生涯学習基盤経営研究 第 37 号 2012 年度

Web-based Archiving of Parallel and Comparable Documents for Online Translators

Kyo KAGEURA † Ryo MURAYAMA †

†Graduate School of Education, The University of Tokyo

This paper introduces a Web-based system that archives parallel and comparable online documents. The system, QRpac, is specifically designed to meet the requirements of online volunteer translators who need to refer to relevant translation document pairs as well as comparable document sets. While many systems have been proposed so far to construct parallel and/or comparable corpora from the web, there have been none that directly respond to online translators' needs. The system is currently fully operational, and operates in a seamless manner upon request by the user. This paper puts emphasis on the context within which the system is required, designed, and used, which we believe is essential and more important than "rote" evaluations.

online translator, terminological resources, comparable archive, parallel archive

Contents

1 Introduction

In recent years we have witnessed a growing number of translation activities carried out online in which online documents in a variety of fields such as politics, culture, sports, computing and so on, are translated. With the ever-expanding breadth of global communication on a wide range of topics and issues and in a wide variety of languages on the net, this trend is set to continue or accelerate for the foreseeable future. As a result, the activities of online translators (we define "online translators" here loosely as people translating online or electronic texts, publishing their translations online, and relying substantially on online resources in the process of translation) are becoming more and more important, while a number of online or installable platforms have recently become available to support certain aspects of translation activity, such as managing translation memory[1] [2] [3] and managing translation workflow and collaborative processes[4] [5] [6]. There are, however, not many platforms, services or sites which provide online translators with basic reference resources such as standard dictionaries or parallel or comparable corpora suitable for translators, although exceptions do exist[7]. We are developing a system that helps online translators by providing them with necessary reference resources.

From the point of view of constructing lan-

guage resources, research into the construction of parallel or comparable corpora tends to be regarded as an issue related to linguistics, machine translation (MT) or cross-lingual IR rather than human translation activities[8] [9] [10] [11] [12] [13] [14] [15] [16] [17], although here again there are some exceptions[18] [19]. In discussions with online translators, we found that they need stratified reference resources, and that we need to be aware of the position of parallel and comparable documents within this stratified reference resources. Taking into account this position of parallel and comparable documents within reference resources for online translators, we have developed a system, QRpac, which enables users to create parallel and comparable archives. QRpac combines established technologies and software as modules. What is important and innovative in QRpac is thus not the technologies for archiving but the perception of the usefulness of parallel and comparable archiving in the translation process and the functional design related to the division of labour between users and the system. This paper presents the system together with the basic design concept that reflects translators' needs.

2 Archiving and translation

2.1 The position of "archives" in the translation process

Although many non-translators tend to regard translation more or less as a linguistic process, translation deals first and foremost with texts[20]. This can be illustrated by taking a situation in which a translator is translating a document which contains a citation from, say, an international treaty. A proper translator would (a) identify that the passage is an extract from an international treaty, (b) refer to the official translation in the target language, if there is one, and (c) decide whether the direct use of the official translation is required or some adaptation of it is possible or preferable. This implies two other important traits of the translation process, i.e. translators do not simply transform language structure in a rote manner, but make decisions, and translators deal with texts as singular products rather than as language samples.

This explains why being bilingual does not on its own make a person a good translator – translation is not so much concerned with transforming source language (SL) expressions into target language (TL) ones using linguistic rules as with giving the target text a place in the set of existing related texts in the TL that corresponds to the place of the source text in the set of existing related texts in the SL. Given this nature of translation, we can say that translators deal with three different levels of language in its broader sense[21]:

1 Linguistic level: translators should have sufficient command of the grammar in both the SL and TL and need to refer to high-quality lexical resources.
2 Text-archive level: translators need to refer to a group of historically and socially accumulated texts in the SL and TL which are relevant to the text they are translating. The relevance here is ultimately supported not by linguistic or topical similarity but by the unique names of authors and translators as well as the unique time of production. From the point of view of linguistic units, this level is concerned not only with passages but also with technical terms, for which translators *must use* established TL terms used in related TL texts.
3 Text-corpus level: translators need to refer to general corpora to check the possible range or usage patterns of TL expressions. Unlike data in archives, data in corpora do not need to be identified as singular existences in history.

