Cross-Language Information Retrieval

Synthesis Lectures in Human Language Technologies

Editor

Graeme Hirst, University of Toronto

Synthesis Lectures on Human Language Technologies publishes monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is placed on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin, Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen, Michael McTear
2010

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenji Li, Ruifeng Xu, Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Cross-Language Information Retrieval
Jian-Yun Nie

ISBN: 978-3-031-01010-1 paperback

ISBN: 978-3-031-02138-1 ebook

DOI 10.1007/978-3-031-02138-1

A Publication in the Springer series

SYNTHESIS LECTURES IN HUMAN LANGUAGE TECHNOLOGIES

Lecture #8

Series Editor: Graeme Hirst, University of Toronto

Series ISSN

ISSN 1947-4040 print

ISSN 1947-4059 electronic

Cross-Language Information Retrieval

Jian-Yun Nie
University of Montreal

SYNTHESIS LECTURES IN HUMAN LANGUAGE TECHNOLOGIES #8

ABSTRACT

Search for information is no longer exclusively limited within the native language of the user, but is more and more extended to other languages. This gives rise to the problem of cross-language information retrieval (CLIR), whose goal is to find relevant information written in a different language to a query. In addition to the problems of monolingual information retrieval (IR), translation is the key problem in CLIR: one should translate either the query or the documents from a language to another. However, this translation problem is not identical to full-text machine translation (MT): the goal is not to produce a human-readable translation, but a translation suitable for finding relevant documents. Specific translation methods are thus required.

The goal of this book is to provide a comprehensive description of the specific problems arising in CLIR, the solutions proposed in this area, as well as the remaining problems. The book starts with a general description of the monolingual IR and CLIR problems. Different classes of approaches to translation are then presented: approaches using an MT system, dictionary-based translation and approaches based on parallel and comparable corpora. In addition, the typical retrieval effectiveness using different approaches is compared. It will be shown that translation approaches specifically designed for CLIR can rival and outperform high-quality MT systems. Finally, the book offers a look into the future that draws a strong parallel between query expansion in monolingual IR and query translation in CLIR, suggesting that many approaches developed in monolingual IR can be adapted to CLIR.

The book can be used as an introduction to CLIR. Advanced readers can also find more technical details and discussions about the remaining research challenges in the future. It is suitable to new researchers who intend to carry out research on CLIR.

KEYWORDS

cross-language information retrieval; multilingual information retrieval; query translation; document translation; translation model; machine translation / statistical machine translation; dictionary-based translation; parallel corpus; comparable corpus; query expansion; transliteration; mining of translation relations / resources

Dedication

To my dear son Guillaume (子吟).

Contents

Preface

Searching for information is part of our daily life in this information era. Ideally, we are interested in information written in our native language. However, relevant information is not always available in our native language, and we are also interested in finding information written in other languages in many situations. This gives rise to the problem of cross-language information retrieval (CLIR), whose goal is to find relevant information written in a different language to a query. In addition to the problems of monoligual Information Retrieval (IR), translation is the key problem in CLIR. The goal of this book is to provide a comprehensive description of the specific problems that have arisen in CLIR, the solutions proposed in this area, as well as the remaining problems.

The book is organized into the following chapters:

Chapter 1 contains a general description of the IR and CLIR problems. We first provide a description of general IR problems and the approaches proposed in monolingual IR. This description provides the necessary background knowledge on IR for readers who are not familiar with IR. Specific problems to CLIR are then introduced. We will discuss the general strategies that we can use to solve these problems. Readers familiar with IR and CLIR problems can skip this chapter or some sections of this chapter.

Chapter 2 focuses on a family of approaches based on manually constructed translation resources and tools. Namely, we will describe the general approaches to machine translation (MT) as well as their suitability to CLIR. Approaches based on bilingual dictionaries will be presented as possible alternatives.

In Chapter 3, we describe approaches exploiting parallel and comparable texts. We also describe attempts to mine translation resources automatically from the Web.

Chapter 4 describes some approaches to further improve CLIR effectiveness.

Finally, in Chapter 5, we provide a view of CLIR for future developments based on the parallel between query expansion in monolingual IR and query translation in CLIR and propose that query translation can inspire much from query expansion. An example is given to illustrate it.

Acknowledgement

A large part of this book is based on the author's work with his graduate students and collaborators. Without the contributions from the students and collaborators, it would be impossible for him to write this book. The author would like to thank his students Jing Bai, Guihong Cao, Jiang Chen, Lixin Shi, and Michel Simard and his collaborators, in particular, Jianfeng Gao and Wessel Kraaij, for their great work.