Blog hosts are limited to major ones, namely, 12 for English[4] and 11 for Japanese[5].

Next, we employ the following procedure for the blog distillation:

i) Given a topic keyword, a ranked list of blog posts are returned by a Web search engine API.
ii) A list of blog feeds is generated from the returned ranked list of blog posts by simply removing duplicated feeds.
iii) Re-rank the list of blog feeds according to the number of hits of the topic keyword in each blog feed. The number of hits for a topic keyword in each blog feed is simply measured by the search engine API used for collecting blog posts above in i), restricting the domain of the URL to each blog feed.

[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`

4.2 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.

4.3 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, two or three word sequences are extracted as candidate terms. Then,

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

English Topic / Japanese Topic	# of terms		# of blog feeds	# of blog posts	# of total words/morphemes
	Topic-related terms from Wikipedia	characteristic terms (manually selected)			
Whaling / 捕鯨	174 / 162	56 / 50	239 / 121	6532 / 2232	2611942 / 5024966
Organ transplant / 臓器移植	231 / 100	55 / 68	206 / 89	1301 / 696	781476 / 995927
Ethanol fuel / バイオエタノール	289 / 46	57 / 50	20 / 108	178 / 657	216138 / 699580
Genetically modified organism / 遺伝子組換え食品	93 / 51	50 / 50	132 / 32	230 / 205	285693 / 191839
Abortion / 中絶	409 / 77	41 / 53	27 / 75	3477 / 731	1430019 / 1839281

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. In our evaluation, for each topic, we use the top 300 frequent terms.

b) Cross-lingual rates $R_E(X_E, X_J)$ and R_E of term probabilities below, where term probabilities P_E and P_J are measured against the whole selected blog posts:

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

Here, the pairs X_E and X_J, Y_J and J_E 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.

Certain number of terms do not have translation into the other language, or even if they have translation in Wikipedia or Eijiro, the translation does not appear in the whole selected blog posts in the other language. Although some of such terms are extremely characteristic only for one language, most of other terms are noises that should be ignored here. Therefore, we do not consider all of such terms when ranking them according to the cross-lingual rates here[7].

In our evaluation, for each topic, we use the top 300 highly ranked terms.

After ranking candidate terms according to the above two measures, those ranked lists are shown to an operator (one of the authors of this paper), and the operator manually selects characteristic terms for each topic. Those manually selected characteristic terms are also evaluated and compared with automatically ranked

4.4 Ranking Blog Feeds/Posts

Finally, we use the characteristic terms extracted in the previous section to rank the blog feeds/posts. Here, only the blog posts that are retrieved in section 4.2 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 total frequencies of all the characteristic terms.
- Blog feeds are ranked according to the total frequencies of all the characteristic terms included in blog posts ranked above.

5 Evaluation

5.1 Topics for Evaluation

We first selected about fifty topic keywords from Wikipedia entries, where each of them originated from Japan, but has its English entry in Wikipedia, and has been to some extent popular abroad (mainly, in United States)

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

[7] Among such terms, those which are extremely characteristic only for one language tend to have high frequencies, and thus are ranked high in the ranking a) according to their total frequencies.

Table 3. Samples of Characteristic Terms Extracted from English/Japanese Blog Posts

(a) Whaling / 捕鯨

Ranked by Freq. in EN blogs						Ranked by Cross-lingual rates $R_E(X_E,X_J)$					
Rank	EN Term	Freq. in EN blogs	JP Translation	Freq. in JP blogs	$R_E(X_E,X_J)$	Rank	EN Term	Freq. in EN blogs	JP Translation	Freq. in JP blogs	$R_E(X_E,X_J)$
13	endangered species	302	絶滅危惧種	47	12.362	11	whaling sanctuary	101	捕鯨禁止海域	2	97.154
15	humpback whales	266	ザトウクジラ	98	5.222	30	animal welfare	188	動物保護	7	51.669
19	animal rights	233	動物の権利	0	∞	242	whale watching	117	ホエールウォッチング	18	12.505
Ranked by Freq. in JP blogs						Ranked by Cross-lingual rates $R_J(X_J,X_E)$					
Rank	JP Term	Freq. in JP blogs	EN Translation	Freq. in EN blogs	$R_J(X_J,X_E)$	Rank	JP Term	Freq. in JP blogs	EN Translation	Freq. in EN blogs	$R_J(X_J,X_E)$
62	調査捕鯨	620	research whaling	59	5.462	–	妨害	243	active jamming	0	∞
130	法律	423	statute	23	9.56	–	反日	231	No Translation (anti-Japanese)	–	∞
212	民族	309	Ethnic group	7	22.945	–	反捕鯨国	136	antiwhaling country	0	∞

