Submission Instruction for Dry-Run Test of CLIR for NTCIR Workshop 3

National Institute of Informatics (NII), Japan
National Taiwan University (NTU), Taiwan
Chungnam National University (CNU), Korea
NTCIR: http://research.nii.ac.jp/ntcir/
CLIR: http://reseach.nii.ac.jp/workshop/clir/

2001-08/2002-10

1. Introduction

In order to promote the researches on cross-language information retrieval (CLIR), the program committee of NTCIR 3 workshop is happy to announce a CLIR task. There are 3 tracks in CLIR task: 1) Multilingual CLIR, 2) Bilingual CLIR, and 3) Single Language IR (non-English IR). Participants from all over the world are welcome and both automatic systems and interactive systems are welcome. Participants are asked to report their system specification and technique details. Please visit the official web site (http://research.nii.ac.jp/ntcir/) for further information. The application form is also downloadable at the web site (http://research.nii.ac.jp/ntcir/workshop/application3/appweb3-en.html). Basically, 4 languages will involve in the CLIR task: Chinese, English, Japanese, and Korean. In order to go through every step of CLIR task, a dry run will be carried out. Dry run is a preliminary trial of all steps in CLIR. It is a kind of pre-test or practice. It is not mandatory, but we hope you will utilize this opportunity to learn the process of the experiment. Also you can tune your system based on the evaluation results of the dry run, which will be returned to you on November 30, 2001. This document gives an introduction to the instruction of submission for the dry-run test. If everything is ok, the same submission instruction will be used in the formal run.

2. Document Set

The documents used in dry-run test are the same as in formal run. We list the related information of document set in the following again. The participants should receive the document set in CD-ROM. If you did not, please contact with our secretariat (ntc-secretariat@nii.ac.jp).

Japan	Mainichi Newspaper (1998-1999): Japanese	236,664
	Mainichi Daily News (1998-1999): English	12,723
Korea	Korea Economic Daily (1994): Korean	66,146
Taiwan	CIRB010 (1998-1999): Chinese	132,173
	United Daily News (1998-1999): Chinese	249,508
	Taiwan News and Chinatimes English News (1998-1999): English	10,204

The participants have to sign different contracts for using these materials. Each contract has its own requirements. We hope participants could understand the complicated situations of copyright issues in different countries.

3. Dry-Run Topics

There are two sets of topic sets. One is for 1994 documents (Korean Newspaper), and the other is for 1998-99 documents (Chinese, Japanese and English Newspaper). Please make sure you use the correct set of topics. The topics are available in the http://research.nii.ac.jp/ntcir/workshop/clir/Participant/DryRunTopic-en.html. The uid is **ntcir3-clir** and the password is **getdata**. You can use the topics other than you had specified in the application form. Please try as many as you can.

4. Types of Runs

A run is a specific combination of variant techniques in search using different query types against documents. A result file for a run contains top 1000 retrieved documents for each of the topics in a topic set. Basically, we allow all types of runs using any combination of fields in topic and use the 'T' (TITLE), 'D' (DESC), 'N' (NARR), 'C' (CONC) and any combination of these symbols to name the run types. That is to say, participant can submit T run, D run, N run, C run, TD run, TN run, TC run, DN run, DC run, NC run, TDN run, TDC run, TNC run, DNC run, and TDNC run. Each participant can submit **up to 3 runs** for each language pair regardless of the type of run. The language pair means the combination of topic language and document language(s). Among these run types, D run type is mandatory run type. Each participant has to submit at least a D run for a language pair. Each run has to be associated with a RunID. RunID is an identity for each run. The rule of format for RunID is as follows.

Group's ID-Topic Language-Document Language-Run Type-pp

The "pp" is two digits used to represent the priority of the run. It will be used as a parameter for pooling. The participants have to decide the priority for each submitted run in the basis of each language pair. "01" means the high priority. For example, a participating group, LIPS, submits 3 runs for C-->CJ track. The first is a D run, the second is a DN run and the third is a TD run. Therefore, the Run ID for each run is LIPS-C-CJ-D-01, LIPS-C-CJ-DN-02, and LIPS-C-CJ-TD-03, respectively. Also if the group uses different ranking techniques in D run for C --> CJ track, the RunID for each run has to be LIPS-C-CJ-D-01, LIPS-C-CJ-D-02, and LIPS-C-CJ-D-03.

5. Result Format

Since the TREC's evaluation program is used to carry out the relevance assessment, each participating group has to submit its retrieval result in the designated format. The result file is a list of tuples in the following form:

```
030 0    udn_xxx_19980101_0001   0       4238    LIPS-C-CJ-D-01
qid  iter  docno                 rank    sim     runid
```

giving document "docno" (a string extracted from the <DOCNO></DOCNO> field of each document, e.g. <DOCNO>**udn_xxx_19980101_0001**</DOCNO>) retrieved by query "qid" (an integer extracted from the <NUM></NUM> field of topic, e.g., <NUM>**002**</NUM>, the "qid" is 002) with similarity sim (a float). The result file is assumed to be sorted numerically by "qid". "Sim" is assumed to be higher for the documents to be retrieved first. The "iter" and "rank" could be regarded as the dummy filed in tuples. In

addition, each field in tuples is separated by inserting 'TAB' (\x0A, \t) character. The search result for each run is stored in one plain-text file with RunID as it's file name. Please do not send more than 1000 documents for each topic.

6. Technique Description

In addition to search results, every participating group has to give us a concise description of each run. This description should at least contain the following information.

(1) RunID: as explaining in RunID Section.
(2) IndexUnit: character, bi-character, bi-word, phrase, etc.
(3) IndexTech: the techniques used to process index terms, e.g., morphology, stemming, POS, etc.
(4) IndexStruc: inverted file, signature file, PAT, etc.
(5) QueryUnit: character, word, phrase, etc.
(6) QueryMethod: (automatic or interactive)
automatic: runs without any human intervention during query processing and search.
interactive: all runs other than 'automatic'
(7) IRModel: vector space model, probabilistic model, etc.
(8) Ranking: ranking factor for measuring each term, e.g., tf, tf/idf, mutual information, word association, document length, etc.
(9) QueryExpan: techniques used to expand query or no query expansion
(10) TransTech: the translation technique used to deal with cross-language information retrieval, e.g., dictionary-based, corpus-based, MT, etc. The detailed information are welcome, e.g., select-all, select-top-N, WSD.
(11) TrainCorpus: the information about document collections used to construct translation models, and pre- or post-translation query expansion, and so on.

7. How to Submit the Search Results

To: Kuang-hua Chen and Kazuko Kuriyama,
at khchen@ccms.ntu.edu.tw, khchen@hrc.ntu.edu.tw, and kuriyama@nii.ac.jp

By: email as attachment, an email must contain ONLY ONE result file.
Deadline: Nov.12, 2001 23:59 Japanese Time

8. Contact Information

Scientific and technical enquiries including question about dry run process, evaluation, document data set, topics and so on; ntcadm-clir@nii.ac.jp (Executive Committee of CLIR)

Regarding CD-ROM delivery; ntc-secretariat@nii.ac.jp

ML of CLIR Participants; ntc-clir-participant@nii.ac.jp