KANSHIN: A Cross-lingual Concern Analysis System using Multilingual Blog Articles

Tomohiro Fukuhara
RACE, The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa, Chiba JAPAN

Akifumi Kimura
Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Tokyo JAPAN

Yoshiaki Arai
Tokyo Denki University
2-2 Kandanishiki-cho, Tokyo JAPAN

Takayuki Yoshinaka
Tokyo Denki University
2-2 Kandanishiki-cho, Tokyo JAPAN

Hidetaka Masuda
Tokyo Denki University
2-2 Kandanishiki-cho, Tokyo JAPAN

Takehito Utsuro
University of Tsukuba
1-1-1, Tennodai, Tsukuba JAPAN

Hiroshi Nakagawa
ITC, The University of Tokyo
7-3-1 Hongo, Tokyo JAPAN

Abstract

An architecture of cross-lingual concern analysis (CLCA) using multilingual blog articles, and its prototype system are described. As various people who are living in various countries use the Web, cross-lingual information retrieval (CLIR) plays an important role in the next generation search. In this paper, we propose a CLCA as one of CLIR applications for facilitating users to find concerns of people across languages. We propose a layer architecture of CLCA, and its prototype system called KANSHIN. *The system collects Japanese, Chinese, Korean, and English blog articles, and analyzes concerns across languages. Users can find concerns from several viewpoints such as temporal, geographical, and a network of blog sites. The system also facilitates users to browse multilingual keywords using Wikipedia, and the system facilitates users to find splog features. An overview of CLCA architecture and the system are described.*

1 Introduction

Today many people who are living in the world use the Internet. We can communicate with others via e-mail. We can read news articles on the Web. We can post our thoughts and opinions in Weblogs, social network services (SNS), and so on. These textual information exchanged on the Web is written in various languages. According to the language observatory project[10], 6,000 languages are currently spoken in the world. Although machine-readable languages are limited at this moment, so many information is exchanged in various languages on the Web. We call this phenomenon as *language-explosion*.

We can see the language-explosion phenomenon in (1) Wikipedia, and (2) the blogosphere. Wikipedia, which is a Web-based multilingual free encyclopedia[1], covers 253 languages for describing articles[2]. Another example is the blogosphere. According to the Technorati's report[3] that analyzes the state of blogosphere in April, 2007, the blogosphere is separated into various language spaces such as Japanese(37%), English(36%), Chinese(8%), Italian(3%), Spanish(3%), Russian(2%), French(2%) and so on. Although many spam blogs called *splogs* are contained, the blogosphere can be seen a multilingual space.

We propose a *cross-lingual concern analysis (CLCA)* for facilitating people to find various viewpoints of people in the world. As language-explosion phenomenon proceeds, *multilingual information access (MLIA)* plays an important role on the Web[5]. Oard pointed out the importance of multilingual information analysis[11]. With CLCA, one can find various viewpoints by comparing information written in various languages. The goal of CLCA is to facilitate

[1] http://en.wikipedia.org/wiki/Wikipedia:About
[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias
[3] http://www.sifry.com/alerts/archives/000493.html

people to understand concerns of people across languages by providing hot topics in each language space, clustering and summarizing topics, translating foreign languages into user's mother tongue (see section 2.2).

In this paper, we propose a layer architecture of CLCA, and its prototype system called KANSHIN. The system collects multilingual blog articles, and analyzes concerns of people in each language. The system provides a user several viewpoints such as temporal, focal, geographical, and network viewpoints. The system also provides a cross-lingual keyword navigation tool based on interlanguage links of Wikipedia, and a splog survey tool called SplogExplorer for investigating the splogosphere. Although these tools have not been integrated into a single user interface yet, we are planning to integrate these tools into a single interface.

This paper consists of following sections. Section 2 describes reviews of previous work, and propose a layer architecture of the CLCA. Section 3 describes a prototype system of the CLCA. In section 4, we describe related work. In section 5, we summarize arguments, and describe future work.

2 Cross-lingual concern analysis

In this section, we review previous work (in section 2.1), and propose a layer architecture of the CLCA (in section 2.2).

2.1 Previous work

As previous work, we review (1) the Columbia NewsBlaster, and (2) the language grid project.

