NTCIR advisor report

Yasushi Ogawa
Software Research Center, RICOH Co., Ltd.
yogawa@src.ricoh.co.jp

Abstract

NTCIR is one of the first projects to construct, using a TREC-like contest operation, a practical Japanese IR test collection. This report discusses design and the administration issues, such as the size and genre of the collection, the description of topics, tasks, the pre-test, and the collaboration with the IREX project.

1 Introduction

Test collections are needed when information retrieval (IR) methods/systems are evaluated, and for English IR there are a variety of standard test collections such as the Cranfield collection, which was built more than 30 years ago. In 1992, NIST (National Institute of Standards and Technology) organized TREC, a contest-type project to evaluate IR systems for evaluating the performance of IR systems under realistic condition — that is, when the size of target documents is measured in gigabytes. TREC has promoted the development of new techniques to handle such large collections.

Even though Japanese IR requires distinctive techniques because Japanese is different from English in many respects, there has been no standard test collection for Japanese IR. Researchers have thus to prepare their own test collections. The first Japanese IR test collection, BMIR, was developed by a working group under the SIG-DBS of the Information Processing Society of Japan and was released in 1996. One distinctive feature of BMIR is that topics are grouped according to the functions necessary to properly process them. Because many papers reporting new techniques evaluated using BMIR have been published, we can say that BMIR has greatly contributed the IR community in Japan. Its size, however, is not satisfactory because BMIR contains only 5080 newspaper articles.

NTCIR is a project to construct a more realistic Japanese test collection. Organized by NACSIS (National Center for Science Information Systems), NTCIR also uses a contest-style operation like TREC, and will encourage Japanese IR studies.

This report summarizes, from viewpoint of a advisory member, various issues related to the design and the administration of the NTCIR project.

2 Design issues

- collection size:

  The number of target documents is about 330,000 and their size is more than 300 MB. This means that researchers engaged in Japanese IR are now able to evaluate their methods/systems using a document set two orders of magnitude larger than BMIR.

  The NTCIR collection, however, is still smaller than the TREC collection, which is over 4GB. Therefore, I expect further efforts to enlarge it.

- genre:

  The target documents are technical abstracts taken from the NACSIS's academic conference papers database. Because the questionnaires filled out by BMIR users indicated that technical papers/abstracts are one of the genres most interesting to IR researchers, it might be welcomed by many researchers.

  Because different kinds of studies would be possible if the collection also contained the bodies (contents) of papers, I expect bodies to eventually be included.

- topic description:

  Topics (search requests) in the formal run consist of the following fields: title, description, narrative, concepts (both in Japanese and English) and a topical field. The amount of detail is richer than that in the topics used in the early part of TREC. Thus, researchers may conduct various kinds of experiments by selecting fields based on their interests.

- various tasks:

  There are, in the NTCIR project, three tasks: ad hoc IR, cross language IR, and term recognition and roll analysis. This means NTCIR covers research fields which cannot be evaluated using the existing BMIR.

  NTCIR, however, does not include tasks which evaluate some important aspects of IR, such as its interactive feature. I therefore expect it to be expanded to include a much wider variety of tasks.

The followings are points to be improved:

- description format:

  Although the target documents and the topics are marked up using SGML tags, some tags — for example those used to separate authors in the authors field — are not in accordance with the SGML standards. It might be better to use the standard SGML tags throughout. In addition, it is desirable to provide DTDs of the documents and the topics.

- imbalance in topical fields:

  The documents cover many topical fields — electronics and computer science, chemistry, architecture, and so on — but, almost half of them came from the electronics and computer science field. This kind of imbalance might be unavoidable simply reflecting the number of societies and the number of members in each, but it might be better to select a better balance of documents. Or, it might be better to introduce subcategories in the electronics and computer science field.

- real topics:

  Some topics look somehow artificial. I believe that query logs are collected at NACSIS as NACSIS provides several retrieval services. Therefore, I expect topics might include those real queries found in the query log.

3 Administration issues

- pre-test:

  Prior to the formal test, a pre-test is conducted. As NTCIR is the first contest-type project in the Japanese IR community and almost none of the participants have any experience in the TREC project, the pre-test might be very useful in helping them to understand the flow of the tasks.

- workshop:

  A workshop is organized at the end of the first cycle of the project. Like a TREC conference, it will give the participants a chance to share their experiences and discuss the advantages and disadvantages of their IR techniques.

- publication of observations during the construction of the collection:

  Various observations obtained while the collection is being constructed are reported and published by the NTCIR organizers. Observations of this kind have seldom been published in Japan, and these publication will help researchers who want to establish new test collections.

The following is a issue needing attention:

- collaboration with IREX:

  IREX, another contest-type project to construct a test collection, is scheduled for almost the same period as NTCIR, and many researchers have joined both. These projects adopted the common TREC format for result submission, but the formats of the target documents and the topics differ between the two projects. These kinds of incompatibilities are troublesome for the participants because they require participants to prepare two similar but different sets of software to handle the two kinds of data formats. The organizers of these projects should, therefore, contact each other more often to avoid such incompatibilities.

4 Conclusion

As more than 30 groups joined the NTCIR project, I could say it was a great success. I expect the test collection established to be published and available to other researchers. I also hope such a project as meaningful as NTCIR will be continued in collaboration with related projects like IREX.

Related Information

TREC: `http://trec.nist.gov`

BMIR: `http://www.ulis.ac.jp:9090/~ishikawa/bmir-j2/eindex.html`

IREX: `http://cs.nyu.edu/cs/projects/proteus/irex`