Using seed terms for crawling bilingual terminology lists on the Web

Takeshi Abekawa[†] **and Kyo Kageura**[‡]

[†] Center for Informatics of Association, National Institute of Informatics, Japan
abekawa@nii.ac.jp

[‡] Graduate School of Education, University of Tokyo, Japan
kyo@p.u-tokyo.ac.jp

1 Introduction

This paper examines the potential of a method for harvesting manually made bilingual term pairs and bilingual terminologies available on the Web. The need for up-to-date multilingual terminological reference resources is very high, but it is generally recognised that manually compiled terminological dictionaries cannot keep pace with the speed of terminological growth. To bridge this gap, much effort has been devoted to developing automatic methods of extracting bilingual terminologies from parallel or comparable corpora [1]. The use of comparable corpora is widely held to be especially important, because they are available in a wider range of languages and text types than parallel corpora.

However, human language practitioners, including online translators (by "online translators" we refer to translators working online, and mainly involved in translating online documents), do not make much use of terminological resources constructed from corpora using the automatic methods. Rather, online translators and other language practitioners tend to use a different approach. When they cannot find relevant entries in online and/or off-line dictionaries and terminologies, many of them generate candidate translations and validate these translation pairs using Google search, assuming that many translation pairs co-occur in online texts [2]. In general, it is recognised that online translators depend heavily on Google search [3].

Taking into account this observation, we developed a system, QRpotato, that collects bilingual term pairs directly from the Web, using seed bilingual term pairs, rather than using corpora as an intermediate resource from which term pairs are extracted [4]. Although an overall evaluation we made in terms of the numbers of Japanese and English bilingual terms collected for a given number of seed terms showed that the system is highly effective in collecting term pairs on a large scale, the evaluation did not examine more detailed aspects of system performance or the degree to which bilingual terms and terminologies exist on the Web. For instance, the availability of bilingual terms is likely to depend on the domain. In order to clarify the performance of the system in a real-world setting, we are currently carrying out experiments to validate actual distribution patterns of bilingual terms on the Web for selected domains, and also checking the effectiveness of seed term pairs from the point of view of their their type of origin and their complexity. This paper reports the first results from these experiments.

2 QRpotato

2.1 Basic philosophy

In the field of library and information science, it is widely recognised that reference resources cannot be reduced to the correctness of individual entries [5]. Like libraries, where "book collections themselves are intellectual instruments that transcend even the content that is within them" [6], a terminological dictionary (or, for that matter, dictionaries in general) transcends the content it contains. For terminological dictionaries to be used by language practitioners, they should have due "normativeness" and/or "comprehensiveness" in terms of their stated objective. These constitute "limiting conditions" that enable users to decide what to do both when they find and when they fail to find the entries they are looking for in a dictionary [7].

In the case of the Web, "normativeness" is not satisfied, while some search engines, most typically Google, enjoy exhaustivity at the social (although not at the actual) level, i.e. most people do not turn to other methods of searching the Web even if they cannot find the information they are looking for using Google search. This is the reason why online translators rely heavily on Google. They take care of normativeness themselves, as it is not provided by the Web, i.e. they validate, either consciously or unconsciously, the information they find using Google search. Taking this into account, we can understand some of the reasons why automatically constructed terminological resources are not used by translators: they do not have the clear "limiting conditions" required for reference resources, and quite often they are provided in such a way that users cannot validate the information.

On the basis of this observation, we developed QRpotato, a system that directly (and exhaustively) collects bilingual term pairs from the Web [4]. Let us now turn to a brief description of the system.

2.2 Mechanisms for extracting term pairs from the Web

The mechanism of QRpotato is based on a simple observation: when multiple bilingual term pairs occur in a Web document, they tend to occur in the same pattern. This is especially true for technical terms in such language pairs as Japanese and English, whose character sets are different. It is a convention in academic writing in Japanese, for instance, to show the English equivalent in brackets immediately after a Japanese term.

Based on this observation, the following procedure is used in QRpotato to extract bilingual term pairs from the Web:

1. Input seed term pairs. Users can either input seed terms one by one via the interactive mode interface, or upload a file containing a set of seed terms via the batch mode interface.

