Cross-Lingual Concern Analysis from Multilingual Weblog Articles

Tomohiro Fukuhara
RACE (Research into Artifacts), The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa, Chiba JAPAN
http://www.race.u-tokyo.ac.jp/~fukuhara/

Takehito Utsuro
Graduate School of Systems and Information Engineering, Tsukuba University
1-1-1 Tennodai, Tsukuba, Ibaraki JAPAN
http://nlp.iit.tsukuba.ac.jp/member/utsuro/

Hiroshi Nakagawa
Information Technology Center, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo JAPAN
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/English.html

Abstract

Cross-lingual concern analysis system from multilingual Weblog (blog) articles is proposed. With the wide spread of the Internet, multilingual documents, especially blog articles written in various languages, are appearing on the Internet. We can find concerns of people who are living in various regions and countries from these blog articles. By comparing various concerns across languages, we can find various viewpoints on a topic. The aim of this research is to facilitate people to find differences of concerns from multilingual blog articles. We propose a cross-lingual concern analysis system called KANSHIN that collects and analyzes Chinese, Japanese, Korean, and English (CJKE) blog articles. Users can find differences of focuses on a topic across languages because the system automatically translates keywords into other languages and retrieve articles in each language. An overview of the system, and preliminary results are described.

Keywords: Cross-lingual concern analysis, CJKE blog analysis, multilingual Web

1 Introduction

Today many people communicate with others across regions and countries. Although English is one of major languages on the current Internet, we consider that documents written in various languages would appear in the near future. We call this phenomenon the *multilingual Web*. The tendency of the multilingual Web can be seen at (1) Wikipedia, and in the (2) blogosphere.

In Wikipedia[1], which is a large-scale public encyclopedia on the Internet, there are 6 million articles, and these articles are written in 250 languages[2]. Top 5 languages used in Wikipedia are English (1,663,419 articles), Germany (549,653 articles), French (453,201 articles), Polish (354,394 articles), and Japanese (334,237 articles)[3]. Although there are articles written in several languages on the same topic[4], contents are different by languages. By comparing these differences among languages, we can find various viewpoints for that topic.

[1] http://www.wikipedia.org/
[2] http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics
[3] On March 1st, 2007. (from http://en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics)
[4] For example, descriptions about Shinzo Abe, who is the current Prime Minister of Japan, are different by languages, especially in Japanese, Chinese, and Korean.

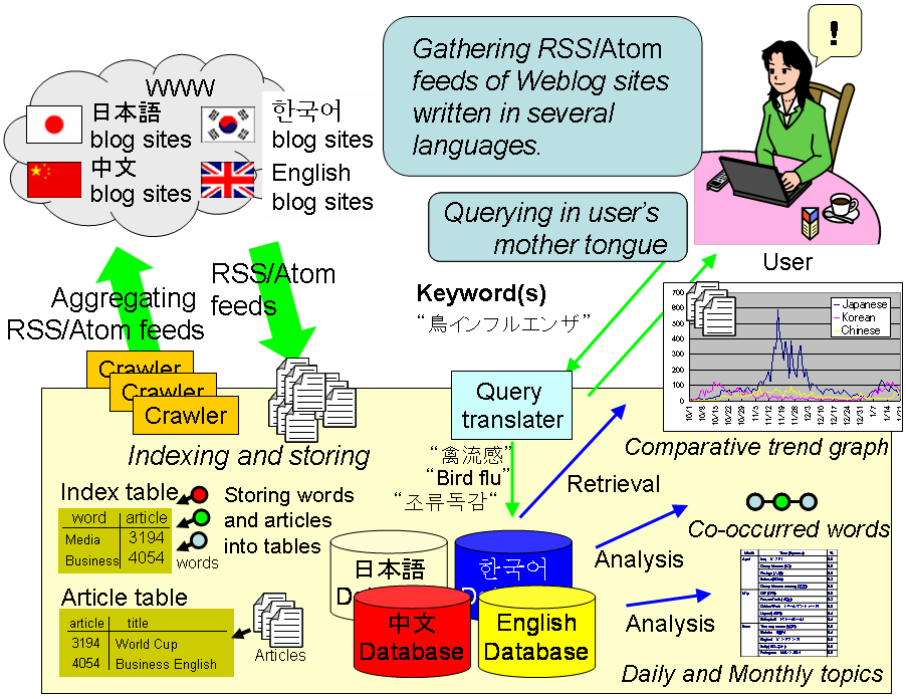

Figure 1: Overview of the cross-lingual concern analysis system using multilingual (CJKE) blog articles.