In relation to these three levels, it is essentially the text-corpus level which can be addressed by parallel or comparable corpora as products of most existing research. This also holds for corpora constructed for use in helping translators[22]

[23] [24]. While there are studies in the fields of NLP, lexicography and translation studies dealing with the construction and use of linguistic level and text-corpus level reference resources, there is a paucity of research addressing reference resources at the text-archive level. QRpac explicitly intends to help online translators construct text-archive level reference resources.

2.2 Useful archives for translators

If we could construct a universal translation archive containing all translated document pairs ever produced in a given SL and TL, with proper identifying information including names of authors and translators, dates, and so on, and from which translators could retrieve a set of translations relevant to their translation task, we might think that that would make an ideal text-archive level reference resource consisting of parallel documents.

There are, however, two problems with this framework. The first is that constructing a universal translation archive is not practically feasible given a range of social limitations, including the issue of copyright[25]. In addition, partly due to the social restrictions represented by copyright issues, there is a massive number of translated documents which have not been digitised. Thus, even if we could make an archive of all the translated documents that exist online, the archive may not match users' expectations, in which case it is unlikely that it would be used[26].

The second issue is related to the fact that not all the SL texts relevant to the particular text that a translator is translating were translated into the TL. Therefore, simply collecting translated documents, even if we could overcome all social barriers and construct a universal translation archive, would only cover a part – however essential this part may be – of all that would be needed by translators. It would be more useful and preferable for the archive to contain relevant SL and TL texts even if they did not match each other – in other words, a comparable archive. Note that this situation is analogous to the position and role of parallel and comparable corpora.

To overcome these problems, we adopted the following goals in our strategic definition of QRpac:

1 Development of a user-driven system. Instead of constructing a large archive for potential users, QRpac helps individual users to construct their own parallel and comparable archives. Although it may seem cumbersome for translators, some online translators make archives of documents relevant to their translation activities using search engines and heuristics. To facilitate the process would thus greatly help these translators.

2 Creation of a combined parallel and comparable document crawler. As will be explained in the next section, QRpac allows users to iteratively archive parallel and comparable documents, in the process letting them make decisions about the direction of archiving and the range of documents to be archived.

By making the archiving process and thus the resultant archive personal, we can guarantee that it will be used, because users will be fully aware of the nature of the archive they constructed and will understand what they can – and cannot – get from it. Note that this is one of a number of general prescriptions that are designed to ensure that reference resources are actually used[27].

3 The QRpac system

3.1 Overall structure

QRpac consists of a parallel document crawler, a comparable document crawler, a term extractor, and a user interface which allows users to interactively control the crawling directions. Users can iteratively construct both parallel and comparable archives in a single session. The archiving process proceeds as follows:

1 As a first step, the user can choose either to construct a parallel archive or to construct a

comparable archive. At the same time, the user can specify the number of iterations for the parallel and comparable archiving process in the session.

2 If the user chooses to construct a parallel archive, s/he specifies a set of monolingual terms which reflect the topic or area of the archive s/he intends to construct.

3 If the user chooses to construct a comparable archive, s/he specifies a set of bilingual term pairs which reflect the topic or area of the archive s/he intends to construct.

4 The user can iterate the process in a session, in other words, s/he can use the set of terms or term pairs extracted from the documents obtained in the previous iteration as seed terms in the new iteration.

Figure 1 illustrates how QRpac works[28]. Essentially, QRpac enables users to iteratively apply both the parallel document crawling and the comparable document crawling process in a seamless manner by allowing users to select seeds at the end/start of each process. For the three modules, i.e. parallel document crawler, comparable document crawler, and term extractor, QRpac uses existing methods and software packages, with some modifications and adaptations.