(b) Organ transplant / 臓器移植

Ranked by Freq. in EN blogs						Ranked by Cross-lingual rates $R_E(X_E,X_J)$					
Rank	EN Term	Freq. in EN blogs	JP Translation	Freq. in JP blogs	$R_E(X_E,X_J)$	Rank	EN Term	Freq. in EN blogs	JP Translation	Freq. in JP blogs	$R_E(X_E,X_J)$
2	organ donation	673	臓器提供	200	4.288	1	Falun Gong	1132	法輪功	4	360.66
6	Falun Gong practitioners	444	No Translation	–	∞	5	human rights	508	人権	9	71.934
7	organ harvesting	270	No Translation	–	∞	20	living donor	61	生体ドナー	3	25.913
Ranked by Freq. in JP blogs						Ranked by Cross-lingual rates $R_J(X_J,X_E)$					
Rank	JP Term	Freq. in JP blogs	EN Translation	Freq. in EN blogs	$R_J(X_J,X_E)$	Rank	JP Term	Freq. in JP blogs	EN Translation	Freq. in EN blogs	$R_J(X_J,X_E)$
17	脳死	541	brain death	92	4.614	2	副作用	328	adverse effect	1	257.37
23	病気腎移植	442	No Translation (transplant using diseased kidney)	–	∞	32	臓器移植法	123	Organ Transplant Law	3	32.172
29	脳死移植	366	brain-dead transplant	0	∞	34	日本臓器移植ネットワーク	38	Japan Organ Transplant Network	1	29.818

and sufficient number of 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 five 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 five topic keywords are closely related to political issues and cross-lingual differences are to some extent related to differences in opinions.

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 procedure given in section 4.2, 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. Finally, as in section 4.3, candidates of characteristic terms are automatically extracted and then, are manually selected, where their numbers are also shown in Table 2.

5.2 Extracting Characteristic Terms

Table 3 shows excerpts of manually selected characteristic terms, as well as their ranking with respect to frequencies or cross-lingual rates $R_E(X_E, X_J)$. Here, most of them are characteristic only in one language, while some are characteristic in both languages. Those terms as well as their cross-lingual rates $R_E(X_E, X_J)$ are useful signal for estimating differences in concerns and opinions in English and Japanese. The followings roughly summarize the findings reported by the operator.

For the topic "Whaling", almost all the terms which are characteristic in English blogs represent against-whaling opinion. On the other hand, almost all the 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

Table 4. Top 10 Ranked English/Japanese Blog Posts ("Whaling" and "Organ transplant")

English Topic (Japan -ese Topic)	English				Japanese			
	rank of posts	total freqs of character-istic terms	author ID as feed rank	description	rank of posts	total freqs of character -istic terms	author ID as feed rank	description
Whaling (捕鯨)	1	87	17th	*against* whaling. Support Sea Shepherd.	1,4,9,10	730,309, 196,194	1st	*for* whaling. Criticizing anti-whaling groups.
	3,4,7	55,52,45	4th	*neutral* with respect to whaling. Author lives in Japan for 30 years.	2,7,8	473,199,196	4th	mainly on American-Japanese relation info. not on whaling
	5,10	50,38	6th	*against* whaling. Support Sea Shepherd.	5	238	6th	*for* whaling. *against* Australia's anti-whaling thought.
Organ trans-plant (臓器移植)	1,3,9,10	379,85,60,59	2nd	Criticizing Chinese illegal organ transplant.	1,2,4,5,8	496,488, 446,435, 297	2nd	Picking up news about kidney transplant using diseased kidney. Criticizing The Japan Society for Transplantation.
	2,7	87,65	5th	Quoting an article *against* Chinese illegal organ transplant.	3,7,9,10	480,340, 283,272	1st	Picking up news article related to medication.
	5	67	4th	Quoting an article *for* improved organ allocation system.	6	412	6th	Recommend organ transplant in China, which is not illegal.