Table 1: Some examples of collocation format

LH symbol	Japanese term	connecting symbol	English term	RH symbol
‘、’	シェーグレン症候群	‘(’	Sjogren syndrome	‘)’
‘ ’	ラポール	‘ ’	rapport	‘と’
‘[’	代謝	‘]](’	metabolism	‘)’
‘●’	アンダーカット	‘【’	undercut	‘】’
‘’	アンタゴニスト	‘’	antagonist	‘’
‘’	イベント	‘ (<i>’	event	‘</i>’
‘’	光	‘</td><td>’	light	‘’

2. Collect Web pages that contain the seed term pair by applying phrase search for each seed term pair. Yahoo! api is used for the Web search.

3. Extract the “collocation format” from the Web pages obtained in step 2. Collocation format is the patterns of occurrence of the seed term pair, which consists of (a) the connecting symbol sequence, i.e. the character sequence inserted between the seed term pair, and (b) the left-hand (LH) and right-hand (RH) terminating symbols that indicate the starting point of the left-hand term and the ending point of the right-hand term in the term pairs. For instance, if the system detects the pattern “, JTERM (ETERM)” on the Web page, it extracts the connecting symbol sequence “ (”, the left-hand terminating symbol “, ”, and the right-hand terminating symbol “)”. The system also analyses HTML tags and uses them for defining the patterns for bilingual term pairs. Collocation format is defined page by page. Some examples of the collocation format are given in Table 1.

4. Using the collocation format, detect term pair candidates from the same Web page.

Steps 2 to 4 are repeated for all the seed term pairs.

We carried out a preliminary experiment at the end of 2009, after the prototype system was developed. In the experiment, we used 210,328 Japanese-English bilingual term pairs taken from the List of Scientific Terms [8] as seed term pairs. The results were as follows [4]:

Number of URLs obtained: 1,425,107

Number of HTML pages obtained: 1,327,180

Number of pages with new term pair candidates: 893,103

Token number of new term pair candidates: 6,567,186

Type number of new term pair candidates: 3,486,125

Manual evaluation of precision using 300 samples showed that 216 (72 percent) were correct pairs, and 22 (7 percent) were partially matched pairs. Assuming that approximately 72 percent of the obtained candidates are correct pairs, 2.5 million term pairs constitute one of the largest bilingual Japanese-English terminologies ever constructed. Using the obtained results as seeds and repeating the steps outlined above, QRpotato should be able to exhaustively collect bilingual Japanese-English term pairs existing in the same Web pages.

3 Viewpoints for evaluation

3.1 Issues related to the actual use of QRpotato

It is expected that QRpotato can provide Japanese and English term pairs in a comprehensive way; that is, as far as searching for term pairs cooccuring in Web pages is concerned, users do not need to turn to other resources even if they cannot find the pairs in the terminology produced by QRpotato. Partly because QRpotato proved useful for real-world use, we detected problems in the system in relation to user expectations and also the system effectiveness and performance not in experimental settings but in actual use.

The first issue is related to the fact that there seems to be a high degree of domain dependency with regards to the effectiveness of the system. Two factors are related to this issue:

(a) The availability of term pairs on the Web seems to be domain dependent. Terms of more practically-oriented domains appear to be abundant, while those of more theoretically-oriented domains seem to be less common, as far as Japanese and English term pairs are concerned. But this common observation has not been empirically examined.

(b) The effectiveness of seed terms for collecting term pairs of the same domain may depend on the nature of the terminology of the domain. In a domain that contains many terms that are used generally, seed terms may be effective in collecting a wide range of term pairs but not in collecting term pairs of that domain. This is related to the second issue.

The second issue is related to system effectiveness. Some terms – such as transliterated terms – may be more effective as seeds than others in collecting more domain-specific term pairs, while others may be more effective in collecting a wider range of pairs. Understanding the behaviour of the types of seed terms will be essential, especially if the obtained terms are to be used for detecting new terms in the second and further cycles.

3.2 Evaluation viewpoints

Against this backdrop, we are currently carrying out experiments to evaluate the real-world system performance and behaviour from the two points of view.