3.2 Parallel archiving

Several systems or methods have been proposed for crawling bilingual parallel documents[29] [30] [31] [32] [33] [34] [35] [36]. While they show good performance, the system or the set of collected documents are not necessarily suitable for online translators, as most of them aim at constructing corpora for computational applications such as MT or IR. In some cases, the system does not allow users to control the crawling, while in other cases the system collects "parallel" segments without distinguishing different span of units or the origin of parallelism. Translators stratify strategies to check reference information[37], and most translators want parallel documents consisting of source and translated document pairs dealing with the topic relevant to the documents that they are translating. Conceptually, this is related to the distinction between corpora and archives. The user's control is critical in giving data the status of a "personal" archive, whereas it is not necessarily relevant in the case of corpora. Taking these factors into account, we decided to use QRselect[38], which aims specifically at translated document pairs relevant to individual users.

Briefly, QRselect works as follows:

1 The user inputs keywords in a language relevant to the topic of the document that s/he is translating.

2 The system retrieves a specified number of web documents in that language relevant to the keywords. When the retrieved documents are evaluated as a translation (see steps 3 to 5), the search is expanded to all documents within the same URL domain.

3 For each retrieved page, the system detects the anchor link and judges whether it refers to the source language document or not, by checking for the existence of "reserved words" or "anchor link words" such as "original," "source language documents," etc[39].

4 For each document retrieved in step 2 and each document detected in step 3, the system removes HTML tags, extracts the textual area, and calculates the similarity of the texts by using word-level translations[40].

5 The system selects pairs whose similarity is above a given threshold.

Experiments showed that the system works very well for crawling translated documents relevant to the user's request[41]. What is important here is that, although there are some topics for which translation document pairs cannot be detected, users can reasonably guess whether this is due to system shortcomings or because such document pairs do not exist in the first place, which is an important condition for reference systems to be used by translators[42].

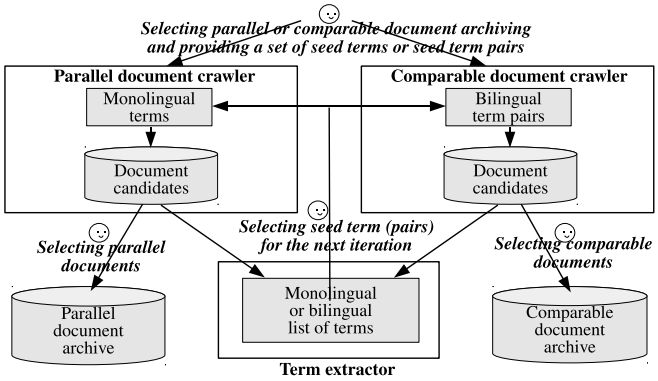

Figure 1: The Overall Flow of Functions in QRpac

3.3 Comparable archiving

For the comparable document archiving module, we use BootCaT[43], a corpus crawler widely used for constructing web-based corpora and used in translation training[44] [45] BootCaT works as follows:

1 The system accepts several terms as seeds from the user.
2 It randomly generates tuples (typically triples) of the seeds and sends them as a query to the search engine.
3 It retrieves the top of hit pages and applies filtering, removing redundancy and cleaning. In the process, it allows the user to choose which hit pages s/he wants to retain.

BootCaT has recently been used to construct comparable corpora from the web[46]. We also extended the use of BootCaT so that it enables users to collect comparable documents, by adding a rapper interface to the BootCaT engine, as follows:

1 The user inputs bilingual term pairs as an initial seed.
2 The system generates parallel tuples (typically triples) of the seeds and sends them as queries to the search engine.
3 The system retrieves the top of hit pages in each language and applies filtering, removing redundancy and cleaning. In the process, it allows the user to select which pages should be retained.

Here again, the distinction between corpus and archive is maintained by the process of the user's decision in choosing the documents.

3.4 Term extraction

Both the parallel document crawling module and the comparable document crawling module start from seed terms (monolingual terms in the case of the former, and bilingual term pairs in the case of the latter) and obtain (parallel or comparable) bilingual documents sets. If we add a term extraction routine to the resultant document sets, we can thus connect these modules so that the user can seamlessly iterate combinations of parallel and comparable corpus crawling to her/his satisfaction.

Many methods of automatic term extraction have been proposed so far. For the term extraction module, we chose a term extractor originally developed and tested for Japanese[47]. According to the developers' experiments, the system shows high performance in extracting especially complex terms. In addition, it is based on a simple, language-independent idea, and works for essentially any language.

