A Parallel English-Japanese Query Collection for the Evaluation of On-Line Help Systems

Richard F. E. Sutcliffe*, Sadao Kurohashi†

* Department of Computer Science and Information Systems
University of Limerick, Limerick, Ireland
Richard.Sutcliffe@ul.ie
†Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-8501, Japan
kuro@pine.kuee.kyoto-u.ac.jp

Abstract

An experiment concerning the creation of parallel evaluation data for information retrieval is presented. A set of English queries was gathered for the domain of wordprocessing using Lotus Ami Pro. A set of Japanese queries was then created from these. The answers to the queries were elicited from eight respondents comprising four native speakers of each language. We first describe how the queries were created and the answers elicited. We then present analyses of the responses in each language. The results show a lower level of agreeement between respondents than was expected. We discuss a refinement of the elicitation process which is designed to address this problem as well as measuring the integrity of individual respondents.

1. Introduction

Information Retrieval (IR) systems typically seek to provide a user with information relevant to their needs, taken from a large document collection which is stored electronically. Users normally specify their requirements by the use of short natural language queries. Our field of interest is the application of information retrieval techniques to the domain of technical software documentation. We are concerned with a number of different approaches, some involving language engineering technology, some using traditional approaches based on stemmed keywords.

In order to evaluate an IR system, it is desirable to obtain a collection of test queries for each of which the 'correct' answers are already known. The performance of a system can then be measured by inputting the queries in turn and comparing the system's results with those provided in the test collection. As part of an earlier project (Hyland et al., 1996), a large set of English queries was gathered for a particular software manual concerned with Lotus *Ami Pro* (Ami Pro, 1993). The first stage of the current project involved elicitation of answers to these queries.

In recent years there has been increasing interest in multilingual information retrieval (e.g. Hull and Grefenstette, 1996). Such systems can take as input a query in one of several natural languages and produce as output a set of documents each of which could also be in one of several languages. To evaluate such systems, a multilingual test collection is needed. These can take two forms, *parallel* and *comparable*. A parallel test collection is one in which an 'identical' set of queries exists in more than one language. For each language there is a set of documents which is 'identical' to those of the other languages. By contrast a comparable collection such as Sheridan et al. (1996) has similar rather than identical queries and documents.

The aim of this work was to augment our original English test collection in order to create a parallel Japanese collection complete with its own queries and responses. Such a resource can be used to evaluate multilingual IR systems and hence to facilitate research which aims to compare the effectiveness of different indexing and retrieval techniques across languages.

The remainder of the paper first describes the preparation of the original English query collection, the elicitation of correct responses to those queries, the preparation of the Japanese query collection and the eliciation of the Japanese responses. A comparative analysis of the results is then presented. Finally, conclusions and suggestions for further work are given.

2. The Engish Query Collection

2.1. The Ami Pro Manual

The *Ami Pro Users Guide Release 3.0* (Ami Pro, 1993) is an instruction manual for users of the Lotus word processor Ami Pro. It is intended to contain everything which a user needs to know in order to use the software, including both elementary and advanced features. The manual contains 621 pages. These are divided into 32 chapters as well as contents pages, a reading guide, four appendices and an index. This work is concerned with the material in the 32 chapters only. Each chapter is divided into sections and subsections, neither of which are numbered. The difference in level of abstraction between sections and subsections is not large and for this reason it was decided to treat sections and subsections equally when carrying out evaluation. For the purposes of retrieval, therefore, each chapter is divided into *n* regions where each region is either a section text or a subsection text. The title of the section/subsection is included in the region. The length of a region varies from one or two lines up to a few pages. However, most regions are around half a page in length. The assumption of this work is that from the perspective of the user, the region is the smallest unit of text with which we need to deal. Thus the correct answer to any query about Ami Pro is considered to be an ordered list of regions.