The first is the effect of term types on system performance. We focused on the following two aspects:

(a) Types of origin of terms. There are three major types of Japanese technical terms: Terms consisting only of elements of Chinese origin (e.g. 情報検索 [information retrieval]); those consisting only of elements transliterated from foreign languages, mainly English (e.g. エントロピー [entropy]); and those consisting of both Chinese-origined and translaiterated elements (e.g. 巡回セールスマン問題 [travelling salesman problem]). It is generally the case that Chinese-origined terms are used to represent established and core concepts, while transliterated terms are used to represent newer concepts. The mixed terms are used to represent derived concepts. We will focus on the role of Chinese-origined and transliterated terms in this paper.

Table 2: Basic quantities of the terminological dictionaries

		Types of origin			Construction	
	All	T	C	M	Simple	Complex
COM	16259	3469	4943	7847	2141	14118
ECN	9120	1453	5774	1893	1013	8107
LAW	10020	204	6584	3232	2431	7589
PHY	11081	994	6303	3784	2096	8985
PSY	7026	427	4993	1606	2050	4976

(b) Construction of terms in terms of their simple complexity. In general, simple terms, consisting of only one lexical item, represent basic or core conceps of the domain, while complex terms tend to represent concepts derived from the core concepts.

The second is the influence of domains on the effectiveness of the system. We chose five domains, i.e. computer science, physics, economics, law, and psychology. We used existing terminological dictionaries of these five domains as resources from which seed terms were selected [9]. We did not observe domain dependency independently, but in relation to the types of terms.

Table 2 gives the basic quantities of these dictionaries. In Table 2, COM, ECN, LAW, PHY and PSY stand for computer science, economics, law, physics and psychology, respectively. "T", "C" and "M" indicate transliterated terms, Chinese-origined terms and mixed terms, respectively.

The following two points should be noted:

1. Chinese-origined terms contain a small number of terms consisting of original Japanese, as we automatically extracted the terms using types of characters. In the case of technical terms, however, they do not need to be distinguished from Chinese-origined terms [10].
2. The distinction between simple and complex terms was made according to English terms. This is due to the fact that distinguishing simple and complex Japanese terms cannot be carried out as mechanically as in English. Though this can be a point of debate from the theoretical point of view, it is not only convenient but also plausible to rely on English terms as QRpotato is intended for practical use.

4 Experiments and observations

For each of the terminological data of the five domains listed in Table 2, we took three random samples each consisting of 100 terms by sampling without replacement. A total of 15 samples (three samples for each of the five domains) were used as seeds for QRpotato.

4.1 Types of origin as seeds

For the experiment observing the effect of types of origin, we took three random samples of 100 terms for both Chinese-origined terms and transliterated terms in each of the five domains. This makes 30 samples (for each of the five domains, three for Chinese-origined and three for transliterated).

Table 3 shows the basic quantities of the results. As in Table 2, T and C indicate transliterated terms and Chinese-origined terms, respectively. The table shows the token number of candidate term pairs (token), the type number of candidate term pairs (type), the number of htmls obtained (#htmls), the mean (mean) and maximum (max) number of terms contained in a page, and the number of htmls from which no new candidate term pairs were obtained ("0"). The figure in brackets under "max" shows the percentage of the number of terms compared to the total token number of terms, and the figure in brackets under "0" shows the percentage of the number of htmls that contain no new candidate term pairs compared to the total number of htmls. In order to see the overall skewness of the distributions or how concentrated the pages from which candidate term pairs can be extracted are, we also calculated the Gini index, which is widely used as a summary index to show the degree of concentration [11]. The Gini index is defined as:

$$G = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{|f_i - f_j|}{2\bar{f}n^2}.$$

In the present case, n is the total number of htmls, and f_i and f_j represent the token number of terms contained in the page i and j, respectively. $\bar{f}$ is the mean token number of terms contained in a page and empirically given by:

$$\bar{f} = \frac{\sum_{i=1}^{n} f_i}{n}.$$

The row indicated by "mean" shows the mean value of the three samples for each observation point.

4.1.1 Characteristics of the domains

The number of term types obtained shows the general tendencies of the domains. For the seeds consisting of transliterated terms, the order of the domains sorted in descending order by the number of term types is:

PSY $\rightarrow$ PHY $\rightarrow$ ECN $\rightarrow$ COM $\rightarrow$ LAW,

and for the seeds consisting of Chinese-origined terms, the order is:

PHY $\rightarrow$ PSY $\rightarrow$ COM $\rightarrow$ LAW $\rightarrow$ ECN.

