Preface

The first NTCIR Workshop, held in Tokyo, August 30–September 1, 1999, is the first evaluation workshop designed to enhance research in Japanese text retrieval. Participating groups enrolled for the Workshop by October 31, 1998. From November 1, 1998, the "Test Collection 1 (NTCIR-1), Preliminary version", which contains more than 330,000 documents, was distributed to each participating group. The documents are summaries of conference papers presented at academic or professional conferences hosted by 65 Japanese academic societies, and more than half are English-Japanese paired. Using this Collection, each participating group has conducted various research studies, including the common tasks set by the Workshop organizers. Research using the Collection will be reported and discussed at the Workshop.

The NTCIR Workshop is co-sponsored by the National Center for Science Information Systems (NACSIS) and the Japan Society for the Promotion of Science as part of the "Research for the Future" Program (JSPS-RFTF96P00602). The goals are as follows:

(1) to encourage research in information retrieval, cross-lingual information retrieval and related areas by providing a large-scale Japanese test collection and a common evaluation setting that allows cross-system comparisons
(2) to provide a forum for research groups interested in comparing results and exchanging ideas and opinions in an informal atmosphere
(3) to improve the quality of the Test Collections based on feedback from participants
(4) to investigate methods for constructing a large-scale test collection and corpus including Japanese text and evaluation methods.

With the prosperity of the Internet, the importance of research in Information Retrieval (IR) and related technologies is increasing tremendously. Research and development of IR systems always require solid evidence based on retrieval testing to show the superiority of the proposed system over previous ones. A test collection is a data set used for such retrieval testing.

The importance of large-scale standard test collections in IR research has been widely recognized. Fundamental IR procedures like stopping, stemming and query analysis are language-dependent. In particular, indexing texts written in Japanese or other East Asian languages such as Chinese or Korean, are quite different from those in English, French or other European languages since there are no explicit boundaries (i.e., no spaces) between words in a sentence. For Japanese, the only standard test collection, BMIR-J2, has been used by more than 50 groups and has contributed greatly to IR research in Japan. However, there are still acute needs for enhancement of the test collection in both varieties of text types and scale. The need for cross-lingual retrieval is also acute in the Internet environment. In order to respond to these needs, we aim to construct a large-scale test collection that will also be usable for cross-lingual retrieval and the application of NLP to IR. So far, IR research in Japan has focused very much on segmentation rather than on the retrieval models themselves. We hope that this Workshop will provide an opportunity to go further into research on models as well as segmentation. We have organized this Workshop as a forum for research groups interested in cross-system comparisons and exchanging ideas and opinions.

Thirty-one groups, including participants from six countries, have enrolled. Among them, 28 groups enrolled in IR tasks (twenty-three in the Ad Hoc Task and 16 in the Cross-Lingual Task), and nine in the Term Recognition Task. Ten are from companies and 21 are from universities or national research institutes. Results were submitted by the following groups and we greatly appreciate the efforts and energy of each group towards research using the Collection.

Fuji Xerox Co., Ltd. (Fuji Xerox ITDC)
Fujitsu Laboratories Ltd. (FUJITSU ONE)
Central Research Laboratory, Hitachi, Ltd.
JUSTSYSTEM Corporation

Department of Information Science, Kanagawa University (Hyper Brain Knowledge Infrastructure Group)
Kanagawa University (Goto Lab.)
Korea Advanced Institute of Science and Technology (KAIST/KORTERM)
Manchester Metropolitan University
Multimedia Systems Research Laboratory, Matsushita Electric Industrial Co., Ltd.
Communications Research Laboratory, Ministry of Posts and Telecommunications
NACSIS (The Structured Index Team)
Natural Language Processing Laboratory, National Taiwan University
Human Media Res. Labs., NEC Corp
NEC C&C MEDIA RESEARCH LABORATORIES
NTT Communication Science Laboratories (AIRCS)
RMIT Computer Science & CSIRO Maths and Information Science
Tokyo University of Technology (Kameda Laboratory)
Toshiba R&D Center
Toyohashi University of Technology (Software System Laboratory)
University of California Berkeley (UC Berkeley Text Retrieval Research Group)
University of Library and Information Science
University of Maryland
Department of Information Science & Intelligent Systems, University of Tokushima (Aoe Laboratory)
Graduate School of Engineering, University of Tokyo
Information Technology Center, The University of Tokyo and Yokohama National University (Nakagawa Lab.)
University of Tsukuba
School of Science and Engineering, Waseda University (Shirai Lab.)

We are extremely grateful to the 65 academic societies (http://www.rd.nacsis.ac.jp/~ntcadm/thanks1-en.html) for their understanding of the Project and for providing permission to use document data for the Test Collection, and to the Information Processing Society of Japan and its Special Interest Group on Infology for their co-operation and providing various opportunities of presentation on the NTCIR Project. We would like to thank Dr Donna Harman (NIST, USA), Dr Mun-Kew Leong (KRDL, Singapore), Dr Sung Hyun Myaeng (National Chungnam University, Korea), Mr Yasushi Ogawa (Ricoh, Japan), Prof Stephan E. Robertson (Microsoft, UK), Mr Tsuyoshi Kitani (NTT Data, Japan), Dr Ellen Voorhees (NIST, USA) and Dr Ross Wilkinson (CSIRO and RMIT, Australia) for their substantial advice and helpful comments on the NTCIR Project. Special thanks go to Drs Ellen Voorhees and Donna Harman, NIST, USA, for permitting the use of TREC's evaluation program and scripts to check search result files.

Analysts of IR relevance assessments and those who undertook the hand-tagging of the corpus have all worked very hard. Without their contribution, the Workshop could not have become a reality. Finally, hard work by a dedicated group is necessary for the success of any project. The hard work of many people in NACSIS and elsewhere must be acknowledged. We have been able to list some of these people on the "Workshop Organization" page.

We look forward to the Workshop and do hope that you will enjoy the Workshop.

Noriko Kando
August 1999