2.2 Collection of Queries

Four sets of queries were gathered for use with the

Collection	No. Queries	Min. Length	Max. Length	Avg. Length	Std. Dev.
Designer	80	5	26	13.8	4.5
Hyland	300	4	27	11.5	5.2
Orion	125	1	19	6.0	3.4
Schmidt	67	4	20	10.1	4.1
Total	572	1	27	9.0	5.2

Table 1: Analysis of query length across the four query collections

Collection	Q	S	VP	IVP	PVP	NP	MISC	Total
Designer	95	20	0	0	0	0	10	125
Hyland	70	0	0	0	0	0	10	80
Orion	49	53	68	2	26	73	29	300
Schmidt	65	1	0	0	0	1	0	67
Total	279	74	68	2	26	74	49	572

Table 2: Syntactic categorisation of queries across the four English query collections

Ami Pro manual. These were the *Hyland Queries*, the *Orion Queries*, the *Designer Queries* and the *Schmidt Queries*. The Hyland Queries were gathered by Patrick Hyland from Ami Pro experts based at Lotus Development Ireland in Dublin. Each was asked to write a few queries which reflected questions which they had themselves asked about Ami Pro during their years of using it in the office. The Orion Queries were gathered from an email technical support line operated for Ami Pro users by Lotus in the US. The Designer Queries were written by Hyland himself and were designed to pose questions in a way which would make them difficult for a keyword-based text retrieval system to answer correctly. Finally, the Schmidt Queries were written by Ingrid Schmidt at Heidelberg University. She purposely learned Ami Pro from first principles and wrote down questions reflecting her thoughts whenever she encountered a problem. Table 1 shows the breakdown of queries by collection together with an analysis of their length in words.

To give an idea of the linguistic characteristics of the queries, Table 2 shows their breakdown into the following categories: Question (e.g. Designer-35: 'if i wish to have different footers on different pages of my document how do i achieve this'), Sentence (e.g. Orion-9: 'cc:Mail is unavailable?'), Verb Phrase (e.g. Orion-51 'Conserve disk space?'), Infinitive Verb Phrase (e.g. Orion-146: 'To print the current document in the background?'), Progressive Verb Phrase (e.g. Orion-5: 'Creating help files?'), Noun Phrase (e.g. Schmidt-23 'Only one word or any string of characters including blanks?') and Miscellaneous (e.g. Designer-41: 'if i wish to modify an existing dictionary entry'). The main conclusion to be drawn from Table 2 is that queries can come in many syntactic forms and in fact less than half are actually questions.

2.3 Elicitation of Responses

In order to evaluate a system using the queries, the correct responses need to be known. In order to accomplish this, the following method was used:

- Four independent respondents were chosen from the population of postgraduates and staff in the Computer Science and Information Systems Department at the University of Limerick. Respondents were native speakers of English with a strong command of word processing concepts.
- Each respondent was provided with written instructions, four sets of queries (Designer, Hyland, Orion & Schmidt) in electronic form, and a printed copy of the *Lotus Ami Pro Word Processor for Windows User's Guide Release 3*. The order of queries was different for each respondent.
- Respondents first read the instructions. Any queries they had were then answered before they started work.
- Each respondent was asked to look at each query in turn and to identify up to five sections which were relevant to that query. They were then asked to rank-order the list of sections by relevance, placing the most relevant section first and the least relevant section last.
- Respondents completed the task in between forty and sixty hours.

3. Japanese Query Collection

3.1 The Ami Pro Manual

Corresponding to the English *Ami Pro Users Guide Release 3.0* (Ami Pro, 1993) there is the *Lotus Ami Pro Word Processor for Windows User's Guide Release 3.1J* (Ami Pro, 1994). This manual was provided with the Japanese version of the software. It contains over 600 pages divided into 32 chapters. Once again, each chapter is divided into sections and subsections, neither of which are numbered. Surprisingly, the manual no longer exists in electronic form because Ami Pro has since been succeeded by Word Pro. For this reason an electronic version had to be reconstructed by an arduous combination of Japanese optical character recognition, editing and proof-reading. (At the time of writing, final proof-reading is the only part of the work not completed.)

