QRselect: A User-Driven System for Collecting Translation Document Pairs from the Web

Kyo Kageura[1], Takeshi Abekawa[1], and Satoshi Sekine[2]

[1] Graduate School of Education, University of Tokyo,
7–3–1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan
{kyo,abekawa}@p.u-tokyo.ac.jp
[2] Computer Science Department, New York University,
715 Broadway, New York, NY 10003 USA
sekine@cs.nyu.edu

Abstract. In this paper we introduce a system that collects English-Japanese translation document pairs from the Web that are relevant to subject keywords specified by the user. The system, QRselect, is specifically designed to meet the needs of online volunteer translators who, in the process of translation, want to refer to a small and specific set of translation document pairs which are relevant to what they are translating. A system which collects relevant existing translated documents and makes them available for reference in the translation process will therefore greatly help these translators. Against this backdrop, we developed a prototype translated document collection system and evaluated its performance. We also examined the users' role in improving the system.

1 Introduction

The ever-expanding breadth of global communication on the Internet has recently been accompanied by an increase in the number of online volunteer translators [1], who translate online documents in a variety of fields such as politics, culture, area studies, sports, computers, etc. and publish their translations on the Web. Loose networks of translators dealing with similar or related subjects have been and continue to be formed.

These translators often play vital roles in distributing important information that would otherwise not appear in mainstream media and in promoting critical media literacy in the age of Internet. Despite this, and despite the fact that there are many translation-aid systems, there has been no system to date that specifically aims at aiding online volunteer translators.

Against this backdrop, we are currently developing a system that aids online volunteer translators. As a part of this, we developed QRselect, a system that collects translation document pairs from the Web, based on translators' requests or specifications. This is, in a sense, a system that constructs each translator's private digital library of relevant translation document pairs, and enables translators to refer to relevant existing translated documents systematically.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 131–140, 2007.

In this paper we will first explain the basic needs of translators that led us to the development of QRselect. We will then introduce the basic structure of QRselect and how it works, and present the results of an evaluation of the system's basic performance, with a diagnosis. We will also discuss the response of translators to the system and the extent to which they are willing to cooperate with us to improve the system's performance and to make it fully effective.

2 Translators' Need for Existing Translated Documents

One of translators' key needs – we interviewed eight online volunteer translators and also obtained opinions by e-mail from twelve other translators – is the ability to refer to and recycle bilingual translations of various language units, such as proper names, repetitive quotations, domain-dependent expressions, etc., from existing translated texts that deal with the same or similar topics and which are judged to have sufficiently high quality. Two things characterise this need:

(i) What translators are looking for within existing relevant translations are not *linguistically similar examples*, but concrete information showing *translation conventions* relevant to the group of texts to which the text that the translator is translating will belong, a group characterised by such basic traits as subject topics, register, etc. In other words, what translators would like to be able to refer to in the process of translation is an *archive of relevant texts*, not an unanchored *corpus* that represents *language* in general [7]. This information need is different from and complementary to the need to check a broader range of reference sources.

(ii) Translators want a system that helps them refer to what they want to refer to. When they look for existing translation document pairs that are relevant to what they are translating, the documents translators refer to tend to be very dense and few in number – in the order of tens or often less. Size cannot compensate for the relevance to their requirements; it is the fact that they checked the documents that they thought they needed to check, which may be few in number, that enables them to finalise the translation. This incidentally corresponds to the claim made in the field of natural language processing that the usefulness and effectiveness of a corpus depends qualitatively on the aim and that a larger corpus may not necessarily perform better [8,10].

Currently, many translators take several steps in order to refer to existing relevant translation document pairs, including checking pages they know as well as using Google to find new translations. What is desirable here, therefore, is to automatise this process by developing a mechanism to collect a set of translation text pairs which are relevant to the text that the translator is translating, according to the translator's request. As such a group of texts is defined vis-à-vis the translator's need which is mostly determined by the particular text which the translator is translating or the particular subject area with which the translator is mainly concerned, a system which collects such translation document pairs should function in a user-driven manner. This requirement contrasts with

the need to collect large bilingual corpora from the Web for use as a basic resource for natural language processing [4,12] or to obtain a broad-coverage list of bilingual word pairs from a large corpora [2,5,6,9,11,16].

