Interactive Cross-Language Searching: Phrases Are Better than Terms for Query Formulation and Refinement

Fernando López-Ostenero, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
Edificio Interfacultativo, C/ Juan del Rosal 16. 28040 Madrid, Spain
{flopez,julio,anselmo,felisa}@lsi.uned.es
http://nlp.uned.es/

Abstract. This paper summarizes the participation of the UNED group in the CLEF 2002 Interactive Track. We focused on interactive query formulation and refinement, comparing two approaches: a) a reference system that assists the user to provide adequate translations for terms in the query; and b) a proposed system that assists the user to formulate the query as a set of relevant phrases, and to select promising phrases in the documents to enhance the query. All collected evidence indicates that the phrase-based approach is preferable: the official $F_{a=0.8}$ measure is 65% better for the proposed system, and all users in our experiment preferred the phrase-based system as a simpler and faster way of searching.

1 Introduction

In our second participation in the CLEF Interactive track, we have focused on assisted query formulation and refinement for cross-language searching. Previous experiments in this area have mainly concentrated on assisted query translation [4, 1]: for every term in the query, the system displays its possible translations with some information about the meaning of each candidate (via definitions in the source language or inverse translations). The user then selects the most appropriate translations, interacting with the system to overcome translation ambiguity.

Our hypothesis to be tested in iCLEF 2002 was twofold:

- Examining translations in an (unknown) foreign language is a high-load cognitive task, and therefore it is worth exploring alternative ways of assisting cross-language query formulations.
- Selecting relevant phrases for a topic should be easier and faster than selecting translations, and without human intervention phrases can be translated more accurately than individual terms.

In order to test our hypothesis, we have used:
An Interactive Cross-Language system that helps users to provide accurate translations for query terms, based on the information provided by a reverse dictionary. This system is used as the reference one.

An Interactive Cross-Language system that helps users to select appropriate phrases to describe their user needs, and translates phrases in a completely automatic way.

A common document translation strategy for both systems based on our previous iCLEF findings [3].

In Section 2, we describe our experiment design. In Section 3, we discuss the outcome of the experiment, and in Section 4 we provide some conclusions.

2 Experiment Design

Our experiment consists of:

- Eight native Spanish speakers with null or very low English skills.
- The Spanish version of the four official iCLEF topics.
- A reference interactive cross-language search system based on assisted term translation (System WORDS).
- A proposed system based on noun-phrase selections (System PHRASES).
- The official iCLEF latin square to combine topics, searchers and systems into 32 different searching sessions.
- The official iCLEF search procedure.

In this section we describe the most relevant aspects of the above items.

2.1 Reference System

The reference system (WORDS) uses assisted query term translation and refinement throughout the search process:

- **Initial Query Formulation.** The system translates all content words in the Spanish iCLEF topic using a bilingual dictionary, and displays possible English translations to the user. When the user points to an English term, the system displays inverse translations into Spanish. This information can be used by the searcher to decide which translations to keep and which translations to discard before performing the first search. Figure 1 illustrates this initial step.

- **Cross-Language Search.** The system performs a monolingual search of the LA Times collection with the English terms selected by the user.

- **Ranked Document List.** The ranked list of documents displays the (translated) title of the document and a colour code to indicate whether each document has already been marked as relevant, not relevant or unsure. Figure 2A shows a retrieved ranked list.