



A SURVEY OF MULTILINGUAL TEXT RETRIEVAL

By Douglas W.Oard and Bonnie J.Dorr

**M.Berkan Ercan
Enes Taylan**

INTRODUCTION

- A survey paper, 30 pages long
- April, 1996
- Multilingual
 - Retrieving text in one language, based on queries on another language.
- Even in 1996, many online resources were available.
 - Monolingual systems were satisfying the demands.



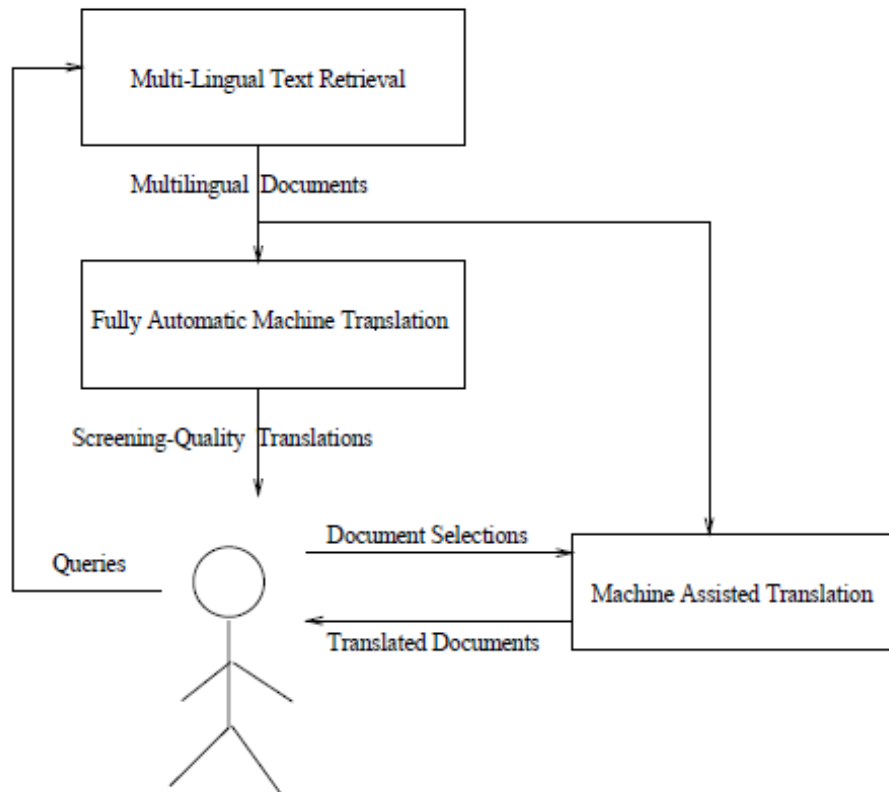
MOTIVATION

- When the documents are desired in another language,
 - It is expected from the user to translate the query.
- However, there are some needs which can not be satisfied by monolingual systems.
 - A collection can contain documents from many different languages. It is impractical to write queries for each.
 - Handling the documents which are expressed with many languages.
- The user's inability to express a query in each language.



MOTIVATION (CONT)

- With Machine Translation, a successful system can be established.



A GOOD DEFINITION OF TEXT RETRIEVAL

- The goal;
 - To present the user a set of items that will satisfy her information need.
 - The approach have to basic components;
 - Representing documents
 - Computing comparison between them
- This process is successful when it produces similar to those produced by human comparison.