Translators regard these two types of information as qualitatively different. This can be understood in analogy to the behaviour of patent translators, who *always need to check* existing translations in the archive of patent documents. They must do so as much to be able to make their own decisions with confidence as to look for translation expressions which they do not know. In order for translators to make their documents authentic and acceptable to readers, the process of situating the translated document within a set of relevant existing documents is an essential process of translation. Irrespective of whether translators adopt existing expressions and phrases or not, and irrespective of whether they have basic information on expressions from other reference sources, this process is therefore a *sin qua non* for translation and cannot be compensated for by other information sources, however large they may be.

Taking the above into account, we have developed QRselect, a user-driven system to collect from the Web a specific set of translation document pairs relevant to the document that the translator deals with.

3 The QRselect Prototype System

3.1 Basic Structure of the System

The QRselect system operates in two different modes, i.e. dynamic mode and batch mode. In the following, we focus primarily on an explanation and evaluation of the dynamic mode. Figure 1 shows the overall framework of the QRselect prototype system.

1. The user inputs Japanese keywords relevant to the topic of the document that the user is translating. In batch mode, the user registers a list of URLs under which translated documents relevant to the translator's interests are published frequently.
2. The system retrieves a specified number of Japanese Web documents relevant to the Japanese keywords, using an existing search engine. When the retrieved documents are evaluated as a translation in steps 3 to 6, the search is expanded to "similar pages", as the site that includes a translated document may well include many translated documents. In batch mode, the system checks the update log of the registered sites, and collects the newly published documents.
3. For each retrieved page, the system detects the anchor link given by the `<a http=""> </a>` tag, traverses the anchor link, and obtains the target page of the link. The system only traverses the anchor links in close proximity to the reserved words, which should indicate that the target page is the original document. After having analysed scores of translated documents, we adopted seven reserved words, i.e. "原文" (original document), "ソース" (source), "英語" (English), "元記事" (source article), "オリジナル" (original), "原著" (source text), "原語" (source language).

4. For each Japanese document retrieved in step 2 and for each document detected in step 3 as a source document candidate, the system applies a simplified version of Webstemmer [15] and extracts the textual area.
5. For the pair of textual areas extracted in step 4, the system calculates the similarity of the texts. This is done by transforming the words in the English text into Japanese using a publicly available large-scale English-Japanese dictionary [3]. Currently, simple content words are used. The similarity is calculated by the ratio of the number of matched word tokens to the total number of Japanese word tokens in the text.
6. The system identifies the pairs whose similarity score given in step 5 is above a given threshold as translation document pairs.

