Error Correction in a Chinese OCR Test Collection

Yuen-Hsien Tseng
Dept. of Library & Information Science,
Fu Jen Catholic University
Taipei, Taiwan, R.O.C. 242
tseng@blue.lins.fju.edu.tw

ABSTRACT

This article proposes a technique for correcting Chinese OCR errors to support retrieval of scanned documents. The technique uses a completely automatic technique (no manually constructed lexicons or confusion resources) to identify both keywords and confusable terms. Improved retrieval effectiveness on a single term query experiment is demonstrated.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval – *clustering*; I.7.5 [**DOCUMENT AND TEXT PROCESSING**]: Document Capture –*Optical character recognition (OCR)*;

General Terms

Algorithms, Documentation, Performance, Languages.

Keywords

Error correction, term clustering, Chinese, confusing pair.

1. INTRODUCTION

Optical Character Recognition (OCR) systems provide a low-cost approach to access retrospective printed documents. However, even the best OCR systems may inevitably produce recognition errors. The techniques that could prevent retrieval effectiveness from being severely degraded by such errors are in great demand.

One might use various retrieval techniques such as n-gram indexing to alleviate the problem of OCR errors [1]. A complementary approach would be to correct the errors before or during the search process. Although OCR systems often have incorporated a word correction process in the final recognition stage, retrieval systems always build a collection-specific word index that might further help correct more possible errors.

Taghva et al had proposed an error-correction method for English [2]. The method first divides index words into *centroids* and *misspellings* based on a dictionary. The misspellings are then clustered around centroids based on similarity measured by approximate string matching. Finally, words in multiple clusters are resolved by their centroids' term frequencies or by a confusion table having information about the errors likely to be made during the OCR process. Comparing to no error correction, the automatic method was able to reduce percentage of missed documents from 2.4% to 1.3% using a Boolean search system.

SIGIR '02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2. CHINESE OCR ERROR CORRECTION

For languages like Chinese, automatic error correction seems much more difficult. Because there is no delimiter between written Chinese words, identifying a correct word automatically is itself a problem, let alone identifying words with characters in error. But if the same error repeats, which often occurs in real-world OCR processes, one can use such information to extract words and their corrupted counterparts as well. Terms like these can then be clustered for correction. Figure 1 shows the proposed algorithm to correct OCR errors for Chinese text.

Step 1:
- Extract maximally repeated patterns as terms from each document, using the algorithm described in [3].

Step 2:
- Collect only those terms with 3 characters or more.
- For each pair of such terms of the same length, compare their constituent characters.
- If two terms only differ in one character, these 2 terms are clustered together with the term having higher document frequency considered as the correct term and the other the erroneous one.
- The two characters that differentiate these 2 terms are reported as a confusing pair and its occurrence is accumulated.

Step 3:
- The confusing pairs are sorted according to their occurrences in descending order.

Figure 1. The OCR error correction algorithm.

By observing that keywords or key-phrases often repeat in texts, an algorithm was proposed to extract maximally repeated strings as keywords without using any human-maintained resources such as dictionaries or corpora [3]. By maximally, it means that either the repeated strings are the longest or have highest term frequency. Surely there are illegal terms that might be extracted as well. But after some simple stop word filtering, the algorithm can achieve more than 84% word accuracy.

In Step 1 one may alternatively use a dictionary to match the document strings to find clean words. Meanwhile approximate string matching can be applied to identify possible corrupted terms. However, the set of Chinese words is open-ended. In a recent experiment [4], 33% keywords extracted by Tseng's algorithm from Chinese news articles are unregistered in a dictionary of 123,226 words. Therefore, although the use of dictionaries may improve word correction capability, we do not use this vocabulary knowledge in this work to see how well we can achieve using only automatic methods.

Next step is to identify correct terms and their corresponding erroneous ones. In printed Chinese documents, each character takes up almost the same width and length. Thus deletion or

insertion errors likely to happen in English OCR texts are almost impossible in Chinese OCR texts. This allows us to compare only those words of the same length to identify the source of errors. Since very short words differing in one character may still be two individual correct words, we compare only those words with 3 characters or more to find one-character error patterns.

Once a confusion character pair is found, we may use a confusion table like that in Taghva's work to determine whether it is a real OCR error or an incorrect pair. We have thus asked the OCR software company from whose software our texts were converted to provide us a confusion table. The table has 4720 confusion pairs. But our experiments showed that many confusion pairs that are not in the table are found by our algorithm, showing that many recognition errors are far beyond the correction capability of the OCR software. Thus we do not rely on the provided confusion table. Instead, we use a simple heuristic rule to determine whether a term is correct or not: the one that has higher document frequency is considered as correct, the other is not.