Another example is the blogosphere. Today, many people read and write blog articles around the world. According to the Technorati's report[5] that analyzes the state of blogosphere in April, 2007, several languages are used in blog articles such as Japanese(37%), English(36%), Chinese(8%), Italian(3%), Spanish(3%), Russian(2%), French(2%) and so on. From these multilingual blog articles, we can find differences of concerns of people on a topic because concerns of people are different by countries and language communities. If we can find differences of concerns across countries or languages, it is useful not only for mutual understanding, but also for solving social problems such as global warming, world poverty, war, conflicts, and so on.

The aim of this research is to facilitate people to find and compare concerns on a same topic across languages. We propose a cross-lingual concern analysis system called KANSHIN that collects and analyzes multilingual blog articles. Figure 1 shows an overview of the system. The system collects multilingual blog articles from Chinese, Japanese, Korean, and English blog sites, and extract keywords from articles, and provides functions for (1) cross-lingual retrieval, (2) finding co-occurred words with a keyword given by a user, and (3) finding daily and monthly topics. Because we aim to facilitate users to find and compare concerns of people across languages, the system translates keywords written in a language into other languages automatically, and retrieves articles across languages.

This paper consists of following sections. Section 2 describes previous work on blog analysis, a definition of a concern in this paper, and requirements for a cross-lingual concern analysis system. Section 3 describes an overview of the prototype system. Section 4 describes preliminary results obtained from the system. In Section 5, we discuss issues to be solved, and related work. In Section 6, we summarize arguments, and describe future work.

[5]http://www.sifry.com/alerts/archives/000493.html

2 Previous work

In this section, we describe previous work on blog analysis systems, a definition of a concern, and requirements for a cross-lingual concern analysis system.

2.1 Previous work

There are several previous work and services on blog analysis systems. Nanno et al. (2004) proposed a system called *blogWatcher* that collects and analyzes Japanese blog articles. Glance et al. (2004) proposed a system called *BlogPulse* that analyzes trends of blog articles. With respect to blog analysis services on the Internet, there are several commercial and non-commercial services such as *Technorati*[6], *BlogPulse*[7], *kizasi.jp*[8], and *blogWatcher*[9]. These services, however, analyze mono-lingual blog articles, i.e., they analyze English or Japanese or other specific language blog articles.

With respect to multilingual blog services, *Globe of Blogs*[10] provides a retrieval function of blog articles across languages. *Best Blogs in Asia Directory*[11] also provides a retrieval function for Asian language blogs. These services are not cross-lingual services. *Blogwise*[12] analyzes multilingual blog articles, but cross-lingual analysis is not realized.

In the language grid project, Ishida and his colleagues aim to support cross-lingual or cross-cultural communication by using computers (Ishida (2006)). Their focuses are on communicative aspect of human to human communication. This approach is important in the real-world communication and computer-mediated communication (CMC) situations using video/audio conferencing tools. On the other hand, our focus is not on communicative aspect, but on analytical aspect of concerns of people across languages.

Cross-lingual blog analysis is needed for understanding differences of viewpoints on a topic. In this research, we focus on a cross-lingual concern analysis that can retrieve and analyze blog articles written in several languages.

2.2 Definition of a concern

We define a word *concern*, and describe a relation between a concern and co-occurred words with a keyword. A concern in this paper is a set of keywords that characterizes documents or people. For example, *folksonomy*, which is an ontology created and maintained in a community, can be seen as an example of a concern because keywords contained in folksonomy represent the interests of the community. We can find concerns of people by extracting keywords that are appeared frequently in blog articles.

In the context of a cross-lingual concern analysis, finding common concerns across languages, and finding unique concerns in a specific language are important. For common (international or multilingual) concerns, 'bird flu (avian influenza)', 'SARS (severe acute respiratory syndrome)', 'terrorism', and 'global warming' might be included. For unique (domestic or monolingual) concerns, cultural events in each country, and domestic problems in a country might be included. For finding unique concerns, we can find unique concerns as daily or monthly topics in our prototype system (Fukuhara et al. (2005)). For finding common concerns, the system has no functions at this moment. As future work, we will develop a function for finding common concerns across languages.

