Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 433–455
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Technical issues of cross-language information retrieval: a review

Kazuaki Kishida *

Faculty of Cultural Information Resources, Surugadai University, 698 Azu, Hanno, Saitama 357-8555, Japan

Received 10 June 2004; accepted 14 June 2004
Available online 23 August 2004

Abstract

This paper reviews state-of-the-art techniques and methods for enhancing effectiveness of cross-language information retrieval (CLIR). The following research issues are covered: (1) matching strategies and translation techniques, (2) methods for solving the problem of translation ambiguity, (3) formal models for CLIR such as application of the language model, (4) the pivot language approach, (5) methods for searching multilingual document collection, (6) techniques for combining multiple language resources, etc.

Keywords: Cross-language information retrieval; Machine translation; Word sense disambiguation; Language model

1. Introduction

Cross-language information retrieval (CLIR) is the circumstance in which a user tries to search a set of documents written in one language for a query in another language. The issues of CLIR have been discussed for several decades. As widely recognized, research efforts for developing CLIR techniques can be traced back to Gerard Salton's articles in the early 1970s (e.g., Salton, 1970).

Especially after the advent of the World Wide Web in the 1990s, CLIR has become more important, allowing users to access information resources written in a variety of languages on the Internet. Since then, the research community of IR has begun to tackle problems of CLIR extensively and intensively. The *Workshop on Cross-Linguistic Information Retrieval* held in August 1996 during the SIGIR'96 Conference is frequently cited as an epochal event for promoting research on CLIR.

* Fax: +81 426 35 9872.
E-mail address: kishida@surugadai.ac.jp

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2004.06.007

Currently, CLIR issues are addressed in workshops of large-scale retrieval experiments such as TREC, CLEF and NTCIR. As described in the introductory paper to this issue, each workshop has been concerned with languages other than English as follows:

TREC: Spanish, Chinese, German, French, Italian, and Arabic.
CLEF: French, German, Italian, Swedish, Spanish, Dutch, Finnish, and Russian so far.
NTCIR: Japanese, Chinese and Korean.

Various research findings on CLIR have been reported at the meetings of TREC, CLEF and NTCIR, and many papers have been published in scientific journals and proceedings.

This article aims at reviewing techniques and methods for enhancing performance of CLIR. We already have a comprehensive review on this topic (Oard & Diekema, 1998). In addition, Peters and Sheridan (2001) cover a wide range of literature and topics on CLIR. The main purpose of this article is to examine literature subsequent to the review by Oard and Diekema and to attempt to organize research results since the mid-1990s in the CLIR field from a technical point of view. For this purpose, some works listed in Oard and Diekema (1998) will be referred to again in this article.

However, it should be noted that this review cannot be completely comprehensive because of the large number of papers on CLIR published in various research areas. The purpose here is to provide a useful map of technical issues of CLIR, rather than extensively enumerating research papers on CLIR. This paper is mainly concerned with "document retrieval," or "text retrieval" issues. For example, CLIR for multimedia data is outside our scope.

The rest of the paper is organized as follows. First, in Section 2, we discuss techniques to match query terms with document representations in the CLIR. More specifically, various methods of translation are described. Section 3 is dedicated to explaining some techniques for solving the problem of term ambiguity, which may occur in the process of translation. Some formal models for CLIR are introduced in Section 4. In particular, we describe the application of the language model (LM), which enables us to combine the retrieval model and the translation model. In Section 5, other important CLIR research topics are discussed: the pivot language approach, search of multilingual document collections, combination of language resources, issues on processing of individual language, user interface for interactive CLIR and evaluation of CLIR. Finally, Section 6 briefly discusses the future direction of CLIR research.

2. Matching strategies and translation

2.1. Matching strategies

2.1.1. Types of matching strategies

The most basic approach to CLIR is to automatically translate the query into an equivalent in the language of the target documents. The translation makes it possible to execute matching operations between the query and each document, and subsequently, compute document scores according to a standard retrieval model such as the vector space or probabilistic model.

However, this is only the starting point. Oard and Diekema (1998) have identified four types of strategies for matching a query with a set of documents in the context of CLIR (Oard & Diekema, 1998, pp. 230–232):

- No translation
  (1) Cognate matching