Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Information Processing and Management 41 (2005) 549–568

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Noun phrases as building blocks for cross-language Search Assistance

Fernando López-Ostenero [*], Julio Gonzalo, Felisa Verdejo

Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), E.T.S. Ingeniería Informática, C/ Juan del Rosal 16, 28040 Madrid, Spain

Received 10 June 2004; accepted 14 June 2004
Available online 12 September 2004

Abstract

This paper presents a *Foreign-Language Search Assistant* that uses noun phrases as fundamental units for document translation and query formulation, translation and refinement. The system (a) supports the foreign-language document selection task providing a cross-language indicative summary based on noun phrase translations, and (b) supports query formulation and refinement using the information displayed in the cross-language document summaries. Our results challenge two implicit assumptions in most of cross-language Information Retrieval research: first, that once documents in the target language are found, Machine Translation is the optimal way of informing the user about their contents; and second, that in an interactive setting the optimal way of formulating and refining the query is helping the user to choose appropriate translations for the query terms.

Keywords: Cross-language information retrieval; Interactive information retrieval; Natural language information retrieval

1. Introduction

Over the last 7 years there has been a great improvement in the techniques to retrieve relevant documents in languages different from the query language. State-of-the-art systems routinely perform above 75% of

[*] Corresponding author.
E-mail addresses: flopez@lsi.uned.es (F. López-Ostenero), julio@lsi.uned.es (J. Gonzalo), felisa@lsi.uned.es (F. Verdejo).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.06.014

the equivalent monolingual retrieval, and occasionally match monolingual performance (Gey, Kando, & Peters, 2002).

But this is just one aspect of the cross-language *Information Access* problem. Take the case of a Spanish journalist that needs to know the local opinion in Japan about a certain event. Let us suppose that the journalist can use a system able to accept queries in Spanish and find documents in Japanese. Let us assume that he is willing to pay for a high-quality manual translation of the documents which are of primary interest to his research. From an initial query in Spanish, the system retrieves a ranked list of documents written in Japanese. How can the user distinguish which ones are really relevant before paying for manual translations? How can he decide whether to stop searching or refine the query? How can he use the information in the retrieved documents to refine the query?

A reason why these problems have rarely been studied from a multingual perspective lies in the implicit assumptions that (a) commercial Machine Translation (MT) systems can be used to translate the documents into the user's native language; and that (b) cross-language document selection and query refinement can be done using such translations.

While there have been some experiments on the use of document translations for cross-language relevance judgment (see Section 6), the assumptions above are still far from being verified experimentally, and there are in fact reasons to question them: first, machine translations are far from perfect, and usually hard to read. Second, it is not evident how the information provided by machine translations can be used to modify and improve the query until the information need is satisfied. Third, machine translation is costly (compared to document retrieval) and may introduce significant delays in a search session.

In this paper, we propose an approach to *cross-language Search Assistance* (as an interactive task, broader than cross-language document retrieval) based on noun phrases as fundamental units for translation and query formulation. This approach consists of:

- An algorithm to align short noun phrases between two languages using only bilingual dictionaries and comparable corpora.
- A system that produces cross-language indicative summaries using a greedy algorithm to translate noun phrases (of any size) using only the previous alignment information and corpus frequencies. These summaries support cross-language document selection.
- An interactive system that supports query formulation and refinement by phrases, where phrases are translated without user intervention.

The use of noun phrases as building blocks for Foreign-Language Search Assistance is essentially novel, but there is evidence that supports the approach:

- While words are optimal indexing units (in non-agglutinative languages), accurate translation demands larger units (Verdejo, Gonzalo, Peñas, López, & Fernández, 2000). (Ballesteros & Croft, 1998) already showed that terms in a phrase could be accurately translated by calculating which combination of candidate term translations occurs most frequently in the target language corpus. Our algorithm to align short noun phrases uses essentially the same approach, although the target language corpus is previously parsed to obtain a target language noun phrase list, rather than a list of individual target language terms for each of the phrase components. In other words, Ballesteros and Croft use source-language phrasal information to translate individual words, while we use source and target phrasal information to map noun phrases between both languages.
- Empirical studies such as (Peñas, Gonzalo, & Verdejo, 2001; Dennis, Bruza, & McArthur, 2002) show that phrases are a natural way of interactively refining queries. They resemble complex searching concepts and they have more semantic content than isolated terms.