UAEMex at ImageCLEF 2016:

Handwritten Scanned Document Retrieval Task

Miguel Ángel García Calderón[1], René Arnulfo García Hernández[2], Yulia Ledeneva[3]

Autonomous University of the State of Mexico (UAEMex), Mexico

[1]tonsquemike@outlook.com, [2]renearnulfo@hotmail.com,
[3]yledeneva@yahoo.com

Abstract. This paper describes the participation of the (UAEMex) at the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. We propose to use a skip-character text search method based on Longest Common Subsequence. Our system split all characters in query to find all Longest Common Subsequence in one line of text.

Keywords: Information Retrieval, Longest Common Subsequence, Free Text Search.

1 Introduction

This paper describes the free text search method used by UAEMex at the ImageCLEF 2016 [3] handwritten retrieval task [4]. The 1st edition of the handwritten retrieval challenge has one task targeted in free text search. Considering transcript text for every character we use a skip-character text search method based on Longest Common Subsequence (LCS) problem.

2 Fixed Gap Longest Common Subsequence

The problem to extract LCS consists of given two sequences find the length of longest subsequence present in both of them. Given a string, a subsequence of the string can be obtained from the string by deleting none or some symbols [2] (not necessarily consecutive ones). To extract non-consecutive subsequences, Iliopoulos [1] proposes a variant to find the LCS, called Fixed Gap Longest Common Subsequence (FGLCS) problem, where a value of k is the fixed gap constraint and the distance between two consecutive matches is required to be limited to at most k+1. Figure 1 shows an example of LCS and FGLCS searching.

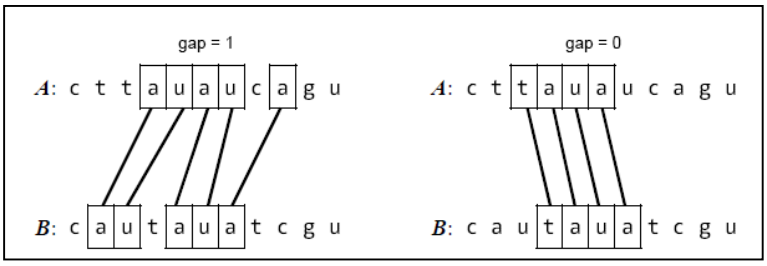

Fig. 1. Example of FGLCS with gap 1 and FGLCS with gap 0.

3 Free Text Search

The proposed method is based on FGLCS search and is divided into three phases. The system is proposed for transcriptions of incomplete or non-existent words.

3.1 Preprocessing Phase

1. Delete non-alphabetic characters in transcript file.
2. Delete line breaks on every segment to get one line segment.
3. Split line by every char.

```
1  no man being bound to subject of arrear as A indi=
2  manner
3  as being in truth of fact or a man A the proportion
4  been , they can a portion no relative to An refund
5  : . a
```

Fig. 2. Example of text segment.

```
1  no man being bound to subject of arrear as A indi= manner as being in truth ...
```

Fig. 3. Example of text segment after line break deletion.

3.2 String Matching Phase

At first step, each query is divided by a space, and then the FGLCS is searched in the actual segment for every word in the query.

3.3 Ranking Phase

Every FGLCS is revised to have the same order of words that in the query, in such case, the confidence score is calculated using equation (1). The system considers that a result is relevant if confidence is more than 0.5.

- q = *chars in query*

- s = *chars in the longest sequence*
- c = *confidence*

$$c = \frac{q-(q-s)}{q} \quad (1)$$

We prove confidence threshold with values 0.9, 0.8, 0.7, 0.6 and 0.5. The best confidence threshold was 0.5.

3.4 Submitted Runs

In this section, the nine free text search runs submitted by UAEMex are presented. Considering bad transcribed words, we change gap value to retrieve more words, however retrieval performance decrease.
Run1: FGLCS search with gap = 0.
Run2: FGLCS search with gap = 1.
Run3: FGLCS search with gap = 2.
Run4: FGLCS search with gap = 3.
Run5: Union of Run1 + Run2.
Run6: Union of Run1 + Run2 + Run3.
Run7: Union of Run1 + Run3.
Run8: Union of Run1 + Run2 + Run3 + Run4.

3.5 Results

In this section, the results of submitted runs by UAEMex are presented. The results with '-' could not be analyzed. Only the measures based on segments are included, and the ones for bounding boxes were omitted. The presented results are extracted only using the *n*-best No.20 of the *n*-best providers by the organizers.
The results of the runs in development the following set of four metrics: Global Average Precision (Segm_gAP), Mean Average Precision (Segm_mAP), Global Normalized Discounted Cumulative Gain (Segm_gNDCG) and Mean Normalized Discounted Cumulative Gain (Segm_mNDCG) have been used to evaluate the accuracy of submissions (see Table 1 and Table2).

Table 1. The results of the development set.

	Segm_gAP	Segm_mAP	Segm_gNDCG	Segm_mNDCG
RUN1	**61.11**	**38.55**	**69.08**	**41.69**
RUN2	47.61	32.33	59.39	37.56
RUN3	30.22	20.32	43.64	27.11
RUN4	-	-	-	-
RUN5	51.21	36.92	64.55	40.70
RUN6	27.62	19.82	53.82	28.90
RUN7	0.15	1.69	1.93	2.82
RUN8	26.24	19.64	53.37	28.81

The results of the runs in test the following set of four metrics: Global Average Precision (Segm_gAP), Mean Average Precision (Segm_mAP), Global Normalized Discounted Cumulative Gain (Segm_gNDCG) and Mean Normalized Discounted Cumulative Gain (Segm_mNDCG) have been used to evaluate the accuracy of submissions (see Table 1 and Table2).

Table 2. The results of the test set.

	Segm_gAP	Segm_mAP	Segm_gNDCG	Segm_mNDCG
RUN1	0.26	0.39	1.22	0.39
RUN2	-	-	-	-
RUN3	-	-	-	-
RUN4	-	-	-	-
RUN5	**3.51**	**0.94**	**10.15**	**1.52**
RUN6	-	-	-	-
RUN7	-	-	-	-
RUN8	-	-	-	-

4 Conclusions

This paper presents results in free text search using LCS. We describe the joint participation of the UAEMex at ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. The proposed method works with words of dictionary and non-existent words. There are big differences between the results of development set (Table 1) and test set (Table 2).
We assume we got bad results because we only use one n-best file provided by the organizers.

References

1. C. S. Iliopoulos and M. S. Rahman, “Algorithms for computing variants of the longest common subsequence problem,” Theoretical Computer Science, vol. 395, pp. 255–267, (2008)
2. H. Lin, M. Lu and J. Fang, "An optimal algorithm for the longest common subsequence problem," Parallel and Distributed Processing, 1991. Proceedings of the Third IEEE Symposium on, Dallas, TX, pp. 630-639 (1991).
   doi: 10.1109/SPDP.1991.21820.
3. Villegas, M., Müller, H., Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science. Springer International Publishing (2016)
4. Villegas, M., Puigcerver, J., Toselli, A.H., Sánchez, J.A., Vidal, E.: Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016).