Range	0	1	2	3	4	5	Tot
English	115 (5)	542 (24)	545 (24)	386 (17)	262 (11)	435 (19)	2,288 (100)
Japanese	683 (30)	607 (27)	489 (21)	280 (12)	124 (5)	105 (5)	2,288 (100)

Table 3: Counts of the number of times a question was assigned a particular number of response sections by a respondent (percentages in brackets). For example, 545 (24) in the column marked 2 means that in the English elicitation as a whole (four respondents taken together) 545 of the 2,288 questions answered by a respondent (i.e. 24%) were assigned two response sections.

Range	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Tot
English	222	196	136	65	26	20	9	7	8	6	3	4	1	0	1	704
Japanese	150	340	109	59	22	8	5	3	3	0	0	0	0	0	0	699

Table 4: Analysis of average section selection frequency by band. For example, 136 in the column marked 2 means that in the English elicitation, each respondant selected on average 136 different sections with a frequency > 1 and <= 2.

Range	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0	>	Tot
English	222	0	0	76	47	73	0	37	44	32	23	0	23	17	11	14	85	704
Japanese	150	88	72	70	67	43	27	24	27	15	16	12	15	15	7	10	41	699

Table 5: More detailed analysis of part of Table 4 with narrower bands

Range	0	10	20	30	40	50	60	70	>	Tot
English	0	4	7	8	4	2	2	5	0	32
Japanese	0	9	9	5	6	1	2	0	0	32

Table 6: Analysis of average chapter selection frequency by band. For example, 9 in the column marked 20 means that in the Japanese elicitation, 9 chapters were cited > 10 and <= 20 times on average per respondent.

3.2 Creation of Queries

There are two possible approaches to the creation of a set of Japanese queries. Firstly, a *comparable* collection of queries could be elicited from Japanese users, following the English query collection method outlined earlier. This approach has the advantage that the Japanese queries are naturally occurring instances of Japanese information needs. On the other hand it has the disadvantage that the queries are only generally comparable to the English ones. This makes comparisons across languages a lot less direct.

The second approach is to translate the original English queries into Japanese, creating a *parallel* collection. This has the disadvantage that the resulting queries are not naturally occurring but on the other hand there is an exact correspondence between each English query and its Japanese counterpart. This makes certain kinds of analysis possible which could not be undertaken using a comparable query collection. In considering comparable vs. parallel approaches, we should also point out that the translation method assumes that the content of the two manuals is identical. This is not quite the case. For example, the Japanese manual necessarily coveres topics particular to Japanese such as the various input methods (e.g. Roma-ji, Kana Keyboard, JIS code etc) as well as the manipulation of the different character encoding schemes (e.g. SJIS, JIS, EUC etc.) The parallel collection as conceived here contains no queries relating to such topics because they are not part of the English manual.

In summary, therefore, translation of the queries into Japanese was carried out by a native speaker who also had a command of word processing terminology and concepts in both languages. The translator was asked to reflect both the style and content of the original queries as closely as possible in the translation. This meant, for example, that if an English query was ungrammatical or elliptical, so was its Japanese counterpart.

3.3 Elicitation of Responses

Elicitation of Japanese responses was carried out in a manner analogous to that for the English collection, as follows:

- Four independent respondents were chosen from the population of postgraduates in the Graduate School of Informatics at Kyoto University. All respondents were native speakers of Japanese with a strong command of word processing concepts.
  Each respondent was provided with written instructions, four sets of Japanese queries (Designer, Hyland, Orion & Schmidt) in electronic form, and a printed copy of the *Lotus Ami Pro Word Processor for Windows User's Guide Release 3.1J*. The order of queries was different for each respondent.
- Respondents first read the instructions. Any queries they had were then answered before they started work.
- Each respondent was asked to look at each query

Frequency (English)	4	3	2	1
Designer	0.66	0.62	1.15	3.21
Hyland	0.94	0.59	0.93	2.79
Orion	0.61	0.75	1.18	3.57
Schmidt	0.84	0.76	1.19	3.48
Average	0.76	0.68	1.11	3.26

Frequency (Japanese)	4	3	2	1
Designer	0.18	0.50	0.78	2.46
Hyland	0.39	0.5	0.84	2.16
Orion	0.19	0.32	0.60	2.54
Schmidt	0.27	0.39	0.93	2.97
Average	0.26	0.43	0.79	2.53

Table 7 (left) **and Table 8** (right)**:** Average number of sections cited with a particular frequency in the English and Japanese data. For example 0.75 in the column marked 3 in Table 7 means that in the Orion collection, an average of 0.75 sections are cited three times by English respondents in answer to a query. Similarly, in Table 8, 0.18 in the column marked 4 means that in the Designer collection, an average of 0.18 sections are cited four times by Japanese respondents.