The *Columbia NewsBlaster*[4] is one of related work. The system collects news articles written in several languages, and translates multilingual articles into English, and classifies and summarizes the translated articles[2]. The focus of this work is on NLP techniques of clustering, summarization of multilingual texts. Meanwhile, our focus is on the combination of various representations, i.e., we use textual information, and graphical representations such as geographical and network views (see section 3).

In the *language grid project*, Ishida and his colleagues proposed tools to support cross-lingual and cross-cultural communication[6]. The focus of the language grid project is on the human-to-human communication in the real-world and computer-mediated communication (CMC). Meanwhile, our focus is on an assist of users to find and compare concerns of people across languages. We consider that the combination of several represetations is important for CLCA. In the next subsection, we describe a layer architecture of CLCA.

[4]http://newsblaster.cs.columbia.edu/

2.2 Layer architecture of CLCA

For facilitating people to find concerns across languages, we propose a layer architecture of CLCA[5]. Figure 1 shows the layer architecture of CLCA. This architecture consists of the following seven layers.

1. Fundamental layer
2. Linguistic resource layer
3. Search layer
4. NLP application layer
5. Machine translation layer
6. Integration layer
7. Trust layer

The first layer is the *fundamental layer* in which character codes and communication components are provided. At the bottom line of Figure 1, there are two components: (1) Unicode, and (2) Internet components. These are fundamental components for CLCA for describing texts in various languages, and exchanging textual information with other people.

The second layer is the *linguistic resource layer*. There are two components, i.e., (1) linguistic resources such as thesauri, dictionaries, and NLP tools, and (2) multilingual textual data resources such as news articles and blog pages written in various languages.

The third layer is the *search layer* in which there is a cross-lingual information retrieval (CLIR) component. In this layer, one can retrieve multilingual textual data on the Web.

The fourth layer is the *NLP application layer* in which various NLP applications such as text summarization, clustering, opinion analysis, sentiment analysis, and visualization are provided. These NLP applications are applied to each language.

The fifth layer is the *machine translation (MT) layer*. This component translates the output of NLP applications into user's mother tongue.

The sixth layer is the *integration layer* in which various outputs of NLP applications, which are translated into a user's mother tongue, are integrated into a report. Intelligent user interface would be needed for this layer.

The seventh layer is the *trust layer* where trust among people will be established. In this layer, one can find concerns across languages, and s/he can have a better conversation with other people who are living in different areas or countries.

[5]The idea of this architecture is borrowed from the layer cake architecture of the Semantic Web (http://www.w3.org/2001/sw/).

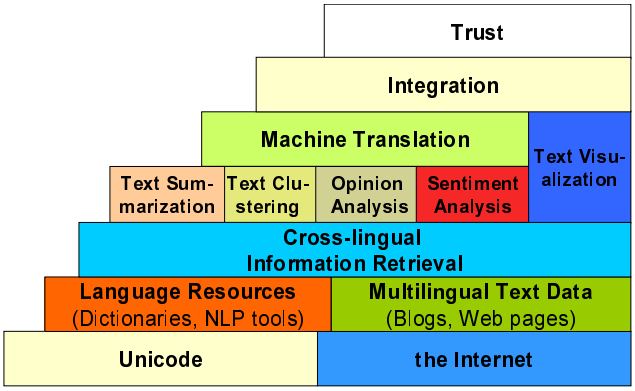

Figure 1. Layer architecture of the cross-lingual concern analysis.

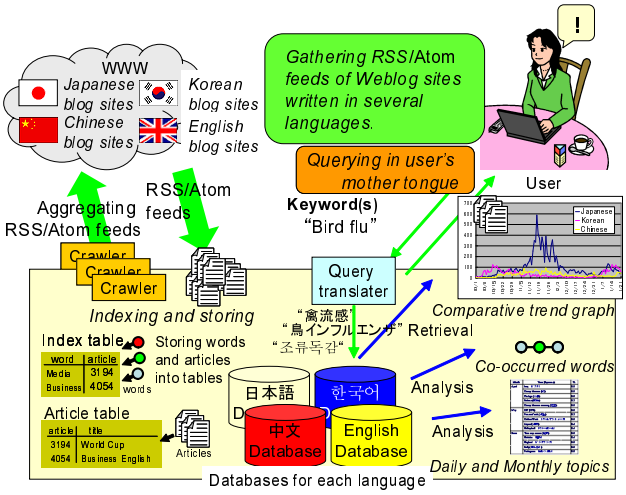

Figure 2. Overview of the KANSHIN system. The system collects Japanese, Chinese, Korean, and English blog articles. A user can retrieve articles across languages. She can also find co-occurred words with a keyword, and topic-words of the day.

