OCR Correction and Query Expansion for Retrieval on OCR Data —CLARIT TREC-5 Confusion Track Report

Xiang Tong[1], ChengXiang Zhai[1], Nataša Milić-Frayling[2], and David A. Evans[2]

[1] Laboratory for Computational Linguistics
Carnegie Mellon University
Pittsburgh, PA 15213
{xt22@andrew.cmu.edu, cz25@andrew.cmu.edu}

[2] CLARITECH Corporation
5301 Fifth Ave.
Pittsburgh, PA 15232-2124
{natasa@clarit.com, dae@clarit.com}

1 Introduction

In CLARIT^{TM1} TREC-5 confusion track experiments, we explored two techniques for improving retrieval performance over corrupted data: (1) OCR word error correction to improve OCR text accuracy, and (2) query expansion by adding query term variants found in the corrupted text. The OCR word correction technique is based on statistical word bigram modeling [Tong & Evans 1996]. The variants of a query term are terms similar to the query term, as measured by the edit distance [Wagner 1974]. While the official runs were based on the first approach, in our follow-up experiments we tested the second approach as well.

In this report we first give a brief description of the OCR correction and query expansion techniques, and then discuss the results of our experiments.

2 The Automatic OCR Correction System

The OCR correction technique used in the CLARIT confusion experiments processes the whole text, sentence by sentence. Figure 1 shows the architecture of the system. For a given text, the correction procedure has three main steps:

1. Read a sentence from the OCR text.
2. Retrieve up to M candidates from the lexicon for each detected error that does not match any lexicon entry. Rank the M candidates by their conditional probabilities with regard to the error. Keep only the top N candidates for the next processing step.
3. Use the Viterbi algorithm to get the maximum likelihood word sequence for the sentence.

In the experiments reported in [Tong & Evans 1996], this method corrected up to 50% of errors in the OCR texts, when applied with $M = 2000$ and $N = 10$. In our TREC-5 confusion experiments we used $M = 200$ and $N = 10$ in order to speed up the correction process. We were aware that this parameter setting may result in a performance inferior to the one observed in earlier experiments.

[1] CLARIT is a registered trademark of CLARITECH Corporation

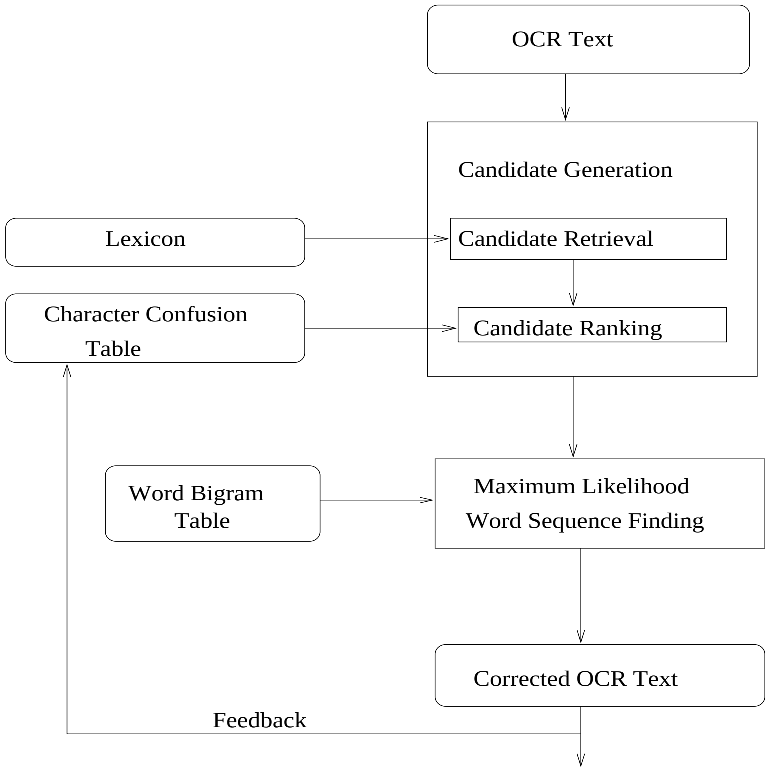

Figure 1: OCR correction system architecture

3 Query Expansion

The original query is expanded by adding variants from the corpus that are similar to the query terms as measured by edit distance [Wagner 1974]. For each query term, at most 10 variants within edit distance of 3 are added. The original query term is assigned a weight of 9 and each of the variants is assigned a weight of $(4 - edit\_distance)$. This method is similar to the query expansion performed by Cornell in TREC 4 [Buckley et al. 1996]. The main difference between the two methods is that the Cornell group expanded queries with all variants with edit distance 1, and used the combined *idf* scores of original and added terms to weight the query terms.

