

# Using N-gram Models to Combine Query Translations in Cross-Language Question Answering

Rita M. Aceves-Pérez, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

Language Technologies Group, Computer Science Department,  
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.  
{rmaceves, mmontesg, villasen}@inaoep.mx

**Abstract.** This paper presents a method for cross-language question answering. The method combines multiple query translations in order to improve the answering precision. The combination of translations is based on their pertinence to the target document collection rather than on their grammatical correctness. The pertinence is measured by the translation perplexity with respect to the collection language model. Experimental evaluation on question answering demonstrates that the proposed approach outperforms the results obtained by the best translation machine.

## 1 Introduction

A question answering (QA) system is a particular kind of search engine that allows users to ask questions using natural language instead of an artificial query language. In a cross-lingual scenario the questions are formulated in a language different from the document collection. In this case, the efficiency of the QA system greatly depends on the way it confronts the idiomatic barrier. Traditional approaches for cross-lingual information access involve translating either the documents into the expected query language or the questions into the document language. The first approach is not always practical, in particular when the document collection is very large. The second approach is more common. However, because of the small size of questions in QA, the machine translation methods do not have enough context information, and tend to produce unsatisfactory question translations.

A bad question translation generates a cascade error through all phases of the QA process. This effect is evident in the results of cross-lingual QA reported on the last edition of CLEF [4]. For instance, the results from the best cross-lingual system (that uses the French as target language) were 64% of precision for the monolingual task, and 39.5% when using English as question language. In this case, the errors in the translation of the question cause a drop in precision of 61.7%.

Recent methods for cross-lingual information access attempt to minimize the error introduced by the translation machines. In particular, the idea of combining the capacities of several translation machines has been successfully used in cross-lingual information retrieval [2]. In this field, most works focus on the selection of the best

---

\* This work was partially financed by the CONACYT (grants 43990 and 184663). We also like to thanks to the CLEF for the provided resources.