Finding Chance in Multi-lingual Translation Site

Akinori Abe, Chen Kak Toong, Masashi Nakamura, Mitsuo Tuskada, and **Hiroshi Kotera**
NTT MSC
No. 43000, Jalan APEC, 63000 Cyberjaya
Selangor Darul Ehsan, Malaysia
ave@cslab.kecl.ntt.co.jp, toong@arc.net.my,
msc-naka@mbf.sphere.ne.jp, msc-tskd@mbd.sphere.ne.jp,
kotera@arc.net.my

Abstract

Recently, the data on web site has been thought of as meaningful data to analyze our society. It is used to see the trends in specific field. In addition, it can be used to predict the next trend or current hidden needs. This is because the data is natural (unintentional). We run the multi-lingual machine translation service site. In fact, in more than one year, we have collected huge mount of log data. They are user's access data, user's translation data (word, text, url, language pair etc.), and user's feedback comments. From these data, we can obtain a lot of information. First, this paper shows details of our multi-lingual translation site and some statistically analyzed results from the log data. Then, from the viewpoint of chance discovery, the analysis of the log data is shown.

Introduction

From the marketing viewpoint, it is important to analyze real (electronic) information or data like POS data to predict future trend, to control supply, or to produce a new service. For example, as for shops, if they collect and analyze their selling data, they can find a certain relationship between climate and goods to be sold. Then, they can use efficient selling strategies. These sorts of strategies have been mainly planed by human experts. However, recently, computers try to perform like the human experts by analyzing electronic POS data. One of the examples will be data mining from POS data. By data mining from such data, such relations as shown above and current trend will be estimated. In addition, recently, we have been able to deal with log data in the web site. The data in the web site seem to be treasure from the viewpoint of chance discovery. This is because the data is a result from the user's behaviour in the web site. In addition, since it reflects user's daily behaviour, the data is natural (unintentional). The results include, for example, online shopping data, internet searching procedure, user's interest transition etc. In addition, important data may hide there. In fact, some companies do internet audience research with the contracted person (Hagiwara, 2000). They use special device to analyze the behaviour of user in the internet. Their result shows current trends in the internet. In fact, the user frequently accesses 'search engine sites' and the sites that provide online community (in 2000). Anyway, though we do not deal with in this paper, we can also find novel or rare events from the results. From the computational viewpoint, from simple ones to graphical ones, there exist a lot of web log analyzers (either commercial or free). They are usually programmed and used for general purpose. Therefore, in general, they are used to find trends from web logs. On the other hands, Matsumura proposed *ChanceFinder* to find new (potential) topics from the user log file of Web site (Matsumura, 2000),(Matsumura, 2001), (Matsumura, 2002). By their definition, chance pages are the page that has novel topic and that has some links to such pages. Thus, it is very important to deal with web log file to predict human behaviour and social situation and affair.

We run the multi-lingual machine translation service site `http://sangenjaya.arc.net.my/`. In fact, in more than one year, we have collected huge mount of log data. They are user access data, user's translation data (word, text, url, language pair etc.), and user's feedback comments. From these data, we can obtain a lot of information. Of course, we must consider the user's severe comments to improve the service. However, we think that there hides more important information. During analyzing the log data, we will find a sort of chance (definition will be shown later).

This paper, first, shows details of our multi-lingual translation site and some statistically analyzed results from the log data. Then, from the viewpoint of chance discovery, another analysis of the log data is shown.

Multi-lingual translation site

Overview of multi-lingual translation site

Our multi-lingual translation site consists of online dictionary, text translation, and web translation. The translation language pairs include translation between English and Japanese, English and Malay, English and Indonesian, and English and Chinese etc.[1] Since the site is located in Malaysia where people speaks at least four different languages (Malay, Mandarin, English, Tamil (or Hindi) etc.), this site has been focused on the translation between English and Asian language. Part of the web page interface is shown in Fig. 1.

Table 1 shows access ratio in terms of user's countries. Since our site introduced in Japanese homepage magazine,

[1] Since the end of April 2002, translation between Japanese and Chinese, Japanese and Korean, and that from English to Thai have been added. Also, translation from English to Japanese restarted.