The following is an example of query expansion for Topic 36. There is only one word, "headband", in the original query.

```
Topic 36: headband
Final vector after expansion:
        <t cf="9">headband</>
        <t cf="2">headbands1</>
        <t cf="2">headboard</>
        <t cf="2">headend</>
        <t cf="3">headland</>
        <t cf="2">headlands</>
        <t cf="2">headpond</>
        <t cf="2">headwind</>
        <t cf="2">hedland</>
        <t cf="2">leadand</>
        <t cf="2">neadland</>
```

4 Experiments and Results

The TREC-5 confusion task was named "poorly identified known item search". It involved 49 queries, each uniquely identifying a single document in the FR94 database. The data for searching are provided in three forms: the original uncorrupted data, slightly degraded OCR data with 5% word error (degrade5), and significantly degraded OCR data with 20% word error (degrade20). A participating group is allowed to submit up to 1,000 documents for each topic. The rank of the target item in the retrieved list is recorded. If the known item is not retrieved, a rank of 2000 is assigned. The performance of the system is evaluated using average expected search length and mean reciprocal, which are the mean rank and the mean of the reciprocal of the rank calculated over all the topics respectively.

4.1 Official Runs

In the CLARIT official runs, the two degraded databases were first corrected by our OCR correction module and then indexed using the standard CLARIT indexing module. We used Federal Register data from 1988 and 1989 as the training data to create a lexicon and estimate word uni-gram and bi-gram probabilities. The correction was run on a DecAlpha workstation with 128 MB RAM and it took about 5 days to correct each of the two 250M degraded databases.

The original natural language queries were submitted to the CLARIT retrieval system, which parsed the queries into single terms and phrases.

For comparison purposes we performed the retrieval experiments over all three sets of data: uncorrupted, degrade5, and degrade20 data. Search over the uncorrupted data was performed using the original queries without automatic feedback. In the experiments over degrade5 and degarde20 data, we evaluated system performance for both the initial queries and the enhanced queries from automatic feedback. The results of the simple initial retrieval (CLCON5 and CLCON20) and those with the automatic feedback (CLCON5F and CLCON20F) are in Table 1. Figure 2 shows the number of documents retrieved at different ranks. Table 2 contains the performance based on the results of all the participating groups.

Table 1: Results of all the CLARIT official runs

	Uncorrupted	CLCON5	CLCON5F	CLCON20	CLCON20F
Average Rank	9.39	84.92	66.82	282.16	286.694
Mean Reciprocal	0.73	0.40	0.23	0.21	0.19
Rank 1–10	44	30	24	21	21
Rank 11–100	4	15	21	15	13
Rank 101–1000	1	3	3	8	10
Not Found	0	1	1	5	5

Table 2: Best, worst, and median average rank and mean reciprocal scores across all the groups

	Uncorrupted			Degrade5			Degrade20		
	Best	Med	Worst	Best	Med	Worst	Best	Med	Worst
Average Rank	8.24	89.18	228.49	25.10	116.73	776.55	115.41	280.78	837.51
Mean Reciprocal	.74	.39	.20	.57	.30	.19	.50	.21	.12

Our performance on uncorrupted data is very close to the best systems in terms of both the expected run length and mean reciprocal. The performance over the slightly degraded data (CLCON5) is better than the median.

For both CLCON20 and CLCON20F, the evaluation statistics are worse than the median. We suspected degrade20 text contains many word errors involving space boundaries. Indeed, there are numerous run-on ("of the"/"ofthe") and split-word ("training"/"train ng") errors. The current OCR correction method can not deal with such boundary errors.

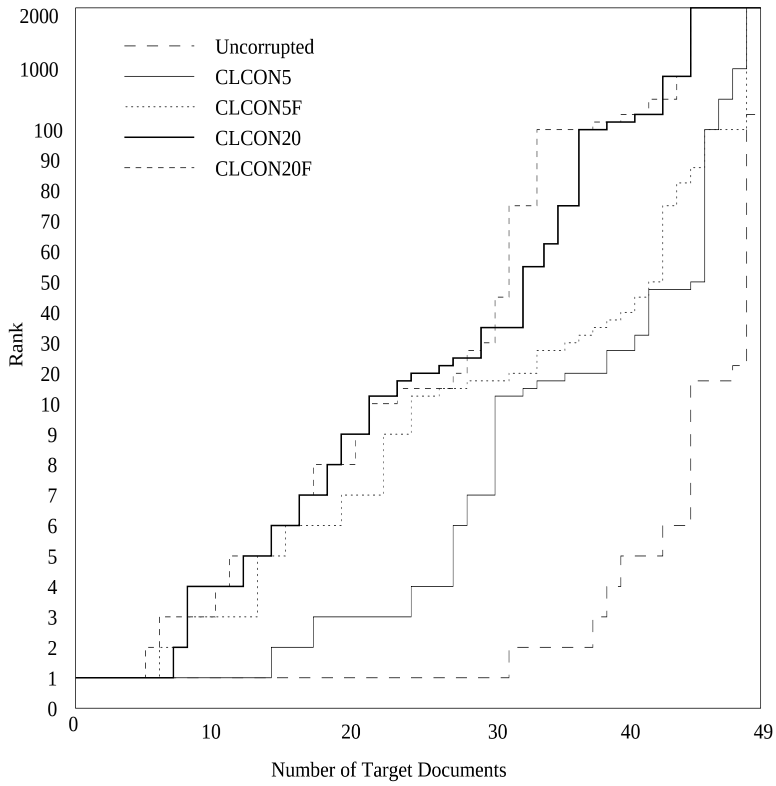

Figure 2: Number of target documents retrieved at a given rank for all the official runs

4.2 Follow-up Experiments on Degrade5 Data

In TREC-5 follow-up experiments, we tested the query expansion technique and compared it with the OCR correction approach. These experiments were run on Degrade5 data. In contrast to the official runs, in these experiments, we did not use phrases to represent queries but only single words. Table 3 contains the retrieval results and Figure 3 shows in detail the number of target documents retrieved at different ranks.

