# End-to-End Neural Cross-Lingual IR by Sharing Representation

Rui Zhang
Yale University
LILY Lab

## Cross-lingual Information Retrieval (CLIR)

Information Retrieval is the task to retrieve relevant documents from a corpus for a given user query. Examples of information retrieval include Google Search Engine. Usually, information retrieval is performed in a monolingual setting, i.e., the documents and user queries are written in the same language. Traditional monolingual information retrieval technique use term-based approaches such as TF-IDF, BM25.

Cross-lingual Information Retrieval (CLIR) is the information retrieval task when the documents are in a language different from that of the user's query. As multilingual document collections are becoming prevalent, this task is an important application. For example, imagine an investor wish to monitor the consumer sentiment on Twitter from around the world. He or she can issue a user query in English and would like to find all relevant tweets in different languages.

## Translation-based Methods for CLIR

The main method to build CLIR system is the translation-based approach. The translation-based approach is the idea to solve the translation problem separately, so CLIR becomes the document retrieval task in the monolingual setting. Thus it is a pipeline of two components: translation + monolingual IR.

Depending on the translation, this approach can be further divided into the document translation approach and the query translation approach. For example, suppose the user query is in English and documents are in Swahili. We can do 1) query translation from English to Swahili using a bilingual dictionary, or 2) document translation from Swahili to English using a machine translation system.

## Neural (Monolingual) Information Retrieval

Recently, many successful neural IR systems have emerged:

- DUET (Mitra et al., 2017)
- PACRR (Hui et al., 2017)
- DSSM (Huang et al., 2013)
- DESM (Mitra et al., 2016)
- MatchPyramid (Pang et al., 2016)
- DRMM (Guo et al., 2016)

… ...

For example, The DUET (Mitra et al., 2017) model is a hybrid approach that combines signals from a local model for relevance matching and a distributed model for semantic matching.

But, they are mainly evaluated in Monolingual IR settings.

## Research Goal and Challenge

Research Goal: Build an end-to-end neural CLIR that

- models local information including unigram term match and position-dependent information such as proximity and term positions.
- models global information as semantic matching in distributed representation space.
- directly learns from (query, document, relevance) supervisions.
- performs better than the pipeline translation-based approach because it avoids cascading errors.

Research Challenges

- How can we capture local information and global information when query language and document language are different?
- How can we use and learn shared representation for multiple languages?

## Proposed Method

1) Use multilingual word embeddings to build a similarity matrix. This models local information.

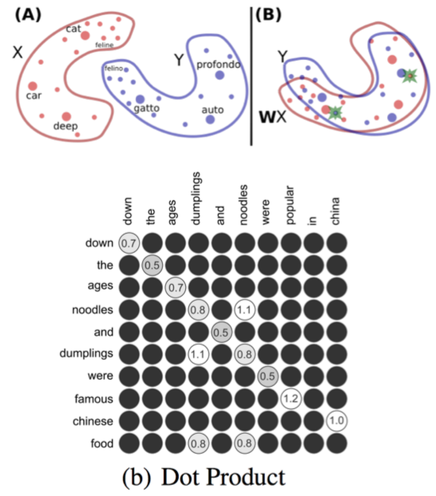


Figure 1. Similarity Matrix based on Dot Product (Pang et al., 2016) on Facebook MUSE embedding.

2) Use monolingual or multilingual embedding to learn a shared distributed representation. This models global information.

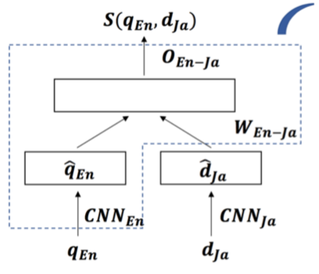


Figure 2. Shared Distributed Representation (Sasaki et al., 2018)

## Data Sets

WikiCLIR (Sasaki et al., 2018)

- Automatically created from parallel wiki pages
- Large-scale, 25 languages
- Recently published

Standard CLIR task data sets

- CLEF
- NTCIR
- TREC

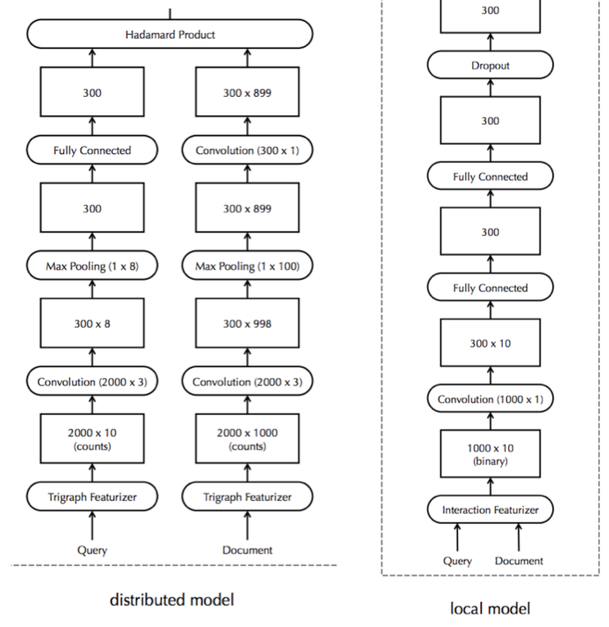


Figure 3. DUET system (Mitra et al., 2017)