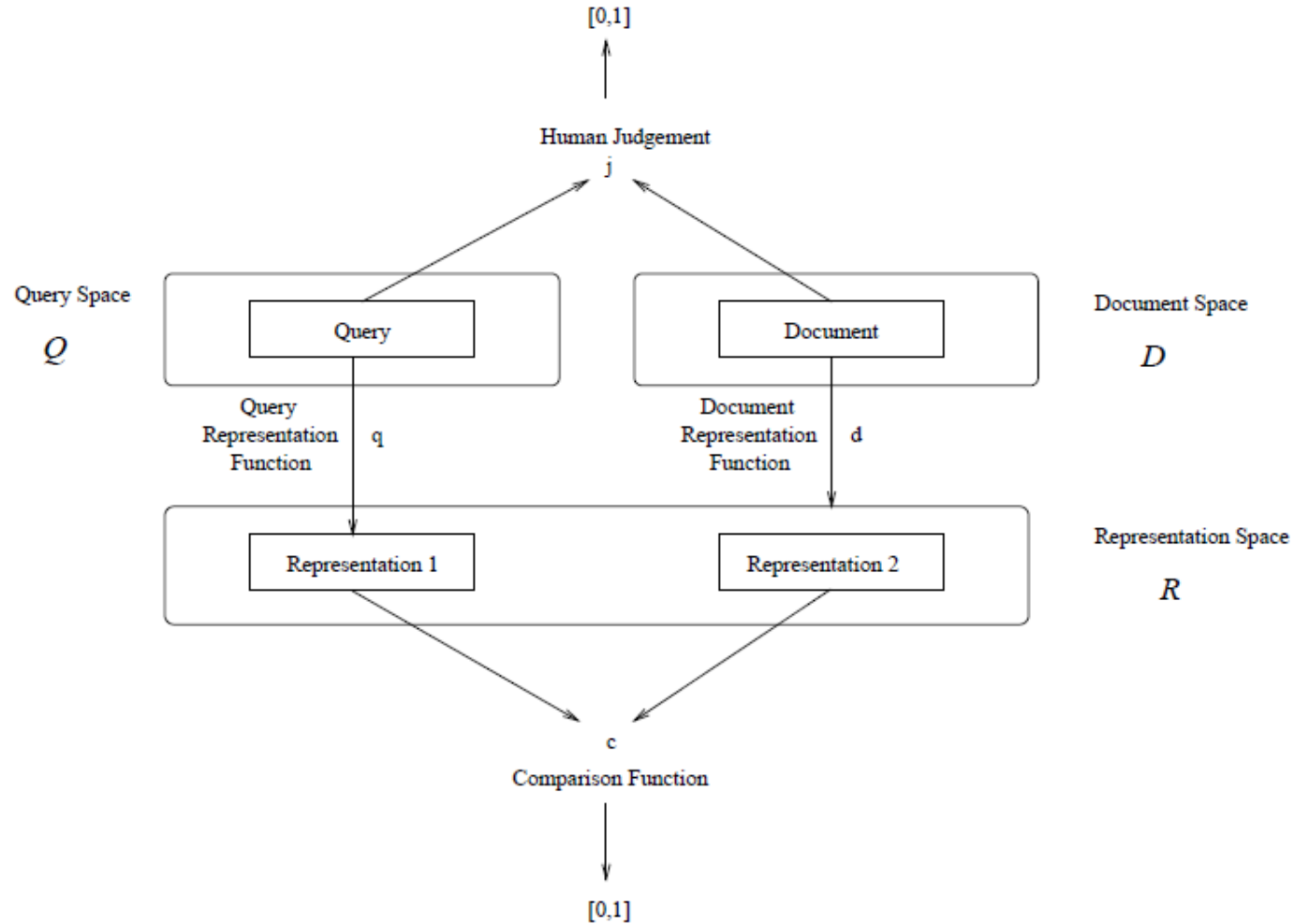
TEXT RETRIEVAL SYSTEM MODEL

- A typical query is one or two words while a document can be a hundred pages long.
- The words (query) that the user provides can change in the documents although they have the same information. “paraphrase problem”
- In an ideal system;

$$c(q(\text{query}), d(\text{doc})) = j(\text{query}, \text{doc}), \forall \text{query} \in Q, \forall \text{doc} \in D,$$



TEXT RETRIEVAL SYSTEM MODEL



COMPARISON

- Precision & Recall
 - Nothing new here, since at least 1996



APPROACHES TO MULTILINGUAL TEXT RETRIEVAL

○ Text Translation

- The most straightforward approach
- Translate either d or q and meet them in the same language. (by using an fully automatic machine translation)
- One weakness; this works only in limited domain.
 - In most cases encoding the domain knowledge is expensive.
 - In this case semantic translation errors can not be tolerated



APPROACHES TO MULTILINGUAL TEXT RETRIEVAL (CONT.)

○ Multilingual Thesauri

- A dictionary that keeps the “cross – linguistic synonymy”
 - The representation of the relations between terms found easy to understand. Thus leads the user to form better queries.
 - Significant domain knowledge can be encoded into thesaurus.



EARLY WORK & STUDIES ON THE SUBJECT

- It is a survey paper, nothing technical... Easy to read!
- So read yourself



CONCLUSION

- Multilingual text retrieval is possible
 - Text translation
 - Multi lingual Thesauri
- Can be used implement a system like that.