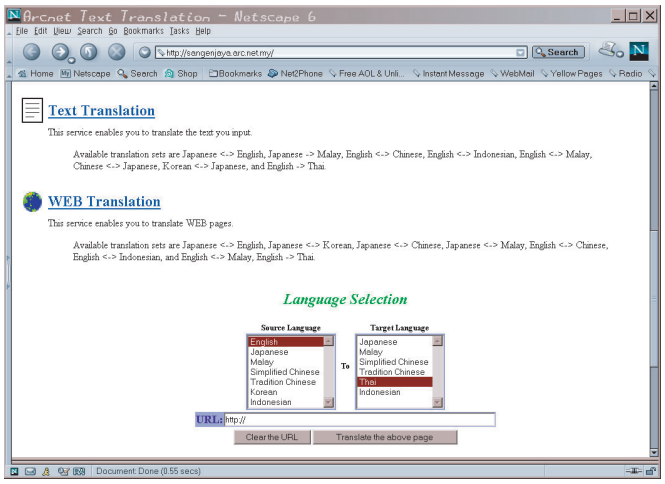

Figure 1: The translation page interface.

around half of access comes from Japanese site. In addition, due to the focusing on Asian language service, access ratio from Asian countries are more than 75%. Therefore, the following results might be influenced by this access ratio.

countries	ratio 10-01	ratio 7-02
.com, .org, .net	19.35	16.41%
Japan	64.03	61.72%
Malaysia	6.82	13.99%
Singapore	2.90	1.76%
Taiwan	1.54	1.33%
Hong Kong	0.73	0.86%
Other Asia	0.57	0.85%
Europe, Russia	1.06	1.18%
USA, Canada, S-America	1.84	0.96%
Australia, New Zealand	1.09	0.75%
Arab, Africa etc.	0.07	0.20%

Table 1: The access ratio in October 2001 and July 2002 (user's countries).

Some statistical results from logs

Our site collects various sorts of log files. They are user access log data, user's translation log data (word, text, url, language pair etc.), and user's feedback comments.

Fig. 2 shows an average access ratio in a day (GST+0800). This result is quite different from that in (Hagiwara, 2000). This is because since services in our site are specified to translation, many of the users seem to use our service in thier office or school. In fact, if we refer to Fig. 3, we can find that access ratio in the weekend goes down. This result can be used for a general design of the system. That is, we can estimate the busiest hour in a day or a week to distribute tasks (i.e. to increase or decrease working computers).

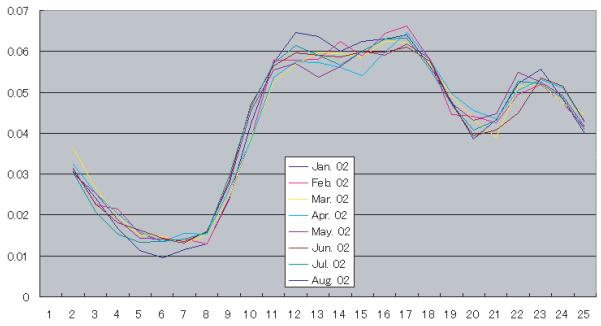

Figure 2: Average access ratio in a day

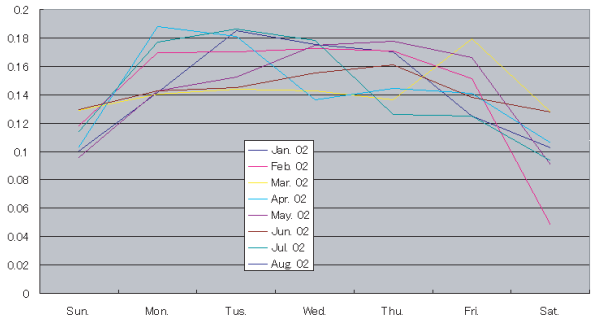

Figure 3: Average access ratio in a week

As shown above, our site has three types of services. The access ratio of each services during October 2001 and July 2002 is shown in Table 2. To tell a truth, this ratio is almost the same from the beginning of the service.

service	Web trans.	Text trans.	Online dictionary
ratio	8%	78%	14%

Table 2: Access ratio according to service.