Table 3: Results from follow-up experiments on degrade5 data

Database	uncorrupted	uncorrected		corrected	
Query	unexpanded	unexpanded	expanded	unexpanded	expanded
Ave Rank	5.69	143.4	27.8	80.1	66.0
Mean Rec	0.77	0.618	0.556	0.456	0.376
1–10	46	39	38	32	31
11–100	2	4	10	13	12
101–1000	1	3	1	3	5
Not Found	0	3	0	1	1

The degrade5 baseline run, which uses the unexpanded queries over the uncorrected database, has the best mean reciprocal among all the degrade5 runs. For 35 queries, the target documents are among the top 3 documents in the rank lists. This indicates that CLARIT retrieval system is robust enough to retrieve the relevant documents based on query terms that remain uncorrupted in the target documents. The system is not able to retrieve the target documents for only 3 out of the 49 queries.

Using expanded queries over the original database, the system finds all the target documents with the mean reciprocal 0.556 only a slightly worse than the baseline 0.618 (see Table 3). Furthermore, all but 1 of

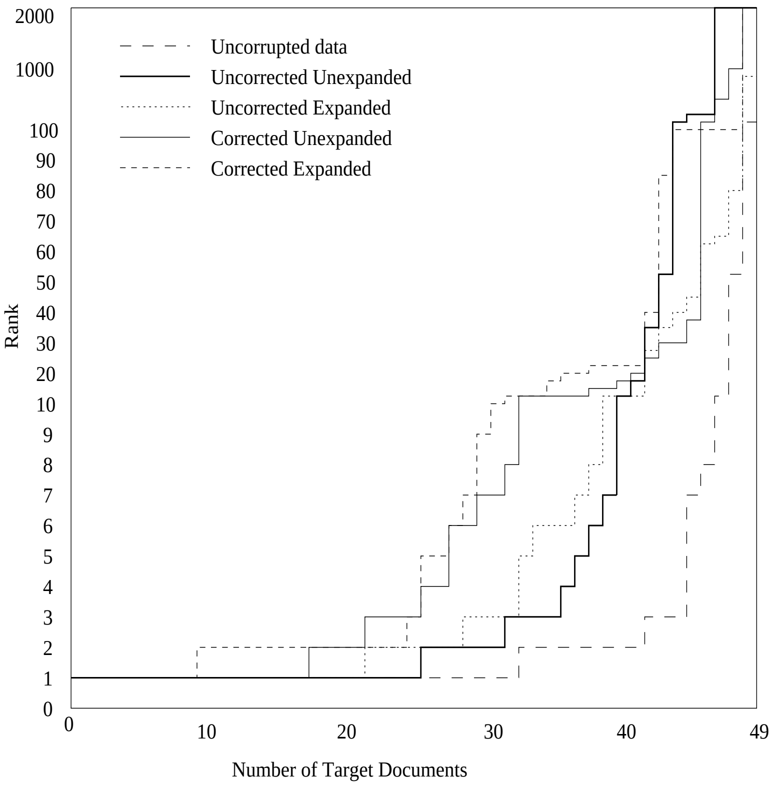

Figure 3: Number of target documents retrieved at a given rank for all the follow-up experiments

these known items are retrieved within rank 100. This shows that our query expansion technique is effective for solving the problem of mismatch between clean queries and corrupted data.

The effectiveness of our OCR correction method is not clear in these experiments, although it is worth noticing that both runs over the corrected degrade5 database miss only one target document each.

5 Conclusions

In TREC-5 confusion experiments, we tested OCR word error correction and query expansion approaches to enhance the retrieval on OCR texts. The corpus-based query expansion technique was quite effective; the OCR correction technique needs to be investigated further.

References

[Buckley et al. 1996] Buckley, C., Singhal, A., Mitra, M., and Salton, G. 1996. New Retrieval Approaches Using SMART: TREC 4. In: D.K Harman, (Ed), *Proceeding of the Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, 1996, 25-48.

[Tong & Evans 1996] Tong, X. and Evans, D. A. 1996. A Statistical Approach to Automatic OCR Error Correction in Context. In: *Proceedings of the Fourth Workshop on Very Large Corpora*, 88-100. Copenhagen, Denmark, 1996.

[Wagner 1974] Wagner, R. A. 1974. The String-to-string Correction Problem. *J. ACM, 21, 1,* Jan. 1974, 168-173.