With respect to the relation between a concern and co-occurred words with a keyword, we consider that their relation represent a *topic and sub-topic* relation. For example, a topic 'disaster' might contain sub-topics such as 'earthquake', 'typhoon', 'tsunami' and so on. We can find co-occurred words with a keyword from blog articles. Co-occurred words are important for finding

[6] http://technorati.com/
[7] http://www.blogpulse.com/
[8] http://kizasi.jp/ (in Japanese)
[9] http://blogwatcher.pi.titech.ac.jp/ (in Japanese)
[10] http://www.globeofblogs.com/
[11] http://www.misohoni.com/bba/
[12] http://www.blogwise.com/

focuses of a topic in each language. We can find differences of focuses of a topic among languages by looking at co-occurred words. Examples of co-occurred words are described in Section 4.

2.3 System requirement

Following functions are needed for a cross-lingual concern analysis system.

(1) Automatic translation of keywords into other languages.

(2) Clustering and summarizing articles.

(3) Automatic translation of content of blog articles.

(1) is needed for realizing a cross-lingual concern analysis because the system is required to retrieve and analyze articles written in various languages. (2) is required for supporting users to find important clusters of articles easily. (3) is also needed for facilitating users to understand content of blog articles.

In this paper, we focused on the first function. We implemented this function in the prototype system. An overview of the system is described in the next section.

3 Cross-lingual concern analysis system

In this section, we describe a prototype system of the cross-lingual concern analysis. We describe (1) an overview of the system, (2) summary of blog data, and (3) an automatic translation of keywords using Wikipedia.

3.1 Overview

The prototype system called KANSHIN collects blog articles written in Chinese, Japanese, Korean, and English. The system provides users with functions for retrieving and analyzing articles.

Figure 1 shows an overview of the system. The system has lists of blog sites for each language. By using these lists, the system collects RSS[13] and Atom feed files provided by blog sites, and extracts keywords from feed files by using morphological analysis tools, and store keywords and articles in databases. We describe current state of the blog data in Section 3.2.

The system uses several linguistic tools for extracting and indexing keywords from blog articles for each language. For Japanese, we used a morphological analysis tool called *Juman*[14]. For Korean, we used a morphological analysis tool called *KLT*[15]. For Chinese, we use a morphological analysis tool called *ICTCLAS* (Zhang et al. (2003)). There are two types of Chinese characters: (1) traditional Chinese, which is mainly used in Taiwan and Hong Kong, and (2) simplified Chinese, which is mainly used in mainland China. We treat both of types in the system. For English, we use a part-of-speech tagger called *SS tagger* (Tsuruoka and Tsujii (2005)).

Users can retrieve and analyze blog articles for each language (monolingual analysis), and across languages (cross-lingual analysis).

With respect to a monolingual concern analysis, users can retrieve blog articles, find co-occurred words with a keyword[16], and find daily and monthly topics for each language. We implemented these functions into the previous system (Fukuhara et al. (2005)).

With respect to a cross-lingual concern analysis, users can retrieve articles across CJKE languages. The system retrieves articles across languages by translating keywords into other languages. The system provides a *comparative trend graph* that shows a daily trend of articles containing the keywords. Figure 2 shows a screen image of a result of cross-lingual concern analysis. The system provides a user with graphs on (1) articles by language, (2) daily trend of articles, (3) comparison of number of articles, and (4) a list of retrieved articles. We describe the translation procedure in Section 3.3.

[13] Several references such as RDF Site Summary or Really Simple Syndication or Rich Site Summary are existed.
[14] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html (in Japanese)
[15] http://nlp.kookmin.ac.kr/HAM/kor/ (in Korean)
[16] For calculating co-occurred words, we use the Dice coefficient (Manning and Schütze, 1999, p.299).

Figure 2: Screen image of the prototype system.

Table 1: Summary of blog data (at March 10, 2007, 0:00)

Language	# of blog sites	# of articles	Days
Chinese	633,678	6,970,596	775
Japanese	2,648,008	139,712,963	1,088
Korean	471,869	28,081,261	587
English	70,167	6,081,399	121
Total	3,823,722	180,846,219	—

3.2 Data

Table 1 shows the summary of blog data stored in the system[17]. 2.6 million sites and 139 million articles are registered for Japanese since March 18th, 2004. 633 thousand blog sites and 7.0 million articles are registered for Chinese since January 25th, 2005. For Korean blog, 471 thousand blog sites, and 28 million articles are registered since August 1st, 2005. For English, 70 thousand blog sites, and 6.0 million articles are registered since November 11th, 2006. As a whole, 3.8 million sites, and 180 million articles are registered in the system.