In fact, before starting the service, we thought Web translation service will be used much more. However, the user did not seem to use Web translation service so frequently. The simple reason will be that the user does not know all url that has topics that he/she is interested in. In fact, at the beginning of the starting the service, most of users using Web translation service translated 'search engine' page like `yahoo`. In addition, in Text translation service, some of the users seemed to translate a part of article copied from a certain web site like newspaper. Another reason why they did not use Web translation may be caused by the translation speed. This is because Web translation requires slightly long time and translates all pages that may include unnecessary information.

Fig. 4 shows access frequency in terms of language-pair during October 2001 and July 2002. Actually, since English to Japanese translation service has been off from April 2001

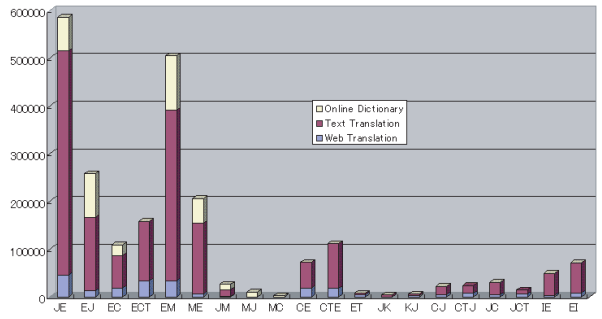

Figure 4: Access frequency in terms of language-pair

to March 2002, the access frequency of it is slightly lower than that to be estimated. Anyway, the fact that the access frequency of English to Malay translation service is high is natural and predictable.

From another log analysis (not shown in this paper), we can find an access frequency of translation between Indonesian and English has been increasing. Indonesia has more than 200,000,000 populations, but internet infrastructure has not yet been maintained well. We think that in the future there will be a certain chance in translation between Indonesian and English. Even now, some user like to use this service to read article written in Indonesian. In this sense, translation between English and Hindi or Tamil might have chance. However, we do not have translation system between English and Hindi or Tamil. In another reason that Hindi movies are quite in boom in the world, that translation will be important.

Table 3 shows the access ratio in terms of the domain in April 2002. We show the result in terms of language-pairs. The result can be easily predicted[2].

Next, we will show the frequently translated sites. When we started the service (Dec. 2000), the most translated site was 'search engine' site like `www.yahoo.com`. However, recently, the tendency has changed. As shown in table 4, the users input the real url to be translated. These facts show that the internet users have interests on a certain topic before accessing the internet. They would previously check some information like url from magazines or other media. In fact, search engine is important to do internet-surfing to find unknown site, however, we think chance will exist other than search engine.

We show that the user's action to the translation site has changed. In addition, if we check the url and tranlated sentences, we can find long term transition or change in the user's interests. Accordingly, we can also guess the next trend. Of course, this is a chance.

[2]You can find that .sa (Saudi Arabia site) quite frequently used the translation service. However, this is not chance. This is because, as shown in the next section, they used our service as a proxy.

Unpredicted usage in our site

After starting the service, we found very unique but unpredicted usage of our site. We do not think they are chance, however, we think that these experiences will be useful in risk management in the future. However, since they are risk management, they will be chances.

Used as proxy Our site was used as proxy to see sites that include erotic contents. In fact, in Malaysia, Singapore, and Islamic countries, seeing erotic media is banned. Therefore, as to internet, at the entrance of the country, such erotic sites are almost removed and cannot be easily accessed. However, since our company does not have such restriction, some users used our site as proxy to access erotic sites. They use wrong translation pair not to translate the page.

Piratical use of our site The translation engine in our site was directly used. From some site piratically used our site. They include translation function of our site in their page. Since more than one site used the same script, the script to use our site seemed to be delivered in the internet.

Does chance hide in the user's feedback?

What is a chance in multi-lingual translation site?

Ohsawa defined chance as follows(Ohsawa, 2002):

> A chance (risk) is a new event/situation that can be conceived either as an opportunity or a risk.

In the previous sections, we show some cases as chance. Even from the statistical log analysis that is thought of as average or trends, if we change our viewpoint, we can find a chance. In the paper, we define chance in multi-lingual translation as next trend or future demands.

Obtaining some data from the user's feedback

Our site has user's feedback. Usually, the users send us their feeling on our service and their requests. For example, we have been missing English to Japanese translation service. During the period, we have received a lot of requests to restart English to Japanese translation service. Moreover, the users require to start new translation services. They are Chinese to Japanese translation service, Korean to Japanese translation service, etc.[3] It is natural request from Japanese. These requests are major requests. However, we also received minor requests like translation service between Arabic and English. Malaysia where our site is located is Islamic country, therefore, a lot of Islamic persons seem to visit our site. Another reason will be that the Islamic person should read Koran, therefore, they need translation from or to Arabic.