Table 1. Summary of the collected data

Language	# of blog sites	# of articles
Chinese	895,128	9,419,747
Japanese	3,994,473	210,532,827
Korean	106,389	7,609,625
English	110,716	16,422,286
Total	5,106,706	243,984,485

3 Prototype system of CLCA

In this section, we describe (1) an overview of the system, (2) a cross-lingual keyword navigation tool, and (3) a splog survey tool called *SplogExplorer*.

3.1 Overview of the system

We created a prototype system of CLCA called KANSHIN. The system collects and analyzes multilingual blog articles[3, 4]. Figure 2 shows an architecture of the system. The system collects and analyzes Japanese, Chinese, Korean, and English blog articles. Table 1 shows the summary of collected data at this moment[6].

The system provides several viewpoints for finding concerns across languages. Users can find following viewpoints.

1. Temporal view of concerns
2. Focal view of a topic
3. Geographical view
4. Network view

We describe each viewpoint in the following.

[6] on March 9, 2008, 18:00 JST

Temporal view of concerns

Users can find temporal trends of concerns across languages. Figure 3 shows daily trends of concerns over 'crude' in the four languages (from October 6, 2007 to January 25, 2008). In this figure, two points are indicated. At the point A, we can see a burst of Korean articles around December 10, 2007. The reason of this burst is an oil spill accident occurred in Korea[7]. Meanwhile, at the point B, we can see another burst of Japanese blog articles around January 5, 2008. This is because crude oil prices past $100 a barrel in the New York mercantile exchange. Thus, we can find differences of concerns over the same topic. We can also compare concerns by using co-occurred words (see next subsection).

Users can retrieve blog articles across languages by using his or her mother tongue. Because the system automatically translates a keyword into other three languages[4], users can retrieve articles without translating the keyword manually. The system translates a keyword by using bilingual dictionaries and Wikipedia. Furthermore, a multilingual keyword navigation tool based on the interlanguage link of Wikipedia is provided for supporting users to select an appropriate translation from keyword candidates (see section 3.2).

[7] http://en.wikipedia.org/wiki/2007_Korea_oil_spill

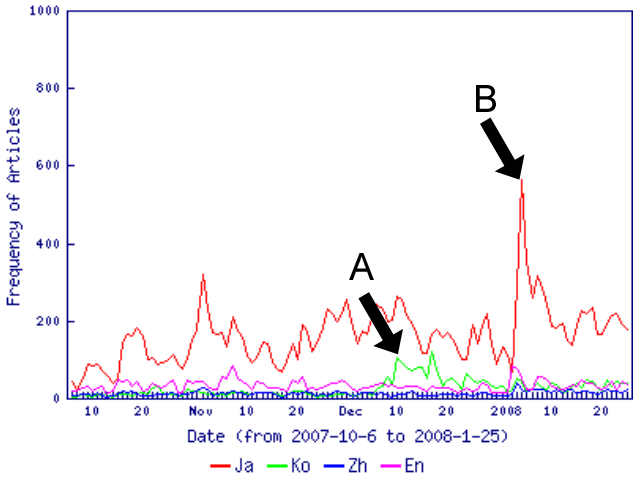

Figure 3. Daily trends of concerns over 'crude' in four languages ('Ja' stands for Japanese, 'Ko' for Korean, 'Zh' for Chinese, and 'En' for English blogs). Point A of Korean blog shows concerns over an oil spill accident in Korea. Point B of Japanese blog shows concerns over news stories about the rise of crude oil prices in New York.

