Preface

The goal of this workshop is to foster research and development of the technology for patent corpus processing, by providing a forum in which researchers and practitioners can exchange and share their ideas, approaches, perspectives, and experiences from their work in progress.

The processing of intellectual property (IP) documents, including patents, is important in the scientific, business, and law communities. Much of the focus for patent and IP processing has been in the database and information retrieval communities, but not in the computational linguistics (CL) and natural language processing (NLP) communities.

In 2000, the first ACM SIGIR 2000 Workshop on Patent Retrieval was held. In this workshop, patent retrieval systems in use at EPO (European Patent Office) and JAPIO (Japanese Patent Information Organization) were introduced, and a number of issues related to patent retrieval (e.g., producing ontologies, cross-language retrieval, and evaluation methods) were proposed/discussed.

In 2001-2002, the NTCIR workshop (the National Institute of Informatics, Japan), which is a TREC-style evaluation forum for research and development on IR/NLP, first performed the patent retrieval task. Two years of Japanese patents (approximately 7M documents published in 1998-1999; 18GB) were used to evaluate mono/cross-lingual patent retrieval systems. In addition, approximately 17M Japanese/English parallel patent abstracts were used to evaluate the effectiveness of extracting translation lexicons. These experiences stimulate us to further explore this exciting research area.