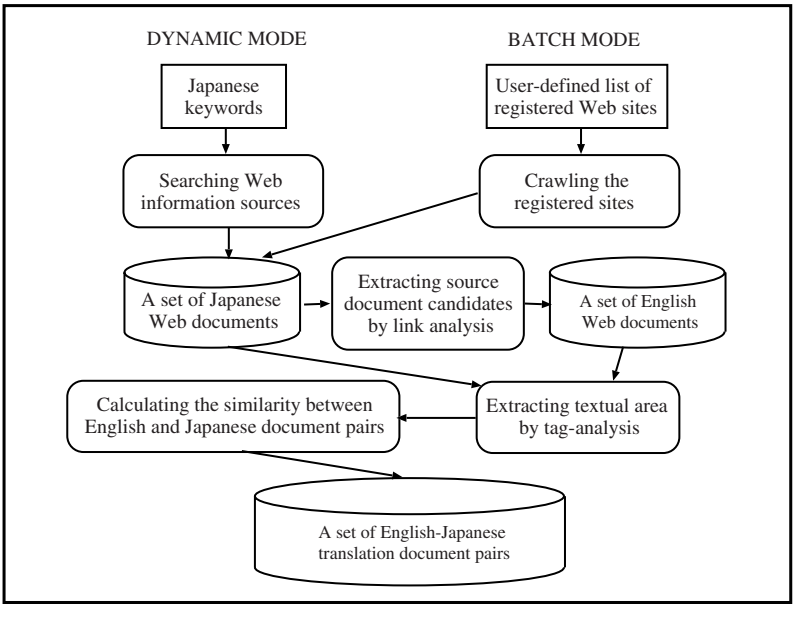

Fig. 1. The Overall Framework of the QRselect Prototype System

The prototype is implemented in Java and operates on Tomcat. Note that the search proceeds from Japanese documents to English documents, which is more efficient because the number of Japanese documents translated from English is much larger than the vice-versa.

3.2 Quantitative Evaluation

We evaluated the dynamic version of the QRselect prototype using 33 keyword sets provided by two translators and two evaluators. The keywords were roughly categorised into five groups, i.e. (a) geographical areas or countries; (b) current affairs; (c) information technology; (d) law, culture and sports; and (e) others.

The keyword sets are given in Table 1. The experimental settings were as follows: (i) the target number of Japanese pages to be retrieved in the experiment was set to 100, which means that for each keyword set, we retrieved 100 Japanese Web pages and detected the translation pages among them; (ii) Google was used as a search engine; (iii) the similarity threshold was set to 0.1, on the basis of a preliminary analysis of the performance. We decided on 100 Web pages because all the translators we consulted scan less than 100 snippets when they check related translations. Due to the system configuration, the total number of retrieved pages may not be exactly 100. Note also that, in servicing the system, we may use the Yahoo Japan search engine because the API provided by Yahoo Japan allows more searches per day than the Google API. There is not much difference in the performance of these two search engines as far as the Japanese pages are concerned. The choice of search engine is external to the system.

Table 1 shows the results of the evaluation. We only give basic figures for precision and recall, and do not give other IR-like performance measures applied to the ordered list of outputs [13], even though it is possible to order the output by means of similarity scores. This is because the user requirement for QRselect is to provide a sufficient amount of information with a manageable level of precision, and the concept of a "trade-off" between precision and recall is not relevant. Each row indicates a keyword set (33 in total). The meaning of the signs in the column are as follows:

- A: The total number of Japanese pages retrieved.
- T: The number of pairs consisting of Japanese translations and their English originals which are accessible through the link (= MH + Y). This is the target that the QRselect prototype should cover.
- C: Translation pair candidates output by QRselect. This is divided into:
  - CY: Correct output.
  - CE: Error, i.e. they are not a translation pair.
- M: Miss, i.e. when QRselect does not output the translation pair but the Japanese page is a translation of some English documents available online. This is further divided into[1]:
  - MH: Original page exists in HTML or related tagged forms.
  - MI: No tagged links, erroneous links or original pages.
- N: The number of non-translation Japanese pages which are correctly identified by QRselect as non-translations.
- P: Precision = CY/C.
- R: Recall = CY/T.

All in all, the system gave a modest performance, with the overall precision being 0.74 and recall 0.35.

[1] We also checked for misses caused by the fact that the document was in pdf format, but there were none in the data we evaluated.

