Cross-Lingual Blog Analysis based on Multilingual Blog Distillation from Multilingual Wikipedia Entries

Mariko Kawaba[⋆] **Hiroyuki Nakasaki**[⋆] **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 distillation (feed search) procedure. Mulitlingual 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.

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, & 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 distillation (feed search) procedure recently studied in TREC 2007 Blog track as one of its task (Macdonald, Ounis, & 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, mulitlingual queries for retrieving blog feeds are created from *Wikipedia* (English and Japanese versions http://{en,ja}.wikipedia.org/) 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 distillation framework in more 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.

Multilingual Blog Distillation from Multilingual Wikipedia Entries

Blog Distillation Task in TREC 2007 Blog Track

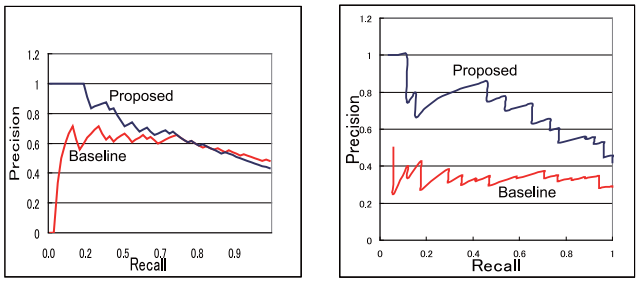
The Blog distillation task (Macdonald, Ounis, & Soboroff 2007) can be summarized as *Find me a blog with a principle, recurring interest in X*. For a given target X, systems should suggest feeds that are principally devoted to X over the timespan of the feed, and would be recommended to subscribe to as an interesting feed about X. As reported in (Macdonald, Ounis, & Soboroff 2007), 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.

Multilingual Blog Distillation

For the purpose of cross-lingual blog analysis, in our framework, mulitlingual 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, we use existing Web search engine APIs ("Yahoo!" API (http://www.yahoo.com/) for English, and "Yahoo! Japan" API (http://www.yahoo.co.jp/) (in Japanese) for Japanese), which return a ranked list of blog posts, given a topic keyword. Blog hosts are limited to major ones, namely, 12 for English and 11 for Japanese. 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 ranked list of blog feeds is generated from the returned ranked list of blog posts by simply removing duplicated feeds, keeping the original ranking of blog posts. (This original ranking of blog feeds are used as baseline.)

English Topic (Japanese Topic)	Short Description Differences in Facts/Opinions (English Blogs)	(Japanese Blogs)
Dragon Ball	A Japanese manga series. Very popular in over 40 countries.	
	Few blogs are about carddas. Instead, many blogs are with reviews of games, and some are with videos uploaded.	Many blogs are about carddas. Some are with reviews of game, and few are with videos uploaded because of Japanese legal regulation.
Wii (Wii)	A video game console recently released by Nintendo. Sold all over the world.	
	Some blogs are about Wii hack. Many blogs are with videos and images of Wii. Some blogs are about games.	No blogs are about Wii hack because of Japanese legal regulation. Many blogs are about games and their walk-through.
Neon Genesis Evangelion	A Japanese anime. Its international releases include an English version as well as those in Europe, Latin America, Asia.	
	Many blogs are with videos and images. Some blots are with personal reviews. No blogs by people playing with pachinko, no splogs.	Many blogs are with personal reviews. Many blogs are by people playing with pachinko gaming devices with Evangelion contents. Some splogs linked to affiliated sites selling Evangelion goods.
Whaling	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.
Yasukuni Shrine	A Shinto shrine located in Tokyo, Japan. Included in the Book of Souls are 1,068 people convicted of war crimes by a post World War II court. Visits to the shrine by cabinet members, and various Prime Ministers in particular, have been a cause of protest at home and abroad.	
	Most blogs are *against* visits to the shrine by cabinet members.	Most blogs are nationalistic, right-wing, and *for* visits to the shrine by cabinet members.

Table 1: Samples of Cross-Lingual Blog Analysis



(a) English (b) Japanese

Figure 1: Performance of Blog Distillation: Sample for "Yasukuni Shrine"

iii) Re-rank the list of blog feeds according to the number of hits of the topic keyword in each blog feed.

Preliminary 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) 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, Japanese and English blog posts show clear differences in facts and opinions included in those posts. In the process of manual relevance judgements, we make the criterion of the TREC 2007 Blog distillation task mentioned above less strict[1]. For the topic "Yasukuni Shrine", Figure 1 gives recall-precision curves for both English and Japanese.

Cross-Lingual Blog Analysis

For the five topic keywords selected above, Table 1 shows English and Japanese topic keywords, their short descriptions, and characteristic cross-lingual differences in facts / opinions included in the retrieved blogs. The first three topic keywords are from entertainment genre and most differences are in cultural facts. The last two are more closely related to political issues and cross-lingual differences are directly related to the polarity in opinions.

Conclusion

This paper proposed a multilingual blog distillation framework and employed it in a more higher level application of cross-lingual blog analysis. Future works for cross-lingual blog analysis on facts and opinions include linking Wikipedia entries to multilingual blog feeds in a much larger scale.

References

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.

[1]The whole of each feed is not necessarily principally devoted to X over the timespan of the feed, but at least certain portion of it is principally devoted to X within certain timespan of the feed. For example, feeds which have a category closely related to X.