4 Interactions with users

In the QRpac system, the user specifies the file name for the archive, and chooses whether s/he

Figure 2: The Basic Interface of QRpac (Parallel Archiving Mode)

wants to start by collecting parallel documents or comparable documents. Figure 2 shows the basic interface for archiving parallel documents. Here, seed terms can be specified either directly on the page or by uploading the file containing seeds. In Figure 2, parameters reflect the parameters of QRselect as the interface shown is for parallel archiving. In the comparable archiving interface, the parameters reflect BootCaT parameters. In both modes, the user can choose whether s/he wants to make a choice among retrieved documents or accept the result straightforwardly. In the parallel archiving mode, the user can also specify "anchor link words" or reserved words for improving the precision of detecting source language texts.

The user can iterate in one session parallel and comparable archiving. If the user clicks on the "Add Iteration" button, a new interface panel is added at the bottom of the page. The archiving processes are connected by the term extraction module: the system extracts terms from the parallel or comparable archive constructed from the previous iteration, and the user can choose seed terms from them to be used for the next archiving process. Figure 3 shows the interface for term selection. We are currently planning to ask several online translators to test the user friendliness of the interface.

5 Remaining issues

5.1 Evaluation

As QRpac uses established mechanisms for parallel and comparable document crawling and for term extraction, we can reasonably rely on the quantitative evaluations given in the work reported for the original methods[48] [49] [50]. We are

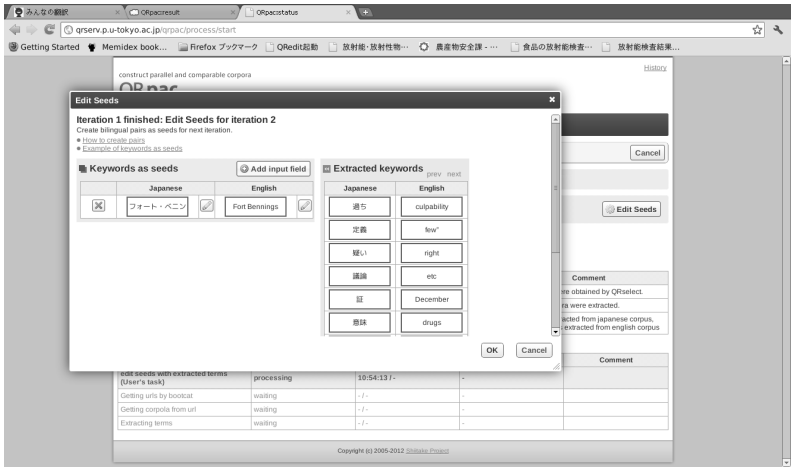

Figure 3: The Term Selection Popup (with Background Window Showing Processing Status)

therefore focusing on qualitative evaluation, by means of feedback from translators who are expected to use the system. Full qualitative evaluation is, however, being hampered by certain technical problems.

5.2 Technical issues

QRpac currently has two main technical challenges. The first is the processing time. At present, it takes several to scores of minutes to finish the first crawling, especially for parallel archiving, and makes archiving a full and independent task rather than a side task in the process of translation for the user.

Another major problem is the text alignment for parallel documents. While parallel document crawling gives satisfactory performance, the precision of the paragraph alignment is currently low, due to the complexities involved in removing unnecessary tags and strings and extracting corresponding textual areas. Improving the paragraph alignment for parallel documents remains an issue for QRpac.

6 Outlook

While QRpac is fully operational in terms of its functions, the above-mentioned technical issues need to be addressed for it to be accepted by online translators who are interested in archiving parallel and comparable documents for their own aims. Once these performance and precision improvements have been made, we will make QRpac publicly accessible without restrictions, and also incorporate it into Minna no Hon'yaku (Translation of/for/by All), an integrated translation-aid platform for online translators in which more than 2,000 users are currently involved in translating online documents[51].

Acknowledgements

This work is supported by the Japan Society for the Promotion of Sciences (JSPS) grant-in-aid (A) 21240021 "Developing an integrated translation-aid site which provides comprehensive reference sources for translators". The short version of this paper appeared as[52].