Table 1. Evaluation of the QRselect Prototype System for 33 Keywords

Keyword set	A	T	C	CY	CE	M	MH	MI	N	P	R
(a) Colombia	92	1	1	0	1	2	1	1	89	0	0
(a) Colombia, drug	58	25	17	17	0	11	8	3	30	1	0.68
(a) Colombia, Uribe	32	20	19	19	0	1	1	0	12	1	0.95
(a) Venezuela	95	3	1	1	0	3	2	1	91	1	0.33
(a) Venezuela, Chavez	81	3	3	3	0	0	0	0	78	1	1
(a) Falluja	97	14	3	3	0	19	11	8	75	1	0.21
(a) Falluja, Aljazeera	96	31	7	4	3	29	27	2	60	0.57	0.13
(a) Baghdad, resistance	98	36	17	17	0	20	19	1	61	1	0.47
(b) Abu Graib, human rights	97	26	14	7	7	19	19	0	64	0.5	0.27
(b) separation wall	93	11	6	3	3	12	8	4	75	0.5	0.27
(b) Chomsky, Iraq, invasion	96	14	9	9	0	7	5	2	80	1	0.64
(b) Katrina, Hispanic	93	4	21	1	20	3	3	0	69	0.05	0.25
(b) China, censorship	98	30	12	12	0	20	18	2	66	1	0.4
(b) Sellafield, BNG	93	8	1	1	0	12	7	5	80	1	0.125
(b) Said, Arafat, Zionism	51	9	8	7	1	5	2	3	38	0.88	0.78
(b) Catholic, contraception	94	9	6	4	2	6	5	1	82	0.67	0.44
(b) veterans, suicide	91	3	4	1	3	4	2	2	83	0.25	0.33
(c) Torvalds	97	40	4	4	0	36	36	0	57	1	0.1
(c) Stallman	94	17	13	10	3	7	7	0	74	0.77	0.59
(c) Napster, file exchange	97	52	31	31	0	22	21	1	44	1	0.60
(c) Halloween document	79	2	5	0	5	7	2	5	67	0	0
(c) Linux, developing countries	93	13	15	10	5	4	3	0	1	74	0.67
(c) Google, library, scan	96	16	2	1	1	15	15	0	79	0.5	0.06
(d) Free culture	94	25	6	4	2	21	21	0	67	0.67	0.16
(d) Krugman, column	97	15	22	7	15	10	8	2	65	0.32	0.47
(d) Seattle Post, Mariners	94	36	1	1	0	40	35	5	53	1	0.028
(d) China, football	99	11	1	1	0	17	10	7	81	1	0.09
(d) Shunsuke, local, media	87	0	0	0	0	0	0	0	87	—	—
(d) F1, interview, driver	99	62	1	1	0	64	61	3	34	1	0.02
(d) Ghibli, export	63	1	0	0	0	2	1	1	61	—	0
(d) Hollywood, star, article	97	1	1	1	0	0	0	0	96	1	1
(e) Nablus report	100	23	11	10	1	23	13	10	66	0.91	0.43
(e) John Pilger	95	26	19	18	1	10	8	2	66	0.95	0.69
Total	2936	587	281	208	73	451	379	72	2204	0.74	0.35

3.3 Diagnosis

The overall figure, however, means little because the user is concerned only with the performance for a specific subject topic. Of much greater importance are the causes of errors (CE) and misses (MH).

Errors (CE) can be divided into the following patterns:

1. The Japanese pages are not translations, but refer to English documents as an information source or as related information. This pattern accounted for 67 cases, 41 of which were caused by a single Web site.

2. The Japanese pages are translations, but the system traversed the wrong link and detected false pages. This pattern accounted for six cases, four of which did not have correct links to the original. In two cases the system traversed the wrong link because many translated texts were contained in a single Japanese page.

Misses (MH) can be divided into the following patterns:

1. The link was not detected by the QRselect system, because the anchor link to the original document was not accompanied by the reserved words assumed by the QRselect system. This type of miss accounted for 187 cases.
2. The system properly detected the original text by traversing the link but identified that the pages were not translations at the stage of similarity calculations. This type of miss accounted for 192 cases. This was caused by (i) poor performance in the extraction of the textual area by tag analysis, and/or (ii) the limitations of the simple English-to-Japanese word transformations in the dictionary-based similarity calculation. Although these two causes are interdependent and it is difficult therefore to specify which is the main cause, in at least 25 cases the improvement of tag analysis is essential because in these cases the tag analysis failed to identify the main textual area. On the other hand, in 60 cases the similarity score was above 0.08 (note that the threshold was set to 0.1). For these cases it can reasonably be predicted that the improvement of dictionary-based matching methods will lead to a reduction in misses.