3.3 Translating keywords using Wikipedia

For translating keywords given by a user, we use *Wikipedia*[18] which is a public online encyclopedia. Because Wikipedia entries are added and modified by many people, new words that are not registered in traditional dictionaries are registered quickly.

Figure 3 shows an overview of the translation procedure. A Wikipedia entry often has hyperlinks to the same entries written in other languages. Therefore, we follow those hyperlinks from an entry written in the source language to the target language. If there are no entries associated with the keywords, or if there are no hyperlinks to the target language, the procedure fails.

[17] Checked at March 10, 2007.
[18] http://www.wikipedia.org/

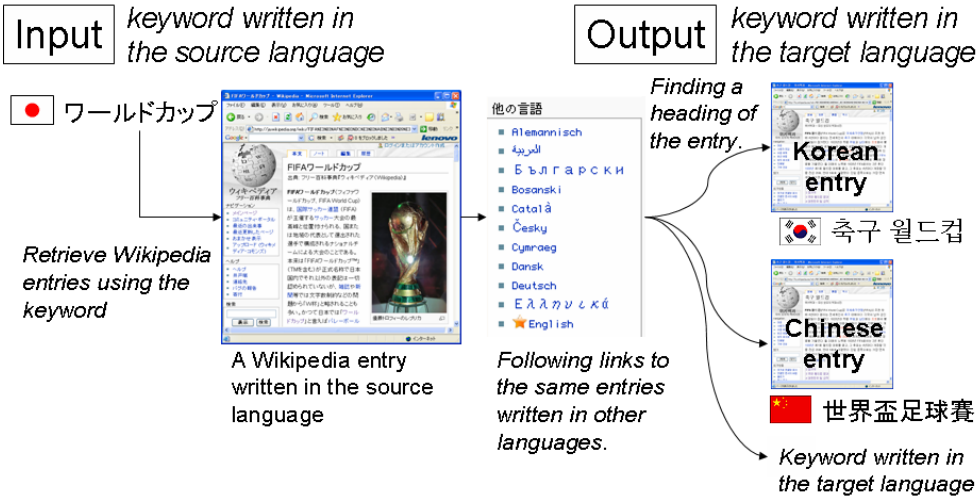

Figure 3: A procedure for finding translation of a keyword using Wikipedia.

4 Preliminary results

In this section, we describe preliminary results obtained by the system.

4.1 Differences of concerns over 'cloning' across languages

We can find differences of focuses on a topic by comparing co-occurred words with the same keyword across languages. Co-occurred words are important for identifying focuses of a topic. For example, if we find that 'hostage' and 'release' are co-occurred words with 'Iraq', we can find the focus is on the *hostage crisis in Iraq* in 2004. Furthermore, if we find 'abuse' as a co-occurred word with 'Iraq', we can find that focus is on the *Iraqi prisoner abuse scandal* occurred in 2004[19]. Therefore, we can find focuses of a topic from co-occurred words.

We show co-occurred words with 'cloning' found in CJKE blog articles from Table 2 to Table 5. The search term is from December 1st, 2006, through January 25th, 2007. Figure 4 compares the frequency of articles containing the keyword 'cloning' for each language. During this period, 5,009 articles are found. Among total articles, 2,873 articles (57.4%) are English, 1,771 articles (35.4%) are Japanese, 271 articles (5.4%) are Korean, and 94 articles (1.9%) are Chinese.

In Japanese blog, we found 'disease', 'technology', 'human', 'mobile (phone)', and 'body' as co-occurred words. Table 2 shows a list of co-occurred words. The first word 'disease' is appeared as 'Crohn's disease'. This is because the pronunciation of the keyword[20] used for retrieving articles resembles to the pronunciation of 'clone'. The fourth word 'mobile (phone)' is appeared as 'cloned mobile (phone)' in blog articles. This is because cellular phone cloning[21] came up in gossip in early December, 2006.

In Chinese, 'human', 'time', 'country', 'problem', 'mode', 'cell', 'period', and 'world' are extracted as co-occurred words in this period. Table 3 shows a list of co-occurred words. We found few interesting articles because only a few articles mentioned about 'cloning'. It seems that Chinese speaking bloggers have little concerns about 'cloning'.