Although the status of the number of term types should ultimately be evaluated according to their relevancy and also in terms of the overall size of the terminology of each domain, it is interesting that more terms in the domains of physics and psychology were obtained than in that of computer science, the terms of which one would expect to exist abundantly on the Web.

Table 3: Results from running QRpotato using different types of origin as seeds

			#term pairs			#terms/html			
Domain	Type	Sample	token	type	#htmls	mean	max	0	Gini
COM	T	1	48897	30776	6257	7.81	1089 (2.23)	2307 (36.87)	0.886
		2	53428	34444	6293	8.49	3296 (6.17)	2933 (46.61)	0.906
		3	39451	28009	3633	10.86	1801 (4.57)	1550 (42.66)	0.912
		mean	47259	31076	5394	9.05	2062 (4.36)	2263 (41.95)	0.901
	C	1	37549	31967	1293	29.04	1494 (3.98)	392 (30.32)	0.890
		2	16876	14925	1042	16.20	875 (5.18)	400 (38.39)	0.883
		3	44342	35960	2617	16.94	2598 (5.86)	986 (37.68)	0.914
		mean	32922	27617	1651	20.73	1656 (5.03)	593 (35.92)	0.896
ECN	T	1	33635	24729	4890	6.88	1801 (5.35)	2350 (48.06)	0.917
		2	43207	29914	4383	9.86	1801 (4.17)	1784 (40.70)	0.908
		3	51435	38940	4000	12.86	3526 (6.86)	1751 (43.77)	0.919
		mean	42759	31194	4424	9.87	2376 (5.56)	1962 (44.35)	0.915
	C	1	11956	9575	434	27.55	1171 (9.79)	144 (33.18)	0.886
		2	11232	9747	543	20.69	772 (6.87)	182 (33.52)	0.894
		3	22232	18895	1549	14.35	1053 (4.74)	458 (29.57)	0.874
		mean	15140	12739	842	20.86	999 (6.60)	261 (31.00)	0.885
LAW	T	1	46116	32182	5157	8.94	2151 (4.66)	2221 (43.07)	0.910
		2	41542	30761	3372	12.32	2151 (5.18)	1500 (44.48)	0.917
		3	33161	24454	6199	5.35	1129 (3.40)	2856 (46.07)	0.891
		mean	40273	29132	4909	8.87	1810 (4.49)	2192 (44.65)	0.906
	C	1	20106	19181	1645	12.22	3708 (18.44)	1457 (88.57)	0.982
		2	15913	12842	1224	13.00	1837 (11.54)	683 (55.80)	0.928
		3	38488	29964	2369	16.25	2882 (7.49)	1536 (64.84)	0.946
		mean	24836	20662	1746	13.82	2809 (11.31)	1225 (70.16)	0.952
PHY	T	1	55576	41532	7003	7.94	1270 (2.29)	2948 (42.10)	0.902
		2	53015	41458	5335	9.94	3549 (6.69)	1873 (35.11)	0.895
		3	44042	32482	3986	11.05	3296 (7.48)	1453 (36.45)	0.906
		mean	50878	38491	5441	9.64	2705 (5.32)	2091 (38.43)	0.901
	C	1	76979	62240	3054	25.21	3082 (4.00)	1556 (50.95)	0.931
		2	63724	52889	1965	32.43	3799 (5.96)	777 (39.54)	0.916
		3	106496	85787	2922	36.45	13148 (12.35)	1200 (41.07)	0.943
		mean	82400	66972	2647	31.36	6676 (8.10)	1178 (44.50)	0.930
PSY	T	1	78088	55660	10031	7.78	1813 (2.32)	4864 (48.49)	0.911
		2	67909	49573	10381	6.54	1661 (2.45)	5201 (50.10)	0.907
		3	58803	41863	8472	6.94	2152 (3.66)	3823 (45.13)	0.898
		mean	68267	49032	9628	7.09	1875 (2.75)	4629 (48.08)	0.905
	C	1	31287	25592	1845	16.96	1800 (5.75)	758 (41.08)	0.909
		2	54687	41721	2555	21.40	3207 (5.86)	1035 (40.51)	0.910
		3	44352	37757	2584	17.16	6382 (14.39)	864 (33.44)	0.904
		mean	43442	35023	2328	18.51	3796 (8.74)	886 (38.06)	0.908

