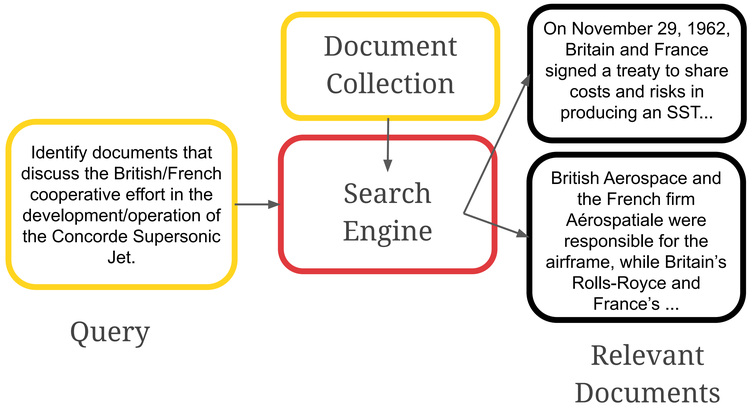
Supporting Global Knowledge Sharing using Cross Language Information Retrieval

Petra Galuščáková and Douglas W. Oard
{petra,oard}@umd.edu

Information Retrieval

Information retrieval is searching for the relevant documents in a large collection of documents using a query input by the user. The aim of the search engine is to return documents relevant to the query.

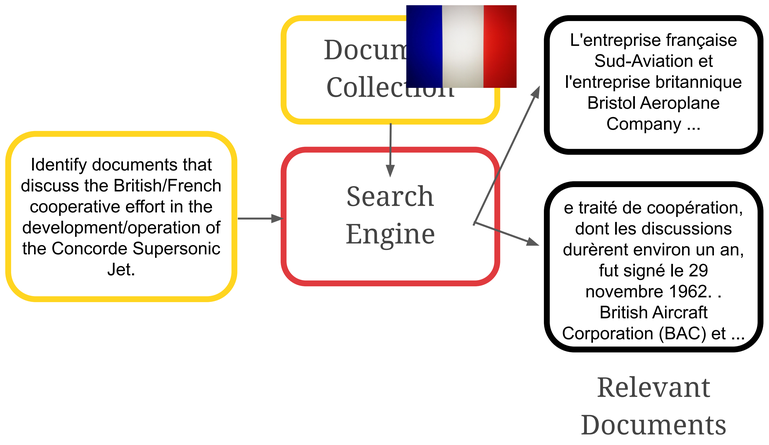


Possible Applications

- Chinese technical publications might be of value to NASA engineers working on similar problems.
- Understanding the global reaction to NASA's activities would benefit from systems that could process Hindi and French as easily as English.
- Allowing search in the oral history archives. The interviews in the Shuttle-MIR oral history collection were conducted in English; people who could speak only Russian simply weren't interviewed,

Cross Language Information Retrieval (CLIR)

CLIR is a special case of Information retrieval in which the language of the documents differs from the language of the query.



CLIR Architecture

- Documents or queries are translated into compatible representations using machine translation or dictionaries.
- Search engines can handle ambiguity and translation errors well using multiple translation variants.
- Ranking based on embeddings (dense vector representations) of terms, sentences or segments.
- Ensemble methods can improve robustness.

CLIR vs. Monolingual IR

System	Mean Average Precision
Monolingual IR (Russian queries) BM25	0.345
Cross-Language IR (English queries) Document translation + BM25	0.336
Cross-Language IR (English queries) Document translation + embeddings	**0.434**

Table 1. Comparison of the monolingual and crosslingual retrieval on the Russian 2003 and 2004 collections. Documents are in Russian, queries are either in Russian or English.

Beyond CLIR

- "Documents" might be speech or video.
- Text might be printed or handwritten.
- Content or queries may include several languages.
- Queries might be structured or simple.
- Information need might be narrow or broad.
- Summaries of relevant documents might be needed.
- Summaries and documents may need translation.
- How quickly can we build systems?

SCAN ME

SCAN ME

This research has been supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.
Retrieval results were provided by the participants of the 2021 Johns Hopkins HLTCOE Mini-SCALE workshop.