

Cross-Language Information Retrieval

Jian-Yun Nie

(University of Montreal)

San Rafael, CA: Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 8), 2010, xv+125 pp; paperbound, ISBN 978-1-59829-863-5, \$40.00; ebook, ISBN 978-1-59829-864-3, \$30.00 or by subscription

Reviewed by

Marcello Federico

FBK-irst Trento

Cross-Language Information Retrieval is a compact book introducing a branch of information retrieval that has gained considerable research interest since the dawn of the World Wide Web in the mid 1990s. Information retrieval is generally concerned with the problem of finding documents within a large collection that are relevant to a given input query. Whereas the original formulation of IR assumes that queries and documents are written in the same language, cross-language IR (CLIR) presumes instead that they are written in two different languages. If the collection contains documents in more languages, then we refer to multi-lingual IR (MLIR), which is typically solved with multiple instances of CLIR. Recently, other variations on the theme have been proposed that address non-textual documents, such as image, music, and speech retrieval. An interesting application of CLIR is the retrieval of images that are provided with textual descriptions in any language.

Computational linguistics could be interested in CLIR for several reasons. CLIR is mainly about the optimal integration of machine translation (MT) and IR, and it presents peculiar and difficult translation issues when short queries are involved, which is the most common case. For such problems, interesting approaches have been developed and refined over time, which mainly build on top of core statistical MT techniques (e.g., word alignment models, translation models) and various lexical resources (e.g., WordNet, dictionaries).

In recent years, several books on IR have been published (e.g., Grossman and Frieder 2004; Manning, Raghavan, and Schütze 2008; Büttcher, Clarke, and Cormack 2010), which devoted at most a section or chapter to CLIR. As specific books on CLIR have been limited so far to edited collections of scientific papers (Grefenstette 1998), it was definitely time for the first monograph on the topic.

Jian-Yun Nie's volume is structured as five chapters, which are organized as follows:

- Chapter 1, "Introduction," covers IR problems, approaches, and models, language problems in IR with European and East Asian languages, CLIR problems and approaches, needs for CLIR and MLIR, and a brief history of CLIR.
- Chapter 2, "Using manually constructed translation systems and resources for CLIR," covers an introduction to MT, basic use of MT in CLIR, and dictionary-based translation for CLIR.
- Chapter 3, "Translation based on parallel and comparable corpora," covers methods for automatic paragraph and sentence alignment, use

of translation models for CLIR, alternative approaches using parallel corpora, discussion of CLIR methods and resources, and mining for translation resources and relations.

- Chapter 4, “Other methods to improve CLIR,” covers pre- and post-translation expansion, fuzzy matching, combination of translations, transitive translation, and integration of monolingual and translanguing relations.
- Chapter 5, “A look into the future: Towards a unified view of monolingual IR and CLIR?” summarizes the state-of-the-art in CLIR and proposals for improvements.

CLIR approaches are in general presented together with their statistical models, whose understanding does not require more than elementary calculus and probability theory. However, the book does not present algorithms or data structures to implement the models, so it might not be a sufficient resource to build an effective CLIR system.

The first two chapters are rather introductory and lead to the conventional CLIR approach, in which MT or dictionary-based translation is simply cascaded with monolingual IR. A discussion on the limitations of using such general translation tools convinces the reader of the need for translation techniques that are more specific to and better integrated with IR. In Chapters 3 and 4, the core of the book, several advanced CLIR models from the recent literature are discussed. In particular, Chapter 3 focuses on the collection and processing of parallel texts and on statistical translation models for query terms. Chapter 4 discusses cross-lingual counterparts of well-established IR techniques (i.e., pre- and post-translation query expansion) as well as CLIR-specific methods to further improve retrieval performance (e.g., fuzzy matching and translation combination). Finally, in Chapter 5 the author, starting from a parallel between query expansion in IR and query translation in CLEF, proposes new directions for future work.

In conclusion, the book presents a body of work in CLIR with a uniform level of presentation and a consistent notation. It is definitely a good reference for an introduction to the field as well as for a survey of the state-of-the-art.

References

- Büttcher, Stefan, Charles L.A. Clarke, and Gordon V. Cormack. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, Cambridge, MA.
- Grefenstette, Gregory, editor. 1998. *Cross-Language Information Retrieval*. Springer Verlag, Berlin.
- Grossman, David A. and Ophir Frieder. 2004. *Information Retrieval: Algorithms and Heuristics*, second edition. Springer Verlag, Berlin.
- Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

This book review was edited by Pierre Isabelle.

Marcello Federico is co-director of the Human Language Technology research unit at FBK-irst, Trento (Italy). He has contributed about 100 scientific papers in the fields of automatic speech recognition, statistical language modeling, cross-language information retrieval, and statistical machine translation. e-mail: federico@fbk.eu.

Information Retrieval Jian-Yun Nie, Jean-Francois Dufort Dept. IRO, University of Montreal, CP. 6128, succursale Centre-ville Montreal, Quebec, H3C 3J7 Canada nie@iro.umontreal.ca. Our experiments on cross-language information retrieval (CLIR) show that such a translation model leads to a much better CLIR effectiveness when compound terms are integrated. Keywords: information retrieval, compound term, translation model, cross-language information retrieval, query translation. 1. Introduction Most information retrieval (IR) systems currently in use are based on simple words, which are used as indexes for documents and queries. The estimation of a document's relevance to a query is based on a sharing of keywords between them. Jian-Yun Nie. university of montreal. Verified email at iro.umontreal.ca. information retrieval natural language processing. Articles Cited by Co-authors. Title. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. JY Nie, M Simard, P Isabelle, R Durand. Proceedings of the 22nd annual international ACM SIGIR conference on , 1999. Jian-Yun Nie. Home. Contact. Interesting Links. Information Retrieval. Organizations. SIGIR. Association for Information Science and Technology. Chinese and Oriental Languages Information Processing Society. Journals. The Public-Access Computer Systems Review. The Information Society. Resources. A book written by C.J. van Rijsbergen Information Retrieval. More resources for IR: corpus and tools for information retrieval. My laboratory and department. RALI lab. Livre Cross-Language Information Retrieval. © 2017 Jian-Yun Nie. Home. Jian-Yun Nie, Jing Bai. Département d'Informatique et de Recherche Opérationnelle, Université de Montréal. Query Translation, Query Expansion, Cross-Language Information Retrieval, Markov Chain, Random Walk. 1. INTRODUCTION. Cross-Language Information Retrieval (CLIR) has attracted a large number of studies, and a variety of methods for query translation have been proposed [1, 5, 10, 34, 16, 30, 31]. Many of these methods rely on dictionaries for query translation due to the simplicity of the methods and the increasing availability of machine readable bilingual dictionaries [10, 11, 14, 15, 33]. Improving query translation for cross language information retrieval using statistical models. In Proceedings of ACM SIGIR, pp. 96-104. [12] Grefenstette, G. (1999).