Chinese illegal organ transplant. On the other hand, many terms which are characteristic in Japanese blogs are closely related to kidney transplant using diseased kidney.

5.3 Ranking Blog Feeds/Posts

For the topics "Whaling" and "Organ transplant", Table 4 lists the summary of top 10 ranked blog posts, along with total frequencies of characteristic terms in each post, blog author identity as its rank in blog feed ranking. Although a few posts are not on the given topic, most other posts are deeply concerned with the given topic, and furthermore, represent clear opinions on *for* or *against* the issue of the given topic[8].

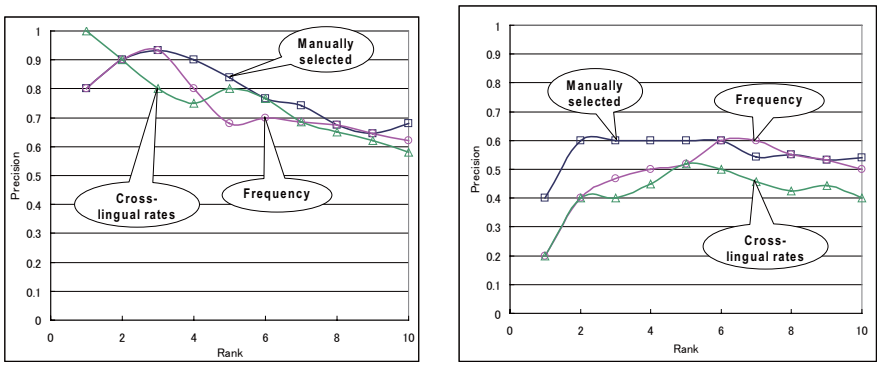
Furthermore, Figures 2 and 3 show evaluation results of ranking English/Japanese blog feeds/posts. Here, we compare three types of characteristic terms, namely, the top 300 frequent terms, the top 300 highly ranked terms in terms of the cross-lingual rates $R_E(X_E, X_J)$, and those manually selected terms. Each blog feed/post is judged whether it is relevant to the given topic. It is remarkable to note here that, in terms of ranking blog feeds, automatically extracted characteristic terms perform almost comparable to manually selected characteristic terms.

Based on those evaluation results, we strongly argue that major contribution of this paper is that we successfully invent an automatic technique of mining cross-lingual differences of opinions and cultural concerns in blogs of two languages. It can be obviously seen from the evaluation 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 [2, 6]. 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.

6 Conclusion

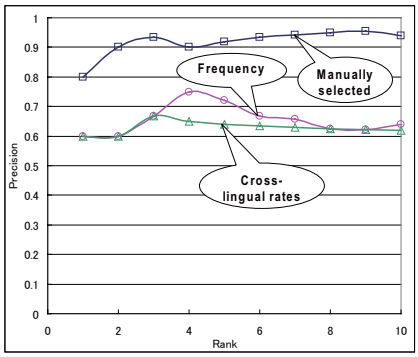
This paper proposed how to cross-lingually analyze multilingual blogs collected with a topic keyword. In addition to proposing the overall framework of cross-lingual and cross-cultural comparison of concerns and opinions in blogs in two languages, this paper showed the effectiveness of the proposed framework with detailed examples of efficiently mining and comparing cross-lingual differences in concerns

[8]For the topic "Whaling", among bloggers whose blog is in English, only one blogger is rather neutral, where his blog is full of comments opposing the claim that anti-whaling against Japan is a kind of racism.