Notes

1) http://www.lingotek.com/
2) http://kilgray.com/products/memoq/
3) http://www.omegat.org/
4) http://www.globalsight.com/
5) http://www.langtech.co.uk/
6) http://www.proz.com/
7) http://trans-aid.jp/
8) Chen, J. and Nie, J-Y. "Parallel web text

mining for cross-language IR," *Proc. of RIAO 2000*, 2000, p. 62-77.

9) Fukushima, K., Taura, K. and Chikayama, T. "Fast and accurate method for detecting English-Japanese parallel texts," *Proc. of COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, 2006, p. 60-67.

10) Hong, G., Li, C-H., Zhou, M. and Rim, H-C. "An empirical study on web mining of parallel data," *Proc. of 23rd COLING*, 2010, p. 474–482.

11) Li, B. and Liu, J. "Mining Chinese-English parallel corpora from the web," *Proc. of 3rd IJCNLP*, 2008, p. 847-852.

12) Ma, X. and Liberman, M. Y. "BITS: a method for bilingual text search over the web," *MT Summit XII*, 1999.

13) Mohler, M. and Mihalcea, R. "Babylon parallel text builder: gathering parallel texts for low-density languages," *Proc. of 6th LREC*, 2008, 1228-1231.

14) Resnik, P. and Smith, N. A. "The web as a parallel corpus," *Computational Linguistics*, vol. 29, no. 3, 2003, p. 349-380.

15) Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M. and Laurikkala, J. "Focused web crawling in the acquisition of comparable corpora," *Information Retrieval*, vol. 11, no. 5, 2008, p. 427-445.

16) http://www.translationautomation.com/

17) Uszkoreit, J., Ponte, J. M., Popat, A. C. and Dubiner, M. "Large scale parallel document mining for machine translation," *Proc. of 23rd COLING*, 2010, p. 1101–1109.

18) Baroni, M., Kilgarriff, A., Pomikálek, J. and Pychlý, P. "WebBootCaT: Instant domain-specific corpora to support human translators," *Proc. of EAMT*, 2006, p. 247-252.

19) Kageura, K., Abekawa, T. and Sekine, S. "QRselect: a user-driven system for collecting translation document pairs from the web," *Proc. of 10th ICADL*, 2007, p. 131-140.

20) Munday, J. *Introducing Translation Studies*. London, Routledge, 2001.

21) Kageura, K. and Abekawa, T. "NLP meets library science: providing a set of enhanced language reference tools for online translators," *A-LIEP 2009*, 2009.

22) Bernardini, S. and Castagnoli, S. "Corpora for translator education and translation practice," in Yuste, E. ed. *Topics in Language Resources for Translation and Localisation*. John Benjamins, Amsterdam, 2008, p. 39-55.

23) Sharoff, S. "Translation as problem solving: uses of comparable corpora," *Workshop on Language Resources for Translation Research and Practice*, 2006.

24) Zanettin, F. "Bilingual comparable corpora and the training of translators," *Meta*, vol. 43, no. 4, 1998, p. 616-630.

25) The result of the Google Editions lawsuit is indicative of this problem.

26) Kageura, K. and Abekawa, T. "On the concept of 'comprehensiveness' in information services: the case of the online translation aid and hosting service Minna no Hon'yaku," *A-LIEP 2011*, 2011.

27) Kageura, K. and Abekawa, T. "NLP meets library science: providing a set of enhanced language reference tools for online translators," *A-LIEP 2009*, 2009.

28) Kageura, K. and Murayama, R. "QRpac: archiving parallel and comparable documents from the Web," *ICADL 2012*, 2012.

29) Chen, J. and Nie, J-Y. "Parallel web text mining for cross-language IR," *Proc. of RIAO 2000*, 2000, p. 62-77.

30) Fukushima, K., Taura, K. and Chikayama, T. "Fast and accurate method for detecting English-Japanese parallel texts," *Proc. of COLING/ACL Workshop on Multilin-*

gual Language Resources and Interoperability, 2006, p. 60-67.

31) Hong, G., Li, C-H., Zhou, M. and Rim, H-C. "An empirical study on web mining of parallel data," *Proc. of 23rd COLING*, 2010, p. 474–482.