Respondent	E1	E2	E3	E4
E1		0.42	0.31	0.40
E2			0.33	0.42
E3				0.36
E4				

Respondent	J1	J2	J3	J4
J1		0.16	0.24	0.25
J2			0.17	0.18
J3				0.31
J4				

Table 9 (left) **and Table 10** (right)**:** Coefficients indicating the degree of match between different pairs of respondents. Respondents E1 to E4 are English while respondents J1 to J4 are Japanese. For example, 0.24 in the column marked 'J3' in Table 10 means that the match of the response list given by J1 for a query compared with that given by J3, averaged over all the queries, was 0.24.

in turn and to identify up to five sections which were relevant to that query. They were then asked to rank-order the list of sections by relevance, placing the most relevant section first and the least relevant section last.

4. Results

4.1 Number of Responses Provided

The first analysis to be carried out was a measurement of the number of response sections which were provided by respondents for each query in the collection. Recall that they were asked to provide 'up to five'. Counts of the number of responses actually given are shown in Table 3. As can be seen, respondents were reluctant to provide large numbers of responses. Only in 19% of the English responses and 5% of the Japanese responses were five sections actually returned. The average number given can be computed from Table 3 as 2.6 for English and 1.5 for Japanese.

4.2 Frequency of Section and Chapter Use

The next analysis was a measurement of the overall usage of sections. For each section, a count was made of the number of times it was used by some respondent as the answer to some query. Naturally, some sections are referred to much more frequently than others because they are more useful and we wished to investigate this phenomenon. The results are summarised in Tables 4 and 5. We state the citation rate per respondent in order to facilitate comparison. Rates are divided into bands and the number of sections cited with a frequency within each band is then given.

Several conclusions can be drawn from these tables. Firstly, 222 English sections out of 704 are not cited as answers to any query. Thus 32% of the entire manual is not referred to. The results for Japanese are lower: 150 sections out of 699, i.e. 21%. Secondly, many sections are cited > 0 and <= 2 times. For English 332 are so cited (47%) while for Japanese 449 are so cited (64%). Thirdly, usage of sections declines rapidly after frequency two. The maximum citation frequency for an English section is 14 while for a Japanese section it is 8.

The overall usage of chapters by respondents was also measured (see Table 6). As the table shows, usage peaks within the range > 0 and <= 30 citations per respondent, for both English and Japanese.

4.3 Individual Responses to Queries

In order to measure the level of agreement between respondents in each language a count was made for each query of the number of sections cited five times, four times, three times, twice or once. An average of these five figures was then computed for each query collection. The results are shown in Table 7 (English) and Table 8 (Japanese). The strongest answers to a query are those which were cited independently by the maximum number of respondents. For the English collection, there were on average 0.76 sections cited by all four respondents. The corresponding figure for Japanese was 0.26 which is considerably lower. The number of sections generally rises much more steeply for Japanese than it does for English as the required frequency falls. For example, in English the number of sections rises from 0.76 to 3.26 as the frequency drops from four to one, a factor of 4.29. In the same interval, the number rises from 0.26 to 2.53 in Japanese, a factor of 9.7. Generally, the level of agreement between Japanese respondents appears to be lower than that among English ones.

Next, a direct comparison was made between responses as follows: For each pair of respondents in a given language, a comparison was made of their responses to each question. This was done by comparing the response lists for a question and computing a coefficient comprising the number of common section names divided

by the total number of distinct section names. The average of these values over all the queries was then determined. The results are shown in Tables 9 and 10. Two points should be noted. Firstly, the overall level of agreement is low – an average match of 0.37 for English and 0.22 for Japanese (computed by taking the mean of the values in Tables 9 and 10 respectively). Secondly, the level of agreement in Japanese is lower then for English.