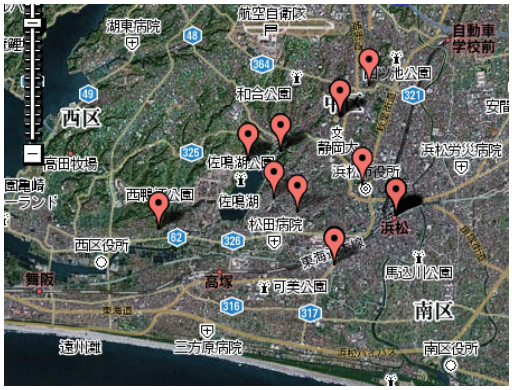

Figure 4. Geographical analysis of blog articles. In this figure, each point represents a location mentioned in blog articles. Points are plotted on the Google map.

Focal view of a topic

Users can find co-occurred words for a keyword. Table 2 shows a list of co-occurred with 'crude' from December 30, 2007 through January 27, 2008. We can find that the term 'crude' is co-occurred with 'oil', 'price', and 'market' in Japanese, Chinese, and English blogospheres. Meanwhile, in Korean blogosphere, 'Taean (태안)', 'Accident (사고)', 'Chung-Nam (충남)', 'Offshore (앞바다)' are extracted as co-occurred words with 'crude' because an oil spill occurred near the Taean county of South Korea in December 2007. Thus, we can see differences of focuses on a topic among languages.

Geographical view

Geographical views for visualizing locations of concerns is also important for CLCA. The system provides a user geographical views[8]. Figure 4 shows a geographical analysis of a blog site. One can find locations mentioned in blog articles instantly. For extracting location information, the system extracts address information by using a named-entity extraction tool called CaboCha[8], and Yahoo! LocalSearch API[9]. Extracted address information is plotted on the Google map[10].

Network view

Network analysis allows us to find implicit structures of blog sites across languages. Therefore, the system provides a network view of relations among blog sites. Figure 5 shows a network view (link structure) of Korean blog sites. One can find a network of blog sites by specifying a seed blog site. From this seed blog site, the system crawls blog sites based on the depth-first search (up to three hops), and extracts hyperlinks from blog articles. We can find similar blog sites that share the same concern.

3.2 Cross-lingual keyword navigator

The cross-lingual keyword navigator[11] is a support tool for browsing translation candidates of a keyword across languages. Figure 6 shows a screen image of the tool. This tool analyzes *interlanguage links (ILLs)* of Wikipedia[12], and shows relations of keywords as a network. Wikipedia has an interlanguage link that enables editors to link two articles written in language A and B together. By analyzing ILLs, we can find translation candidates. Figure 6 shows a network on 'video games'. Users can find several translation candidates for the keyword, and choose one of keywords for retrieval.

[8] This function currently works with Japanese language.

[9] http://developer.yahoo.co.jp/map/localsearch/V1/localsearch.html (in Japanese)

[10] http://maps.google.com/

[11] Available at http://arai.cdl.im.dendai.ac.jp/

[12] http://en.wikipedia.org/wiki/Help:Interlanguage_links

Table 2. Co-occurred words with 'crude' in each language (The period is from December 30th, 2007 through 26th January, 2008).

	Japanese blog		Chinese blog		Korean blog		English blog	
	Term (Japanese)	Articles	Term (Chinese)	Articles	Term (Korean)	Articles	Term	Articles
1	High (高)	746	Oil (油)	132	Taean (태안)	181	Prices	201
2	Price (価格)	678	Full text (全文)	90	Accident (사고)	144	Barrel	187
3	Inflating (高騰)	677	Price (价格)	75	Chung-Nam (충남)	117	Price	127
4	Effects (影響)	237	Country (国)	64	Offshore (앞바다)	85	New	89
5	Future (先物)	235	Future (期货)	58	Damage (피해)	81	Futures	88
6	Market (市場)	234	International (国际)	55	Taean County (태안군)	59	US	81
7	Up (上昇)	233	Market (市场)	51	Incident (발생)	58	Year	64