In summary, errors and misses were caused by three main factors, i.e. (a) mistakes in detecting anchor links to the original English pages, (b) errors or insufficiencies in textual area extraction by tag analysis, and (c) insufficiencies in dictionary-based similarity evaluation between Japanese and English pages.

4 The Social Model: Translators' Potential Contributions

Among the factors that caused the errors and misses just summarised, the latter two ((b) and (c)) are problems that should be technically solved. On the other hand, the first issue, i.e. mistakes in detecting anchor links to the original English pages, can and should be solved socially, although technical refinements are needed. The social solution can be achieved by promoting the involvement of and contributions by translators: If translators realise the merit of recycling and referring to information made by translators working in similar fields through QRselect, it is expected that they will agree to provide anchor links to the original English texts in a controlled and consistent manner. This is essential for such systems as QRselect, because QRselect has specific target users (online volunteer translators) to whose tasks it aims to contribute, and any successful system of this nature should evolve via interaction between the system and its users.

We consulted eight online volunteer translators about the possibility of adding extra tags or keywords to improve the performance of QRselect. Although most

translators we consulted refused to use extra HTML- or XML-based meta-tags, most of them were at the same time happy to provide explicit anchor links to English originals by <a> tags on the translated document page, and a basic word near the anchor link to indicate that the link is to the original English page. Only one translator we consulted was positive about the use of meta-tags. This is probably partly due to the fact that most translators are reluctant to concern themselves with the technical aspects of the publication of their translated documents, and partly due to the fact that most online volunteer translators publish their own essays and comments as well as translations on a single page and do not manage sites specialised for translations.

Five translators also said that they would modify existing translations which they have published online to make them conform to a format that can be dealt with by QRselect.

If that sort of cooperation can be assumed by online volunteer translators, the 187 misses caused by a lack of reserved words near the anchor links would be avoided, as well as a certain number of misses categorised under MI. In addition, such cooperation may well contribute to reducing errors as well, because it would help reduce the erroneous identification of incorrect links to some referred pages. As the development of QRselect was triggered by requests from translators working online, there is a good chance of extending translators' cooperation in improving the system performance, which in turn would contribute to making a wider range of reference functions available to the translation community. Among the eight translators we consulted about this issue, four started adding specific keywords systematically to their anchor links. We are hoping that this cycle will not only enhance the performance of QRselect but also activate further online translators' activities through the use of QRselect.

5 Conclusions and Outlook

In this paper, we have introduced QRselect, a user-driven system for collecting subject-specific translation document pairs from the Web, and evaluated its performance. We also discussed the social aspect of the system, in which translators would contribute to the improvement of the system to their own potential benefit.

Although there is a general trend in the realm of Web-based information systems towards dealing with a huge, ever-increasing amount of data, there are areas where users require a limited but relevant range of data from the Web which is specific to their concerns. This can be understood in analogy to the relation between the quest for the universal library and the necessity for a personal library. These requirements are independent and complementary, and one cannot compensate for the other. Collecting and recycling existing translation document pairs relevant to the document that the translator is translating is one such area where relevance to the user rather than largeness of scale is required. Although the current performance of the QRselect prototype is moderate, there is a good chance of improving the system performance to make the system fully

serviceable, especially given that we can assume translators' cooperation in making the system more effective.

In the evaluation, we focused on the dynamic mode of the QRselect system. In the real-world setting, however, we envisage that these two modes, i.e. the dynamic mode based on keywords and the batch mode based on registered Web sites, will be used in a mutually complementary manner. Although the current system performance is moderate, there is a good chance of improving the performance to a realistically useful level by improving technical aspects and by obtaining translators' cooperation.