5. Conclusions

We have presented an experiment concerning the creation of evaluation data for information retrieval. A set of English queries was gathered for the domain of wordprocessing using Ami Pro. The answers to these queries were elicited from four respondents. A set of Japanese queries was then produced from the English ones. Answers to the Japanese queries were then elicited from four respondents. A preliminary analysis of the results was carried out.

What does the experiment show? There are a number of conclusions which can be drawn:

- The generation of answers to queries is time-consuming and difficult for respondents. At least six minutes per query were required.
- The average number of different responses per query is low. For English it is 5.81 (summing the bottom row of Table 7) while for Japanese it is 4.01 (summing the bottom row of Table 8). This is partly a characteristic of the domain: On the one hand, queries tend to be fairly specific while on the other hand, the manual deals in an orderly way with each topic before moving on to another one. Such a domain is completely different from the usual paradigm in IR where documents tend to be much more heterogenous, meaning that there may many reasonable answers to a general query.
- The level of agreement found amongst respondents in a given language was lower than expected – less than one response per query was agreed upon by four respondents. Agreement is lower among Japanese respondents than among English respondents. While the task of assigning relevance to a particular section is known to be a complex one (Schamber, 1994) the level of consistency in making relevance judgements is much lower than was reported in TREC-4 (Harman, 1995).

One characteristic of the experiment which might explain some aspects of the results is that there is no way for a respondent to know when they have produced an adequate set of responses (sections) for a query. We simply ask for 'up to' five. In practice, a good deal less than five sections were produced on average per query. To counteract this, a two-stage process might be preferable: (1) elicitation in which a list of candidate sections is produced for each query, and (2) elimination of irrelevant sections from this list, followed by a rank-ordering of all remaining members. Such a procedure would force each respondent to make a decision about each candidate section, whereas at present we rely on the diligence of the individual. A further refinement would be to introduce random distractor sections into the data after stage (1). These would be sections added to the list of candidates for a particular query which had been judged by the experimenters to be completely inappropriate as responses to that query. A check on the integrity of Stage 2 could then be made by seeing for each respondent whether all the distractor sections had been eliminated. If they had not, then we could conclude that the respondent either misunderstood the task or was unable to carry it out.

In general, we would like to attain a higher level of agreement between respondents to be confident that the answers they provide are reliable.

6. References

Ami Pro, 1993. *Lotus Ami Pro Word Processor for Windows User's Guide Release 3*. Atlanta, GA: Lotus Development Corporation, Word Processing Division.

Ami Pro, 1994. *Lotus Ami Pro Word Processor for Windows User's Guide Release 3.1J* (Japanese Version). Cambridge MA: Lotus Development Corporation.

Harman, D.K., 1995. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman (ed.) *The Fourth Text REtrieval Conference (TREC-4)* (pp. 1-23). Gaithersberg, Maryland: National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-236. Available electronically at trec.nist.gov.

Hull, D.A., and Grefenstette, G., 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In H.-P. Frei, D.K. Harman, P. Schaeuble, and R. Wilkinson (eds.), *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval* (pp. 49-57). New York: Association for Computing Machinery.

Hyland, P., Koch, H.-D., Sutcliffe, R.F.E., and Vossen, P., 1996. *Selecting Information from Text (SIFT) Final Report (LRE-62030 Deliverable D61).* Luxembourg, Luxembourg: Commission of the European Communities, DGXIII/E5. Available at www.csis.ul.ie /staff/Richard.Sutcliffe.

Schamber, L., 1994. Relevance and information behaviour. *Annual Review of Information Science and Technology*, 29:3-48.

Sheridan, P., Ballerini, J. P., and Schaeuble, P., 1996. Building a large multilingual test collection from comparable news documents. In G. Grefenstette, A. Smeaton and P. Sheridan (eds.) *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval* (pp. 56-65).