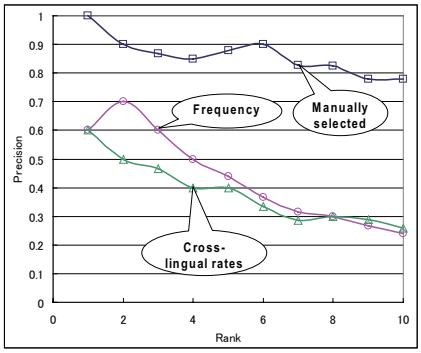


(a) English feed

(b) Japanese feed



(c) English post



(d) Japanese post

Figure 2. Performance of Blog feeds/posts Ranking: Average for 5 Topics

and opinions. Future works for cross-lingual blog analysis on facts and opinions include incorporating multilingual sentiment analysis techniques.

References

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proc. ICWSM*, pages 19–26, 2008.

[2] D. K. Evans, L.-W. Ku, Y. Seki, H.-H. Chen, and N. Kando. 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)*, pages 456–463, 2007.

[3] T. Fukuhara, T. Utsuro, and H. Nakagawa. Cross-lingual concern analysis from multilingual weblog articles. In *Proc. 6th Inter. Workshop on Social Intelligence Design*, pages 55–64, 2007.

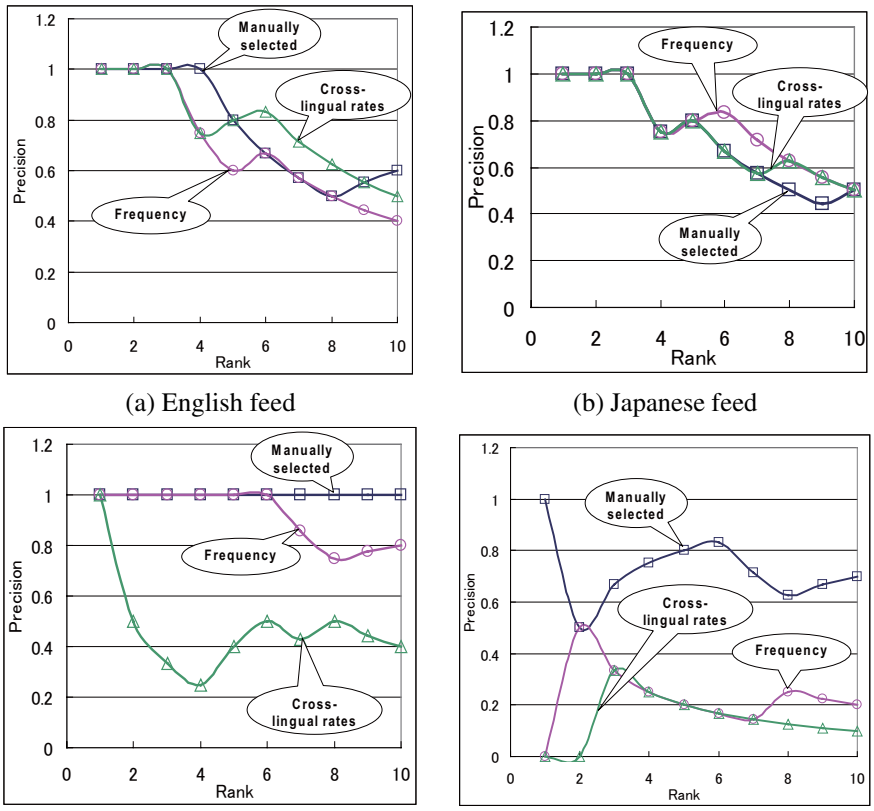
[4] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 blog track. In *Proc. TREC-2007 (Notebook)*, pages 31–43, 2007.

[5] B. Pouliquen, R. Steinberger, and J. Belyaeva. Multilingual multi-document continuously-updated social networks. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pages 25–32, 2007.

[6] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[7] R. Yangarber, C. Best, P. von Etter, F. Fuart, D. Horby, and R. Steinberger. Combining information about epidemic threats from multiple sources. In *Proc. Workshop: Multi-source, Multilingual Information Extraction and Summarization*, pages 41–48, 2007.

[8] M. Yoshioka. IR interface for contrasting multiple news sites. In *Prof. 4th AIRS*, pages 516–521, 2008.



(a) English feed

(b) Japanese feed

(c) English post

(d) Japanese post

Figure 3. Performance of Blog feeds/posts Ranking: Sample for "Whaling"