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 Nishiki-cho, Kanda, Tokyo JAPAN

Takayuki Yoshinaka
Tokyo Denki University
2-2 Nishiki-cho, Kanda, Tokyo JAPAN

Hidetaka Masuda
Tokyo Denki University
2-2 Nishiki-cho, Kanda, 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) from multilingual blog articles, and its prototype system are described. As various people who are living in various regions and countries join the Web, cross-lingual information retrieval (CLIR) becomes important. In this paper, we propose a CLCA as one of CLIR applications for understanding 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, spatial (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 is described.

1 Introduction

Today many people who are living in the world[10]. Although machine-readable languages are limited at this moment, so many information is exchanged by various languages on the Web. We call this phenomenon as language-explosion.

We can see the language-explosion phenomenon on the Web. In Wikipedia, which is a Web-based multilingual free encyclopedia, 253 languages are used for describing articles. Another example is the blogosphere. According to the Technorati’s report that analyzes the state of blogosphere in April, 2007, the blogosphere is separated into various language spaces such as Japanese, English, Chinese, Italian, Spanish, Russian, and French. Although many spam blogs are contained, the blogosphere can be seen a multilingual space.

As language-explosion phenomenon proceeds, multilingual information access (MLIA) would play an important role on the Web. Oard also argued the importance of multilingual blog analysis. Although various MLIA applications are proposed, we propose a cross-lingual concern analysis (CLCA) as one of MLIA applications. With CLCA, one can find various viewpoints by comparing information written in several languages. The goal of CLCA is to facilitate 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 2.2).

In this paper, we propose a layer architecture of CLCA, and its prototype system called KANSHIN. The system collects and analyzes multilingual blog articles, and analyzes articles from temporal, focal, geographical (spatial), and network viewpoints. Users can find various viewpoints on a topic across languages. The system provides a cross-lingual keyword navigation tool based on interlanguage links of Wikipedia, and a splog survey tool. By using the proposed system, users can find and compare concerns of people across languages.

This paper consists of following sections. Section 2 describes reviews of previous work, and propose a layer architecture of CLCA. In section 3, we describe a prototype system of 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, and propose a layer architecture of CLCA.

2.1 Previous work

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

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

In the language grid project, Ishida and his colleagues propose tools to support cross-lingual and cross-cultural communication\(^6\). The focus of this project is on the human-to-human communication in the real-world communication and computer-mediated communication (CMC) using video and audio conferencing tools. Meanwhile, our focus is on analytical aspect of concerns across languages.

We consider that the mixture of several representations of concerns is important for CLCA. In the next subsection, we describe a layer architecture of CLCA.

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

We describe each layer.

The first layer is the fundamental layer in which character code 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 texts with other people and other computers.

The second layer is the linguistic resource layer. There are two components in this layer; (1) linguistic resources, and (2) multilingual textual data resources. The linguistic resources contain thesaurus, dictionaries, and NLP tools. Multilingual textual data contain textual data on the Web.

The third layer is the search layer in which there is a cross-lingual information retrieval (CLIR) component. In this layer, people 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 work with specific language.

The fifth layer is the machine translation (MT) layer. This component translates the output of NLP applications in the fourth layer.

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 would 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.

\(^4\)http://newsblaster.cs.columbia.edu/

\(^5\)The idea of this architecture is borrowed from the layer cake architecture of the Semantic Web (http://www.w3.org/2001/sw/).
3 Prototype system

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

3.1 Overview

KANSHIN is a system that collects and analyzes multilingual blog articles[3, 4]. Figure 2 shows an architecture of the system. The system collects and analyzes four languages, i.e., Japanese, Chinese, Korean, and English blog articles. Table 1 shows a summary of the collected blog data at this moment.

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

In the following, we describe each viewpoint.

Table 1. Summary of collected data (at March 9, 2008, 18:00 JST)

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

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 Korea6. 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 on 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 translates a keyword into other three languages, 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 3.2).

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). The point A in Korean blogs shows concerns over an oil spill accident in Korea, and the point B in Japanese blogs shows concerns over news stories about the rise of crude oil prices in New York.

Focal view of a topic

Users can find co-occurred words for a keyword. Table 2 shows a list of co-occurred with ‘crude’ during December 30, 2007 to January 27, 2008. We can see that the term ‘crude’ is co-occurred with ‘oil’, ‘price’, and ‘market’ in Japanese, Chinese, and English blogosphere. Meanwhile, in Korean blogosphere, ‘Taean (태안)’, ‘Accident (사고)’, ‘Chung-Nam (충남)’, ‘Offshore (앞바다)’ are appeared 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

Providing geographical views is also important for CLCA. The system provides a geographical viewpoint of concerns. Figure 4 shows a geographical analysis result of a blog site. Although this function currently works in Japanese blogs, one can find locations that are 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[7].

Network view

Network analysis is also important for understanding concerns on the Web[1]. Therefore, the system provides a network view of blog sites. Users can find a link structure of blog sites. The system crawls blog sites based on the depth-first search (up to three hops), and extracts hyperlinks from blog articles. A starting URL (seed URL) is specified by a user. Figure 5 shows an analysis result of link structure between Korean blog sites. We can find similar blog sites are connected together in this figure.

3.2 Cross-lingual keyword navigator

The cross-lingual keyword navigator[9] is a support tool for browsing translation candidates of a keyword across languages. Figure 6 shows a network on ‘video games’. Users can find several translation candidates for the keyword, and choose one of keywords for retrieval.

The cross-lingual keyword navigator utilizes interlanguage links of Wikipedia[10]. Wikipedia has an interlanguage link feature that enables editors to link two articles written in language A and B together. By analyzing ILL, we can find translation candidates. By comparing bilingual dictionaries, Wikipedia also contains new words and technical terms. The tool analyzes ILL of Wikipedias, and shows re-
investigating the splogosphere.

3.3 SplogExplorer

One of big issues in the blogosphere is splogs. Because there are so many spam blogs called splogs in the blogosphere[7], filtering out splogs is important for every language spaces. As a first step towards splog filtering, we prepared 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. 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 articles,
Table 3. List of feature variables of splogs

<table>
<thead>
<tr>
<th>Feature Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td># of entries per day</td>
</tr>
<tr>
<td># of characters in an entry</td>
</tr>
<tr>
<td># of HTML tags in an entry</td>
</tr>
<tr>
<td># of within-site links of a blog site</td>
</tr>
<tr>
<td># of external links of a blog site</td>
</tr>
</tbody>
</table>

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

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

Google news\(^{11}\) 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.

In the Global Autonomous Language Exploitation (GALE) project of DARPA\(^{12}\), 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.

TextMap\(^{13}\) and Google Trends\(^{14}\) 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 described a layer architecture of the cross-lingual concern analysis (CLCA), and proposed 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 is to incorporate machine translation systems for supporting users to read multilingual articles.

References

\(^{11}\)http://news.google.com/
\(^{13}\)http://www.textmap.com/
\(^{14}\)http://www.google.com/trends