4.1.2 Tendencies of transliterated and Chinese-origined seeds

For all the domains except physics, seeds consisting of transliterated terms were more effective in obtaining term pairs. This is especially notable in the case of terms in the domain of economics, in which the number of terms collected using Chinese-origined seeds was two-fifths the number of terms collected using transliterated seeds on average. Physics shows a completely different tendency, with the type number of term pairs collected using transliterated seeds less than three-fifths of the term pairs collected using Chinese-origined seeds. This discussion should ultimately be made referring to the intersections and differences between term pairs obtained using transliterated seeds and Chinese-origined seeds if we are to evaluate the absolute effectiveness of exhaustively collecting term pairs. This will be reported soon.

There is a general tendency for the percentage of terms contained in the html which contains the largest number of terms to be higher for Chinese-origined seeds than for transliterated seeds, while the general degree of concentration as measured by the Gini index differs from domain to domain. In computer science and economics, the value of the Gini index is larger for transliterated seeds than for Chinese-origined seeds. If chosen properly, a smaller number of transliterated seeds will be able to cover many term pairs. In law and physics, and to some extent psychology, careful choice of seeds will be more effective for Chinese-origined seeds than for transliterated seeds.

4.2 Simple and complex terms as seeds

For the experiment observing the effect of simple and complex terms as seeds, we took three random samples consisting only of simple terms and three consisting of complex terms. The experimental setup was the same as in the experiment using different types of origin as seeds. Table 4 shows the basic quantities of the results obtained by running QRpotato using seeds consisting of simple and complex terms. Notations are the same as those in Table 3.

4.2.1 Characteristics of the domains

The number of term types obtained shows the general tendencies of the domains, which is roughly in accordance with what we observed in 4.1.1. For the seeds consisting of simple terms, the order of the domains sorted in descending order by the number of term types is:

$$\text{PHY} \rightarrow \text{COM} \rightarrow \text{PSY} \rightarrow \text{ECN} \rightarrow \text{LAW},$$

and for the seeds consisting of complex terms, the order is:

$$\text{PHY} \rightarrow \text{PSY} \rightarrow \text{COM} \rightarrow \text{ECN} \rightarrow \text{LAW}.$$

Here again, we observe that a greater number of terms in the domains of physics and psychology are obtained than in the domains of economics or law.

4.2.2 Tendencies of simple and complex terms as seeds

The difference between simple terms and complex terms as seeds is clear for all the domains. The number of term pairs collected using complex terms as seeds was much smaller than the

Table 4: Results from running QRpotato using different construction of terms as seeds