Currently, we are enhancing the QRselect prototype in three directions:

1. Improving the performance of the module that extracts the textual area from tagged Web documents by tag-analysis.
2. Improving the granularity in calculating the similarity between Japanese and English documents using dictionaries.
3. Incorporating the non-linguistic clues to enhance the performance of similarity calculation.
4. Expanding the system so that it can deal with language pairs other than English and Japanese. We are currently modifying the system to cover English-French translations. We are also developing an interface through which users can modify and adapt the system to language pairs that the user wants.

In the fully operational system into which QRselect is incorporated, recyclable bilingual linguistic units, such as proper names, technical terms, fixed phrases and quotations [14], will be automatically extracted from translation document pairs and these units will be made available to users. In addition, we are also planning to extend the QRselect system by adding a module which automatically generates Japanese keywords when the translator specifies an English document to translate, instead of asking the translator to specify keywords to activate the QRselect dynamic module.

Acknowledgements

This research is partly supported by grant-in-aid (A) 17200018 "Construction of online multilingual reference tools for aiding translators" by the Japan Society for the Promotion of Sciences (JSPS) and the National Institute of Information and Communication Technology (NiCT). The authors would like to thank the anonymous reviewers for their valuable comments.

References

1. Boitet, C., Bey, Y., Kageura, K.: Main research issues in building web services for mutualized, non-commercial translation. In: Proceedings of the 6th Symposium on Natural Language Processing (2005)
2. Cao, Y., Li, H.: Base noun phrase translation using web data and the EM algorithm. In: Proceedings of COLING 2002, pp. 127–133 (2002)

3. Eijiro (2006), http://www.eijiro.jp/
4. Fukushima, K., Taura, K., Chikayama, T.: Fast and accurate method for detecting English-Japanese parallel texts. In: Proceedings of the COLING/ACL Workshop on Multilingual Language Resources and Interoperability, pp. 60–67 (2006)
5. Fung, P.: A statistical view on bilingual lexicon extraction. In: Proceedings of AMTA 1998, pp. 1–16 (1998)
6. Huang, F., Zhang, Y., Vogel, S.: Mining key phrase translations from web corpora. In: Proceedings of HLT/EMNLP 2005, pp. 483–490 (2005)
7. Kageura, K.: The status of "corpus" in human translation. In: Proceedings of the 12th Annual Meeting of the Japan Society of Natural Language Processing, pp. 452–455 (2006)
8. Morin, E., Daille, B., Takeuchi, K., Kageura, K.: Bilingual terminology mining – using brain, not brawn comparable corpora. In: Proceedings of ACL 2007, pp. 664–671 (2007)
9. Nagata, M., Saito, T., Suzuki, K.: Using the web as a bilingual dictionary. In: Proceedings of the Workshop on Data-driven Methods in Machine Translation, pp. 95–102 (2001)
10. Péry-Woodley, M.-P.: Quels corpus pour quels traitements automatiques? Traitement Automatique des Langues 36, 213–232 (1995)
11. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of ACL 1999, pp. 519–526 (1999)
12. Resnik, P., Smith, N.A.: The web as a parallel corpus. Computational Linguistics 29, 349–380 (2003)
13. Sakai, T.: For the realisation of better IR systems. IPSJ Magazine 47, 147–158 (2006)
14. Shinagawa, T., Mori, T., Kageura, K.: Extraction and alignment of textual blocks from online translation document pairs. In: Proceedings of the 12th Annual Meeting of the Japan Society of Natural Language Processing, pp. 520–523 (2006)
15. Shinyama, Y.: Webstemmer (2006), http://www.unixuser.org/~euske/python/webstemmer/index.html
16. Utsuro, T., Kida, M., Tonoike, M., Sato, S.: Collecting novel technical term from the Web by estimating domain specificity of a term. In: Matsumoto, Y., Sproat, R.W., Wong, K.-F., Zhang, M. (eds.) ICCPOL 2006. LNCS (LNAI), vol. 4285, pp. 173–180. Springer, Heidelberg (2006)