In Korean, 'Kang Won Rae', 'attack', 'song', 'star wars', and 'phrase and a clause' are found as co-occurred words[22]. Table 4 shows a list of co-occurred words. The first word 'Kang Won Rae' is the name of a member of the CLON, which is one of popular music groups in South Korea. 'Kang Won Rae' was a topic among Korean bloggers because he met a car accident during this

[19] The change of focuses for a topic 'Iraq' is described in Fukuhara et al. (2005).
[20] We used 'クローン' for representing 'clone'.
[21] http://en.wikipedia.org/wiki/Phone_cloning
[22] We used '클론' for a query string that is a transliteration of 'clone' in Korean.

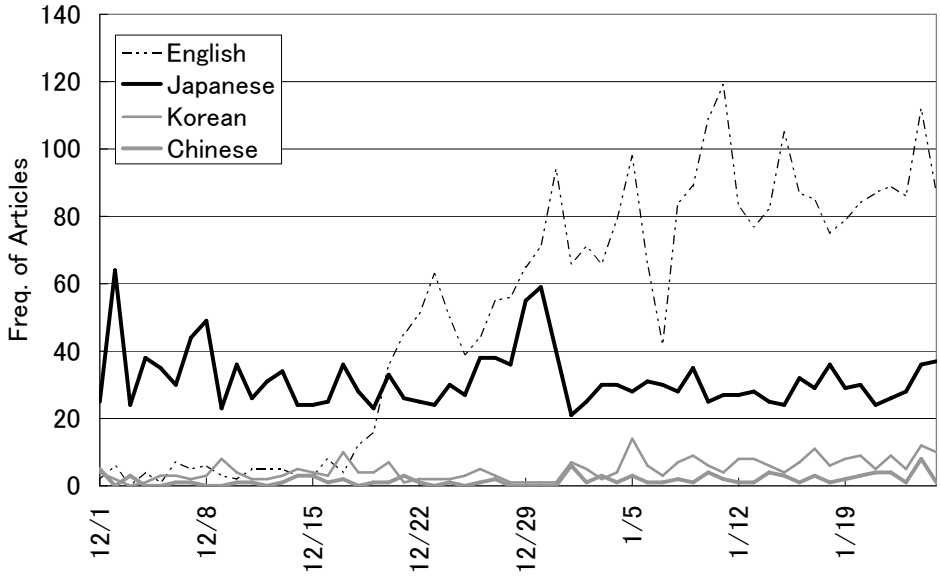

Figure 4: Comparison of concerns about 'cloning' appeared in CJKE blog articles (from December 1st, 2006, through January 25th, 2007).

Table 2: Co-occurred words with 'cloning' in Japanese blog.

Term	# of articles
Disease (病)	222
Technology (技術)	149
Human (人間)	146
Body (体)	78
Mobile (携帯)	56

period. Because his name was a topic in this period, and pronunciations of 'CLON' and 'clone' resemble each other, his name was extracted as a co-occurred word in this period.

In English blog, 'cons', 'pros', 'animal', 'information', 'stem' are found as co-occurred words. Table 5 shows a list of co-occurred words. Although there are many English articles, we found that most of these articles are splogs, which means spam blogs. This is a characteristic phenomenon for English blogs; we hardly paid attention to splogs in other languages. Consequently, 'cons' and 'pros' are strongly affected by splog articles. As described in Kolari et al. (2006), filtering out splog articles is important in English blogosphere. To filter out splogs is our future work.

5 Discussion

In this section, we discuss about the (1) improvement of precisions of a cross-lingual concern analysis, and (2) related work.

Table 3: Co-occurred words with 'cloning' in Chinese blog.

Term	# of articles
Human (人)	21
Time (时间)	7
Country (国)	6
Problem (问题)	5
Mode (模式)	4
Cell (细胞)	4
Period (时候)	4
World (世界)	4

Table 4: Co-occurred words with 'cloning' in Korean blog.

Term	# of articles
Kang Won Rae (강원래)	27
Attack (습격)	20
Song (노래)	14
Star Wars (스타워즈)	13
Phrase and a clause (구절)	11

5.1 Improvement of precisions of a cross-lingual concern analysis

As described in Section 4, the results contained errors such as 'Crohn's disease' and 'cloned mobile (phone)' although we anticipated that results contained the topic of cloning of animals. For improving a precision of the cross-lingual concern analysis, following issues should be solved.