			#term pairs			#terms/html			
Domain	Type	Sample	token	type	#htmls	mean	max	0	Gini
COM	S	1	130207	97759	12668	10.28	2210 (1.70)	5340 (42.15)	0.904
		2	109298	77961	11634	9.39	3279 (3.00)	4558 (39.18)	0.880
		3	162227	119217	14157	11.46	1801 (1.11)	5102 (36.04)	0.886
		mean	133911	98312	12820	10.38	2430 (1.81)	5000 (39.00)	0.890
	C	1	5670	5340	2560	2.21	622 (10.97)	2447 (95.59)	0.993
		2	6402	5474	1632	3.92	1129 (17.64)	1537 (94.18)	0.990
		3	7726	7507	1232	6.27	1053 (13.63)	1175 (95.37)	0.989
		mean	6599	6107	1808	4.13	935 (14.17)	1720 (95.13)	0.991
ECN	S	1	101310	75420	8205	12.35	2721 (2.69)	3065 (37.36)	0.901
		2	111181	81333	8788	12.65	2721 (2.45)	3318 (37.76)	0.904
		3	106980	77932	8624	12.40	3217 (3.01)	3020 (35.02)	0.898
		mean	106490	78228	8539	12.47	2886 (2.71)	3134 (36.70)	0.901
	C	1	9155	8115	231	39.63	1627 (17.77)	112 (48.48)	0.910
		2	6856	6252	392	17.49	1207 (17.61)	207 (52.81)	0.914
		3	3662	3393	259	14.14	373 (10.19)	95 (36.68)	0.865
		mean	6558	5920	294	23.75	1069 (16.30)	138 (46.94)	0.896
LAW	S	1	74085	55240	5489	13.50	6384 (8.62)	2167 (39.48)	0.915
		2	89561	68902	4263	21.01	6384 (7.13)	1661 (38.96)	0.918
		3	56945	41734	2970	19.17	2232 (3.92)	1124 (37.85)	0.894
		mean	73530	55292	4241	17.89	5000 (6.80)	1651 (38.93)	0.909
	C	1	1682	1557	67	25.10	331 (19.68)	22 (32.84)	0.839
		2	2038	1901	112	18.20	386 (18.94)	28 (25.00)	0.829
		3	5422	5322	151	35.91	3207 (59.15)	79 (52.32)	0.934
		mean	3047	2927	110	26.40	1308 (42.93)	43 (39.09)	0.867
PHY	S	1	157752	127481	8970	17.59	13148 (8.33)	3466 (38.64)	0.922
		2	167366	135914	7758	21.57	13148 (7.86)	2989 (38.53)	0.928
		3	160567	118758	8098	19.83	6035 (3.76)	3060 (37.79)	0.922
		mean	161895	127384	8275	19.66	10777 (6.66)	3172 (38.33)	0.924
	C	1	35982	33395	285	126.25	13148 (36.54)	144 (50.53)	0.917
		2	28763	24573	491	58.58	1838 (6.39)	237 (48.27)	0.905
		3	28832	24708	537	53.69	2073 (7.19)	318 (59.22)	0.923
		mean	31192	27559	438	79.51	5686 (18.23)	233 (53.20)	0.915
PSY	S	1	118804	96623	7626	15.58	4340 (3.65)	3441 (45.12)	0.921
		2	101251	79272	6396	15.83	5280 (5.21)	2994 (46.81)	0.925
		3	116136	85611	7566	15.35	3207 (2.76)	3006 (39.73)	0.906
		mean	112064	87169	7196	15.59	4276 (3.82)	3147 (43.73)	0.917
	C	1	9550	6549	559	17.08	2722 (28.50)	350 (62.61)	0.946
		2	10292	9475	494	20.83	2077 (20.18)	347 (70.24)	0.956
		3	11481	8445	650	17.66	2490 (21.69)	281 (43.23)	0.935
		mean	10441	8156	568	18.52	2430 (23.27)	326 (57.39)	0.946

number of term pairs collected using simple terms as seeds. In the domain of law, nearly 20 times more terms were collected using simple terms than by using complex terms. In the domains of computer science, economics and psychology, the number of terms collected using complex terms was more than 10 times larger than the number of terms collected using simple terms. Even in the domain of physics, in which the difference was not as great as in the other domains, nearly five times more terms were collected on average using seeds consisting of simple terms than by using seeds consisting of complex terms.

The number of htmls shows that a much smaller number of Web pages can be obtained using complex terms as seeds in the first place, in all five domains. In addition, the percentage of the number of terms in the page which contains the largest number of terms was much higher for the results obtained using seeds consisting of complex terms than for those obtained using seeds consisting of simple terms. The ratio of htmls that contain no new term pairs also showed the same tendency.

As the Gini index for the seeds consisting of complex terms was much higher than for the seeds consisting of simple terms in computer science and in psychology, we can safely conclude that simple terms contribute more evenly and widely to collecting term pairs than complex terms. In the case of economics, law and physics, the Gini index was larger for seeds consisting of simple terms than for seeds consisting of complex terms, but the ratio of htmls that contain no new term pairs and the ratio of the number of terms contained in the html containing the largest number of terms imply that roughly the same tendencies hold for these three domains.

In any case, the big difference between the results obtained using simple and complex terms as seeds indicates that the choice of seeds will greatly affect the performance of the system. In order to incorporate a routine to select seeds for QRpotato in order to improve its performance and/or effectiveness, we need to delve one step deeper into the effect of seeds on the performance of QRpotato. In order to do so, we need to take into account the nature of the candidate term pairs obtained using different types of seeds, which has not been examined here.