Anyway, from these facts, we can get information that which language written articles do users like to read or guess which language written articles will be frequently read in the

[3]These translations have started. Especially as to Chinese translation, a lot of users used the service. However, Korean translation is not frequently used. We thought that since succor world cup was held in Korean and Japan in 2002, Korean translation will used more. We cannot find an answer, but the reason will be the problem in fonts.

future. Furthermore, we can guess an exact part of the world where peoples are interested in or will be interested in.

Guessing the user's interests from the user's feedback

In order to check the user's situation, we can check the user's action before sending the feedback.

We usually, check when the user returns severe comments or the user complains about mis-translation.

Very typical user's action is as follows. In the followings , "URL JE" means that the user used Web translation service, and the translation pair was Japanese to English.

```
15:10 URL JE http://www.yahoo.co.jp/
15:15 URL JE http://plaza21.mbn.or.jp/~jtomo
15:18 URL JE http://search.yahoo.co.jp/
```

The user first accessed `www.yahoo.co.jp` to search the site that will comtain his/her interested topic. After checking the page, he/she search another topic. This type of behaviour will be categorized in Double Helical Model(Nara and Ohsawa, 2002). We also, show the change of the user's interest causes chance(**?**). Anyway, we can find a certain chance like future boom from the user's behaviour.

In another case, first the user also used search engine in our site, then, he/she read newspapers. "CTE" means Traditional Chinese to English translation service. He/she read these newspapers on 12 Sep 2001. We think that that day, a lot of users accessed the newspaper site to check the news.

```
03:42 URL CTE http://www.orientaldaily.com.hk/
03:46 URL CTE http://appledaily.atnext.com/
03:47 URL CTE http://www.the-sun.com.hk/
03:49 URL CTE http://www.hkdailynews.net/
04:14 URL CTE http://www.visualmedia.com.hk/~kitty/MJ/
```

This user seemed to check Chinese papers. We cannot guess the reason.

We showed two examples. Actually, the user who uses translation will test the site or have some articles written in unknown language. Therefore, we cannot guess the short term transition of the user's interests. However, a certain information can be obtained. For example, the first user checked jewels in Japan. And we did not show the examples, some user read Japanese professional wrestring page, other user read Japanese animation page. These information will be a certain suggestion that person from a certain country is intersted in something in another country. Furthermore, if we also analyze translated texts, we can also find another information. Though, this paper does not deal with it, in the future, we should analyze the texts.

We think this information has relation to chance. Actually, user's comment is used to improve our service and suggestion to what language pair should be prepared. Moreover, if we research the information behind the comment, we can find new bussiness model.

Conclusions

We are now in the globalised world. Therefore, multilingual translation is important to work in globalised society. For individuals, it is also important to watch the world from the international viewpoints. So that, multi-lingual translation service is very important and seems to hide a certain chance in it. Actually, we provide free service, accordingly, a lot of users used our service. However, currently, it is slightly difficult to find charged service. We have been finding business model (this will be also chance) for machine translation.

Then, where is a chance in running multi-lingual translation site? Ohsawa defined chance as "a new event/situation that can be conceived either as an opportunity or a risk(Ohsawa, 2002)". As to web logs, Matsumura defined chance pages are the page that has novel topic and that has some links to such pages. From the viewpoint of commercially maintaining machine translation site, chance will be new business model On the other hands, from the viewpoint of sociology, chance will be results from the change of user's interests. They reflect the situation of society and if we use them correctly, we can guess novel or rare events before they explicitly appear in the world.

Matsuo proposed the method to find a new boom from the messages in an electronic message board(Matsuo, 2002). Though, we have not opened to the public, we prepare multilingual chat system. If we collect data from the chat, we can find a chance like Matsuo's proposal.

Anyway, finally, we show some (business) chance during checking web log in machine translation site.

