Evaluation of the Keyword Extraction Task

NTCIR Workshop TMREC Group
(Koichi Takeuchi, Masaharu Yoshioka, Teruo Koyama & Kyo Kageura)
Department of Research and Development, NACSIS
{koichi,yoshioka,koyama,kyo}@rd.nacsis.ac.jp

Abstract: In this paper we summarise the evaluation of the keyword extraction task. Four result files were submitted by three different groups. The evaluation was carried out using the keywords given by authors and also the keywords manually extracted from the abstracts.

1 Introduction

The evaluation of the keyword extraction task was carried out by matching the extracted keyword candidates with two lists of keywords, i.e. (1) the list of keywords given by the authors, and (2) the list of keywords extracted manually by the TMREC group. In the following, we call the former "author keywords" and the latter "manual keywords".

In constructing the manual keywords, we used the *Encyclopedia of Artificial Intelligence* (Shapiro, 1987) for reference. In the following, we call manual keywords that are listed in Shapiro (1987) "*EAI* keywords", and those which are not listed in Shapiro (1987) "non-*EAI* keywords".

There were a total of 4 results submitted for the keyword extraction task, 3 of which were based on the non-tagged corpus (henceforth we refer to these three by A, B, and C), and one which was based on the tagged corpus (D). Two of them were submitted by the same group, so the number of groups is 3.

2 Evaluation by Author Keywords

Some author keywords do not appear in the abstracts as is. Table 1 shows the number of keywords and the ratio of keywords which occur in the abstracts. Around 40 % of the keywords assigned by authors do not occur in the actual abstracts.

Table 1: Basic Information about Author Keywords

	Author Keywords
Number	8380
Ratio of Keywords in Abstracts	58.3%

Table 2, on the other hand, shows the ratio of extracted keyword candidates that occur in the abstracts.

22 keyword candidates in file C and 42 in file D do not appear in the abstracts. This is because these candidates are constructed by "tranformation" from the original strings in the abstracts.

Table 2: Basic Quantities of Result Files

	Extracted Keywords			
	A	B	C	D
Total Number	18262	48199	29896	30678
Keywords in Abstracts	100%	100%	99.9%	99.9%

Table 3 shows the recall and precision of the results A, B, C, and D, evaluated against the list of author keywords.

Table 3: Recall and Precision by Author Keywords

	Result Files			
	A	B	C	D
recall (%)	23.3	36.7	25.0	30.5
precision (%)	10.7	6.4	7.0	8.3

A gives the best precision, and B gives the best recall. This is natural because the number of all the extracted keyword candidates in B is 2.6 times greater than that of the keyword candidates in A. C is located between A and B. D, which extracts more keywords than C, gives better performance than C not only in recall but also in precision. This may be due to the fact that D uses the tagged corpus.

3 Evaluation by Manual Keywords

There is no manual keyword that does not appear in the abstracts. Table 4 shows the token number of *EAI* keywords, non-*EAI* keywords, and all the manual keywords (*EAI* + non-*EAI* keywords).

Table 4: Basic Quantities of Manual Keywords

	Number of Keywords
EAI Keywords	10290
non-*EAI* Keywords	26039
All Keywords	36329

For *EAI* keywords and all keywords, we calculated recall and precision.

Tables 5 and 6 show the result with respect to *EAI* keywords and all the manual keywords, respectively.

Table 5: Evaluation by *EAI* Keywords

	Result Files			
	A	B	C	D
recall (%)	18.9	71.9	36.6	38.8
precision (%)	7.2	10.4	8.5	8.8

Table 6: Evaluation by Manual Keywords

	Result Files			
	A	B	C	D
recall (%)	27.6	69.0	39.7	51.3
precision (%)	39.9	37.8	35.1	44.3

The number of all the manual keywords is more than four times greater than that of the *EAI* keywords only. As a result, both recall and precision go up in Table 6 in comparison with Table 5. This shows the difference between recall and precision in IR, which reflects the characteristics of the keyword extraction task.

In the evaluation by *EAI* keywords, file B gives the highest performance, both in recall and in precision. However, in the evaluation by all manual keywords, B still gives the highest recall but the precision becomes relatively low.

If we examine the characteristics of the keyword candidates in each file, it can be observed that file B does not have many phrasal keyword candidates, and the length of keywords (in terms of number of letters) is shorter. In files C and D, there are many phrasal keyword candidates and the average length is much longer. Few *EAI* keywords are phrasal, while there are many phrasal keywords if we observe all the manual keywords. Table 7 shows the average length of keywords for each file. The keyword candidates in file B tend to match keywords from reference sources, while the candidates in files C and D tend to match manual keywords.

Table 7: Average Length of Keywords by Letters

	Files					
	A	B	C	D	*EAI*	manual
Average Length	4.4	3.9	5.7	6.6	3.3	5.2

A, which is located in between B and C & D, performs well in terms of precision with respect to all inclusive manual keywords, but the recall is not high.

4 Conclusions

We have briefly observed the quantitative nature of the result files in terms of precision and recall, using both author keywords and manual keywords. Unlike the IR task, relations between recall and precision are not so straightforward, which might reflect the ambiguity of the status of "keywords" in themselves.

As in the results of the term recognition task, differences in performance seem to be more a reflection of the differeing views of different participants about the definition of the concept "keywords", rather than a matter of good and bad methods. Unlike the term recognition, however, differing views seem to be more concerned with the syntagmatic aspect than the lexicological aspect.

References

Shapiro, S. (1987) *Encyclopedia of Artificial Intelligence.* New York: John Wiley. [Ohsuga, S. (trans. ed.) *Jinko Tinou Daijiten.* Tokyo: Maruzen . 1991]