32) Li, B. and Liu, J. "Mining Chinese-English parallel corpora from the web," *Proc. of 3rd IJCNLP*, 2008, p. 847-852.

33) Ma, X. and Liberman, M. Y. "BITS: a method for bilingual text search over the web," *MT Summit XII*, 1999.

34) Mohler, M. and Mihalcea, R. "Babylon parallel text builder: gathering parallel texts for low-density languages," *Proc. of 6th LREC*, 2008, 1228-1231.

35) Resnik, P. and Smith, N. A. "The web as a parallel corpus," *Computational Linguistics*, vol. 29, no. 3, 2003, p. 349-380.

36) Uszkoreit, J., Ponte, J. M., Popat, A. C. and Dubiner, M. "Large scale parallel document mining for machine translation," *Proc. of 23rd COLING*, 2010, p. 1101–1109.

37) Kageura, K. and Abekawa, T. "NLP meets library science: providing a set of enhanced language reference tools for online translators," *A-LIEP 2009*, 2009.

38) Kageura, K., Abekawa, T. and Sekine, S. "QRselect: a user-driven system for collecting translation document pairs from the web," *Proc. of 10th ICADL*, 2007, p. 131-140.

39) The user can specify relevant reserved words at the beginning of the process.

40) This is currently carried out using system-provided dictionaries and the language pairs are limited.

41) Ibid.

42) Kageura, K. and Abekawa, T. "On the concept of 'comprehensiveness' in information services: the case of the online translation aid and hosting service Minna no Hon'yaku," *A-LIEP 2011*, 2011.

43) Baroni, M. and Bernardini, S. "BootCaT: Bootstrapping corpora and terms from the web," *Proc. of 4th LREC*, 2004.

44) Baroni, M., Kilgarriff, A., Pomikálek, J. and Pychlý, P. "WebBootCaT: Instant domain-specific corpora to support human translators," *Proc. of EAMT*, 2006, p. 247-252.

45) Bernardini, S. and Castagnoli, S. "Corpora for translator education and translation practice," in Yuste, E. ed. *Topics in Language Resources for Translation and Localisation*. John Benjamins, Amsterdam, 2008, p. 39-55.

46) Kilgarriff, A., PVS, A. and Pomikálek, J. "BootCatting comparable corpora," *Proc. of 9th TIA*, 2011, p. 123-126.

47) Nakagawa, H. and Mori, T. "Automatic term recognition based on statistics of compound nouns and their components," *Terminology*. vol. 9, no. 2, 2003, p. 201-209.

48) Kageura, K., Abekawa, T. and Sekine, S. "QRselect: a user-driven system for collecting translation document pairs from the web," *Proc. of 10th ICADL*, 2007, p. 131-140.

49) Baroni, M. and Bernardini, S. "BootCaT: Bootstrapping corpora and terms from the web," *Proc. of 4th LREC*, 2004.

50) Nakagawa, H. and Mori, T. "Automatic term recognition based on statistics of compound nouns and their components," *Terminology*. vol. 9, no. 2, 2003, p. 201-209.

51) http://trans-aid.jp/

52) Kageura, K. and Murayama, R. "QRpac: archiving parallel and comparable documents from the Web," *ICADL 2012*, 2012.

オンライン翻訳者による利用を想定したWeb対訳文書及び関連文書のアーカイヴ作成システム

影浦峡 † 村山遼 †

† 東京大学大学院教育学研究科

本研究ノートでは、Web 上で稼働する対訳／関連文書アーカイヴ作成システム QRpac の基本概念と構成について述べる。QRpac は特にオンライン翻訳者を念頭に置いて開発されたシステムで、既往の類似システムが言語研究や機械翻訳のために対訳データをコーパスとして収集するのに対し、明示的に翻訳者の個人アーカイヴ構築を目的としている点に特徴がある。現在、システムは、実効速度の問題から実利用には難があるものの、完全に作動する。本ノートでは、システムの構成だけでなく、背景にある概念を重点的に説明する。

キーワード：オンライン翻訳者，対訳アーカイヴ，関連文書アーカイヴ，専門語彙資源