- The users do not frequently search their interesting page by using search engine, but they check magazine or come from linked pages. Therefore, very famous site like `www.asahi.com` and `www.cnn.com` can live long. However, other infamous site should register not on search engine but on special site and special media like magazine or TV.
- Language pair should be English-centered. Actually, translation services between Chinese and Japanese and between Korean and Japanese are required, if we have translation services between Chinese and English and between Korean and English, it will be better for the users in the world. This is because, recently, Asian countries are thought of as a significant place from the various viewpoint like economy, culture etc.

References

Akinori Abe. User's interests change as Chance Discovery, Proc. of KES2002, 2002.

Analog HomePage, `http://www.analog.cx/`.

Masayuki Hagiwara. The Behavior of Web-user from the Internet Audience Research, In *Proc. of 4th. CmCC Symposium*, 41–48, 2000. (in Japanese)

Naohiro Matsumura, Mitsuru Ishizuka, and Yukio Ohsawa. Discovering promising new topics on the WWW, In *Proc. of KES2000*, 804–807, 2000.

Naohiro Matsumura, Yukio Ohsawa, and Mitsuru Ishizuka. Future Directions of Communities on the Web, In *Joint JSAI 2001 Workshop Post-Proc.*, 435–443, 2001.

Naohiro Matsumura, Yukio Ohsawa, and Mitsuru Ishizuka. Profiling of Participants in Online-Community, In Workshopnote of *Joint PRICAI02 Workshop*, 45–50, 2002.

Yutaka Matsuo, Yukio Ohsawa, and Mitsuru Ishizuka. Mining Messages in an Electronic Message Board by Repetation of Words, In Workshopnote of *Joint PRICAI02 Workshop*, 51–56, 2002.

Yumiko Nara and Yukio Ohsawa. Understanding Internet Users on Double Helical Model of Chance-Discovery Process, *New Generation Computings*, Vol. 21, No. 1, 2002 (to appear)

Yukio Ohsawa. Chance Discovery for Making Decision in Complex Real World, *New Generation Computings*, Vol. 20, No. 2, 143–163, 2002.

Language pair/From domain	
English-Chinese	
.net	26.54%
.tw	17.28%
.com	11.97%
.jp	2.73%
.my	1.42%
.hk	1.38%
Chinese-English	
.jp	29.19%
.net	15.43%
.com	11.31%
.sa	3.8%
.edu	3.28%
.au	2.81%
.ca	1.67%
English-Indonesian	
.id	20.87%
.net	5.31%
.com	2.14%
.jp	1.70%
English-Malay	
.my	23.27%
.sa	3.80%
Indonesian-English	
.com	10.46%
.au	10.10%
.sa	9.81%
.net	9.38%
.jp	6.29%
.sg	4.40%
Japanese-English	
.net	18.25%
.com	16.11%
.jp	13.05%
Japanese-Malay	
.jp	31.06%
.my	10.92%
Malay-English	
.sa	23.24%
.my	9.92%
.sg	8.82%
.net	6.19%
.com	6.03%
.jp	1.94%

Table 3: Who used the service?

Language pair frequently translated sites
English-Chinese http://www.ins.gov/ http://www.yahoo.com/ http://www2.nameplanet.com/mail/ http://web.icq.com/ http://mail.yahoo.com/
Chinese-English http://www.mba8.com/jijing/ http://www.emu-zone.net/ http://www.emuchina.net/ http://cn.yahoo.com/ http://www.taisha.org/ http://mcgi.163.com/term.html http://www.sina.com.cn/
English-Indonesian http://azzam.com/ http://www.yahoo.com/ http://taliban-news.com/ http://www.boycottisraeligoods.org/ http://www.khurasaan.com/ http://www.manutd.com/
English-Malay http://www.yahoo.com/ http://www.telemotive.com/ http://www.penanggolfresort.com/ http://www.bluehyppo.com/
Indonesian-English http://www.safitri.com/Artikel.htm http://www.detik.com/ http://www.awse.com/
Japanese-English http://www.lascachomania.com/ http://www.rakuten.co.jp/ http://www.yahoo.co.jp/ http://www.oudou.co.jp/
Japanese-Malay http://www.sony.co.jp/ http://www.docomo.jp/ http://www.asahi.com/ http://www.yahoo.co.jp/
Malay-English http://thehun.net/ http://www.yahoo.com/ http://www.emedia.com.my/ http://www.utusan.com.my/

Table 4: Frequently translated sites