5 Conclusions and outlook

In this paper, we reported the performance of the translation term pair crawling system QRpotato from the points of view of (a) the effect of domains and (b) the effect of types of terms (types of origin and construction of terms) on system performance based on preliminary experiments we have carried out. While some characteristics were observed, we have not yet clarified the most effective and efficient use of QRpotato, from either the point of view of system performance especially in terms of crawling time or from the point of view of user expectations. In fact, as mentioned in section 4, many analyses remain to be done. For instance, the relationship between the nature of terms detected by QRpotato and the nature of terms used as seeds, and the relationships among terms detected using different seeds, have not yet been fully explored. From the predictive point of view, it would be really useful if we could identify characteristics of term pairs that are effective in detecting new term pairs.

On the basis of what has been reported here and what is currently being examined, and upon consultation with translators, we will carry out practical evaluations of the system in order to define the most useful way of providing potential users with the system and the term pair candidates obtained using the system. QRpotato and the term pair candidates obtained from the Web using QRpotato will ultimately be provided through the translation-aid platfrom Minna no Hon'yaku (MNH: translation of/by/for all, available at http://trans-aid.jp/) that we are running [12].

From a different point of view, we are interested in determining how effective QRpotato is in crawling the Web for Korean-English and Chinese-English technical term pairs. As both Korean and Chinese use character sets which are different from English, and adopt the same conventions as Japanese in indicating English translations of technical terms in certain types of documents, QRpotato will probably work as effectively as it does when crawling the Web for Japanese-English term pairs. But this needs to be properly evaluated if QRpotato and the Korean-English and Chinese-English term pairs are to be used in real-world situations.

Acknowledgements

This work is supported by the 2011 NII research cooperation project "Automatic construction of practically useful English-Japanese terminological lexica using Web information resources" and 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 resources for translators".

Notes and references

[1] There are a great number of papers on this topic and it is not possible to list them all here. Relevant articles are regularly published in such conferences as COLING, ACL, and LREC and such journals as *Terminology* and *Machine Translation*.

[2] Interview with eight online translators by the second author, carried out in 2009.

[3] Alain Désilet, Louise Brunette, Christiane Melanon and Geneviève Patenaude (2008). "Reliable innovation: A tecchie's travels in the land of translators," *8th AMTA Conference*, pp. 339–345.

[4] Takeshi Abekawa and Kyo Kageura (2009). "QRpotato: A system that exhaustively collects bilingual technical term pairs from the Web," *3rd International Universal Communication Symposium*, pp. 115–119.

[5] Editorial Committee of the Handbook of Library and Information Science, ed. (1999). *Handbook of Library and Information Science* (2nd ed). Tokyo: Maruzen.

[6] Sandstrum, J. (2010) "Saving the Warburg library," http://centeredlibrarian.blogspot.com/2010/09/saving-warburg-library.html. Accessed 10 December 2010.

[7] Kageura, K. and Abekawa, T. (2011) "On the concept of 'comprehensiveness' in information services," *Proceedings of the Fourth Asia Pacific Conference on Library and Information Education and Practice*, pp. 496–505.

[8] List of scientific terms provided by the National Institute of Informatics, Japan (http://sciterm.nii.ac.jp/).

[9] Nichigai Associates (1997) *Dictionary of Computer Terms* (2nd ed). Tokyo: Nichigai.

Kanamori, H., Ara, K. and Moriguchi, S. (eds.) (1986) *Yuhikaku Keizai Jiten.* Tokyo: Yuhikaku.

Ozaki, T. (2003) *A Dictionary of English Legal Terminology.* Tokyo: Jiyu Kokuminsha.

Japanese Ministry of Education (1990) *Japanese Scientific Terms: Physics* (2nd ed). Tokyo: Baifukan.

Japanese Ministry of Education (1986) *Japanese Scientific Terms: Psychology.* Tokyo: Gakujutu-Sinkokai.

[10] Kageura, K. (2002) *The Dynamics of Terminology.* Amsterdam: John Benjamins.

[11] Shiba, S., Watanabe, H. and Ishizuka, T. (1984) *Dictionary of Statistical Terms.* Tokyo: Shin'yosha.

[12] Utiyama, M., Abekawa, T., Sumita, E. and Kageura, K. (2009) "Hosting volunteer translators," *MT Summit XII.*