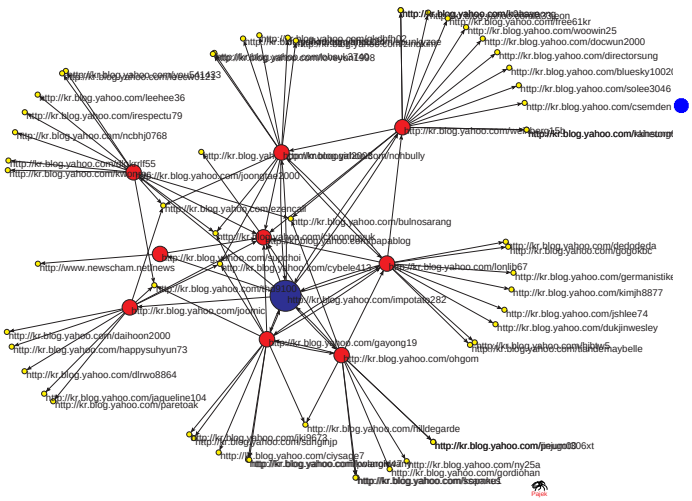

Figure 5. A network view of relations between blog sites (the seed is a Korean blog site). The system follows hyperlinks until three hops. The largest node (blue) represents a seed url specified by a user. The second largest node (red) represents blog sites connected to the seed.

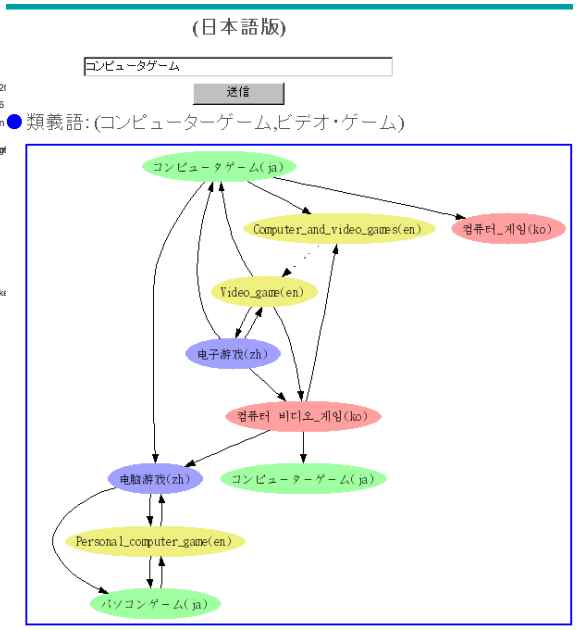

Figure 6. Screen image of the cross-lingual keyword navigator. This tool analyzes interlanguage links of Wikipedias, and shows relations between keywords.

3.3 SplogExplorer

One of major problems in the blogosphere is spam blogs (splogs). Because there are so many splogs in the blogosphere[7], filtering out splogs is important for every language spaces. As a first step towards splog filtering, we developed a splog analysis tool called *SplogExplorer* for investigating the splogosphere.

SplogExplorer has following functions: (1) checking and detecting splogs function, (2) finding *copy-and-paste* splogs function, and (3) collaborative annotation function. Currently SplogExplorer is provided only for researchers. We are planning to simplify this tool so that ordinary users can find and filter out splogs.

Figure 7 shows the screen image of the SplogExplorer. Users can find splog candidates by browsing parameters such as number of articles, number of links in an article, and so on. Table 3 shows a list of parameters that SplogExplorer analyzes. We chose language-independent features so that we can apply this tool to other languages. Although the current dataset used in SplogExplorer is Japanese blog

Table 3. List of feature variables of splogs

of entries per day
of characters in an entry
of HTML tags in an entry
of within-site links of a blog site
of external links of a blog site

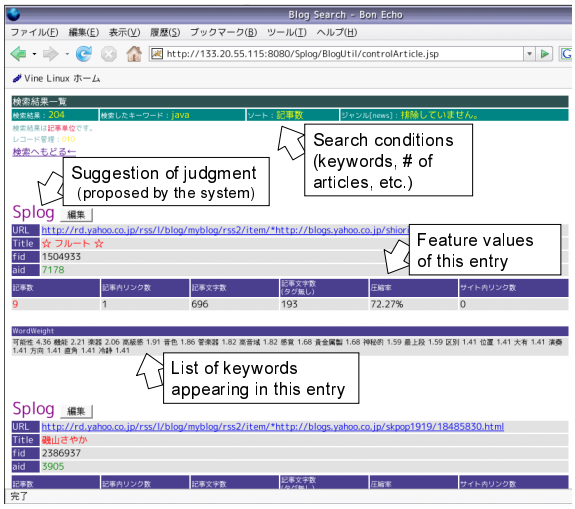

Figure 7. Screen image of the SplogExplorer. Users can find splogs and their feature values.

articles, we will use other languages, and compare features of splogs across languages.