1. Identifying contexts of a keyword
2. Finding adequate translations of a keyword
3. Cleaning of data

The first is to identify and limit contexts of a keyword. From preliminary results, we found several focuses with respect to 'cloning'. These failures were made because we did not know contexts of cloning beforehand, and we did not choose a specific context of cloning. For solving this issue, the system should clarify the context of a keyword by asking a user to specify one of possible contexts. For realizing this function, the system should classify articles retrieved by using a general keyword such as 'cloning' when the system accepts a keyword from users.

The second is to find adequate translations of a keyword. This issue is related to the first one, i.e., we can get more precise translations if we can clarify and specify the context of a keyword. Therefore clarifying contexts of a keyword is important.

In addition to clarification of a context, we should consider the coverage of words for translation. As described in Section 3.3, we use Wikipedia as a cross-lingual dictionary. Although Wikipedia has lots of entries, we often fail to translate a keyword into other languages because there are no entries in the target languages. For compensating this failure, we should adopt other digital resources such as online dictionaries and thesauri. On the other hand, it is not easy to find those resources in some minor languages. Although online dictionaries and thesauri would be prepared in the near future, we should tackle this problem when we try to treat minor languages.

The third issue is the cleaning of data. As shown in Table 5, the effects of splogs became significant in English blog. Splogs are appearing not only in English blog but also in Japanese and Korean blogs recently. Therefore automated identification and removal of splogs are necessary for the system. We are tackling with splog filtering by using word and article frequency. For example, we can find abnormal posting behavior by observing history of postings by a blog site. We can

Table 5: Co-occurred words with 'cloning' in English blog.

Term	# of articles
Cons	1959
Pros	1850
Animal	214
Information	209
Stem	188

also find abnormal words that are heavily appeared in splog articles such as 'pros' and 'cons' in Table 5. These ideas can be applied not only to English but also to other languages. We will develop a method for removing splogs, and implement the method in the system.

5.2 Related work

With respect to cross-lingual concern analysis, Google provides linguistic tools such as *Google Trends*[23] that analyzes trends of keywords fed to Google search, and *Google News*[24] that automatically collects and displays latest news stories written in various languages. In Google Trends, we can compare frequency of documents containing keyword strings specified by a user across languages. However, Google Trends does not translate keyword strings into other languages.

Although Google Trends and Google News are monolingual services, we proposed a cross-lingual concern analysis service.

6 Conclusion

In this paper, we proposed a cross-lingual concern analysis system that collects and analyzes multilingual blog articles. Our prototype system collects and analyzes Chinese, Japanese, Korean, and English blog articles. Users can find differences of focus on a topic across languages because the system automatically translates keywords into other languages by using Wikipedia. We found differences of concerns about 'cloning' among CJKE blog articles. Our future work contains (1) to incorporate more languages into the system, (2) to filter out English splog articles, and (3) to improve and evaluate translation procedure.

References

Fukuhara, T., Murayama, T., and Nishida, T. (2005). Analyzing concerns of people using weblog articles and real world temporal data. In *Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (Available at http://www.blogpulse.com/papers/2005/fukuhara.pdf, Accessed 2007-03-13).

Glance, N., Hurst, M., and Tomokiyo, T. (2004). Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (Available at http://www.blogpulse.com/www2004-workshop.html, Accessed 2007-03-13).

Ishida, T. (2006). An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pages 96–100. (Available at http://langrid.nict.go.jp/publicatione.htm, Accessed 2007-05-28).

Kolari, P., Finin, T., and Joshi, A. (2006). SVMs for the Blogosphere: Blog identification and Splog detection. In *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 92–99. AAAI Press. (Available at http://ebiquity.umbc.edu/get/a/publication/213.pdf, Accessed 2007-03-13).

[23] http://www.google.com/trends
[24] http://news.google.com/

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Nanno, T., Fujiki, T., Suzuki, Y., and Okumura, M. (2004). Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 320–321, New York, NY, USA. ACM Press.

Tsuruoka, Y. and Tsujii, J. (2005). Bidirectional inference with the easiest-first strategy for tagging sequence data. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 467–474, Morristown, NJ, USA. Association for Computational Linguistics.

Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., and Liu, Q. (2003). Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 184–187, Morristown, NJ, USA. Association for Computational Linguistics.