3. EXPERIMENT

The above algorithm was applied to a Chinese OCR test collection. The collection consists of 8,438 scanned clippings collected from newspaper articles between 1950 and 1976 from China, Taiwan and Hong Kong in a mix of traditional and simplified Chinese. A commercial OCR system was used to produce digital texts, with an estimated character accuracy of 69%. Exhaustive relevance assessment was performed by 3 judges over the entire collection for 30 Chinese query topics (in a format similar to the TREC topics) collected from various journal articles published at about the same time as the news stories.

Since incorrect terms due to OCR errors lead to vocabulary mismatch and in turn may finally lead to document miss during document retrieval, the recall capability of single-term queries is evaluated in a Boolean retrieval environment. To prepare the query terms, we segmented the texts from all the fields of the 30 queries using the "correct" (higher frequency) term in each of the term pairs obtained from the algorithm. This collects 20 single-term queries. Since the source of these queries comes from the test collection's query topics, the choice of the single-term queries should reflect no bias in our testing.

Figure 2 shows two query terms and other information. Column 1 (separated by semicolon) lists the queries: "Sino-Japanese Relations" and "Liberating Taiwan". Column 2 is their document frequencies. Column 3 lists all those terms that differ in only one character with the query terms. Column 4 is these terms' document frequencies. Column 5 and 6 lists the corresponding confusion pairs. Column 7 lists the number of times that the confusion pair occurs in all term pairs. Column 8 is the number of documents that contain the erroneous term but not the query term. Column 9 is the sum of column 8 for the same query term. This sum shows the possible number of documents that the query term will miss if it is used in a Boolean search environment.

In Figure 2, the term associated with the first query in row 1 means "Sino-Russian Relations". The term in row 2 is "Sino-Czech Relations". These two terms are incorrectly clustered. The term in row 3 is correctly clustered with the query term. (Despite its confusing character pair occurs only once in all term pairs.) The first query term retrieve 5 documents. If the 3 corresponding terms are used as queries, more 6 documents return, although only 1 of them is the desired (contributed by the term in row 3). Similarly, the second query retrieves 138 documents, while its erroneous counterparts (in the third column from row 4 to12) together retrieve other 76 desired documents. In total, the 20 single-term queries retrieve 3283 documents. Their counterparts retrieve 572 extra desired documents and 164 undesired ones. This amounts to an improvement in gross micro-average recall by 572/(572+3238)=15.01% and a reduction in micro-average precision by 164/(164+572+3238)= 4.13%. Compared to no error correction at all, these results are encouraging.

1.	中 日 关 系	5	中 苏 关 系	3	日	苏	7	3	6
2.	中 日 关 系	5	中 捷 关 系	2	日	捷	3	2	6
3.	中 日 关 系	5	中 日 关 东	2	系	东	1	1	6
4.	解 放 台 湾	138	解 放 台 湛	6	湾	湛	3	6	76
5.	解 放 台 湾	138	解 放 台 沿	5	湾	沿	2	5	76
6.	解 放 台 湾	138	解 放 台 泻	1	湾	泻	2	1	76
7.	解 放 台 湾	138	解 放 台 海	9	湾	海	2	9	76
8.	解 放 台 湾	138	解 放 台 褐	1	湾	褐	1	1	76
9.	解 放 台 湾	138	解 放 台 渴	28	湾	渴	4	27	76
10.	解 放 台 湾	138	解 放 台 溜	24	湾	溜	3	24	76
11.	解 放 台 湾	138	解 放 台 门	2	湾	门	2	2	76
12.	解 放 台 湾	138	解 放 台 沁	2	湾	沁	1	1	76

Figure 2. Two query terms and their corresponding term pairs.

4. CONCLUSIONS

A fully automatic error correction method is proposed for use in Chinese OCR text retrieval. A direct application of this technique would be term suggestion in an interactive retrieval environment. Another application is to strengthen the recognition capability of OCR techniques. The proposed method does not rely on any human-maintained resources. Thus it can be readily applied to other collections of various domains without efforts. This also means that the effectiveness of the automatic approach can likely be further improved if domain-specific vocabulary knowledge is used. Future work will explore this direction to automatically reduce false drops due to incorrect clustering.

5. ACKNOWLEDGMENTS

This work is support in part by NSC 90-2413-H-030-004-.

6. REFERENCES

[1] Yuen-Hsien Tseng, "Automatic Cataloguing and Searching for Retrospective Data by Use of OCR Text", JASIST, Vol. 52, No. 5, 2001, pp. 378-390.

[2] K. Taghva, J. Borsack, A. Condit, S. Erva, "The Effects of Noisy Data on Text Retrieval," JASIS, Vo.45. No. 1, 1994, pp.50-58.

[3] Yuen-Hsien Tseng, "Content-Based Retrieval for Music Collections," Proceedings of ACM SIGIR '99, Aug. 15-19, Berkeley, U.S.A., 1999, pp.176-182.

[4] Yuen-Hsien Tseng, "Automatic Thesaurus Generation for Chinese Documents", to appear in JASIST, 2003.