4 Discussion

In this section, we describe related work.

4.1 Related work

Bautin et al proposed an international sentiment analysis using multilingual news and blog articles[1]. Their approach is based on a machine translation (MT), i.e., their system translates multilingual blog articles into English, and then applies sentiment analysis to the translated articles. On the other hand, our approach aims to apply NLP applications to blog articles written in each language firstly (see section 2.2), and then apply MT to the results of NLP applications. This is because we do not want to lose the quality of contents. Although we have not implemented MT in our system yet, we will evaluate the quality of final results.

In the Global Autonomous Language Exploitation (GALE) project of DARPA[13], researchers are creating systems that assist military persons to understand multilingual speech and text data. The direction of the GALE project is similar to this research, but the focus is different. Our research focus is on analytic aspect in which users can compare concerns across languages.

Google news[14] is one of related work. Google news collects news articles that are written in various languages on the Web, and classifies and summarizes articles. The concept of Google news is different from CLCA because Google news does not compare concerns of people across languages directly. In this research, we proposed a CLCA system that compares concerns of people across languages.

TextMap[15] and Google Trends[16] are also related systems. TextMap shows a geographical viewpoint of concerns[9], and Google Trends shows a temporal viewpoint of concerns. These systems are monolingual systems. We propose a cross-lingual concern analysis using multilingual blog articles.

5 Conclusion

In this paper, we proposed a layer architecture of the cross-lingual concern analysis (CLCA), and its prototype system. The more multilingual texts are available on the Web, the more CLCA becomes important. One of important issues in the next generation search is a CLCA. The prototype system facilitates users to find (1) temporal trends of concerns across languages, (2) focuses of a topic based on co-occurred words with a keyword, (3) geographical and network viewpoints of blog sites/articles. The system also facilitates users to find multilingual keyword candidates, and provides a tool for exploring the splogosphere.

Our future work contains (1) to incorporate machine translation systems for supporting users to read multilingual articles, and (2) to integrate components described in this paper into a single user interface, and (3) to evalute the integrated system.

References

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'08)*, pages 19–26, 2008.

[2] D. K. Evans, J. L. Klavans, and K. R. McKeown. Columbia newsblaster: Multilingual news summarization on the web. In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL*

[13] http://www.arpa.mil/ipto/programs/gale/gale.asp
[14] http://news.google.com/
[15] http://www.textmap.com/
[16] http://www.google.com/trends

2004: Demonstration Papers, pages 1–4, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[3] T. Fukuhara, T. Murayama, and T. Nishida. 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*, 2005. (Available at http://www.blogpulse.com/papers/2005/fukuhara.pdf, Accessed 2008-01-26).

[4] T. Fukuhara, T. Utsuro, and H. Nakagawa. Cross-lingual concern analysis from multilingual weblog articles. In A. Nijholt, O. Stock, and T. Nishida, editors, *Proceedings of the 6th Workshop on Social Intelligence Design*, pages 55–64, 2007. (ISSN: 1574-0846).

[5] F. C. Gey, N. Kando, C.-Y. Lin, and C. Peters. New directions in multilingual information access. *SIGIR Forum*, 40(2):31–39, 2006.

[6] T. Ishida. An infrastructure for intercultural collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pages 96–100, 2006. (Available at http://langrid.nict.go.jp/en/publication.html, Accessed 2008-01-26).

[7] P. Kolari, A. Java, and T. Finin. Characterizing the Splogosphere. In *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wid Web Conference*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County, May 2006.

[8] T. Kudo and Y. Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69, 2002.

[9] L. Lloyd, D. Kechagias, and S. Skiena. *Lydia: A System for Large-Scale News Analysis*, volume 377 of *Lecture Notes in Computer Science*, pages 161–166. 2005.

[10] Y. Mikami, P. Zavarsky, M. Z. A. Rozan, I. Suzuki, M. Takahashi, T. Maki, I. N. Ayob, P. Boldi, M. Santini, and S. Vigna. The language observatory project (lop). In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 990–991, New York, NY, USA, 2005. ACM.

[11] D. W. Oard. Towards analysis tools for a multilingual blogsphere. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 176–178. AAAI Press, 2006. AAAI Technical Report SS-06-03.