NLA Project - Deliverables
Cross Lingual Information Retrieval

Submitted By
Aniruddha Deshpande (20161058)
Samyak Agrawal (20161180)

Project Mentor : Nikhil Pattisapu
Faculty Mentor : Dr. Vasudev Varma

Problem Statement: *Design a Cross Lingual Information Retrieval system considering two languages L1 and L2 of your choice over a specific Domain.*

Preliminary Decisions to be made :

- Choosing the **domain** to be worked on.
- Choosing the language pair **L1** and **L2**.
- Choosing a **Document Ranking Algorithm**.

Interim Tasks/Deliverables :

1. **Crawl and collect documents** based on the chosen domain in both languages L1 and L2. Here, the documents must be semantically similar with respect to that of the domain but not necessarily a corpus of parallel documents. The reason for not needing a parallel corpus is explained in the following point.
2. **Machine Translation Model** for query translation. The approach considered for this **CLIR** system is the *Query Translation* method. Here the input query is translated amongst the language pair (If the query is in L1, it is translated to L2 and vice-a-versa) and these pair of translated and original queries are then used as a search query in the documents of the respective language. Here, we wouldn't require a parallel corpus because independent searches in L1 and L2 would be taking place as the queries have already been translated. Since the documents are semantically similar, data relevant to the independent searches would be outputted post the information retrieval task from the documents in the respective language.
3. **Search and Retrieve** through the corpus of documents with respect to both original and translated queries independently. The retrieval process would make use of a **Document Ranking Algorithm** which ranks and scores the documents based on its relevance with the query.
4. Consider the top N documents based on their rankings/scores from both L1 and L2 language corpora. Now based on the scores of the documents amongst both languages, sort the retrieved documents (considering the scores of the documents retrieved in both L1 and L2 together) and display a **Fused Output** containing the top N' most relevant documents. (N and N' can be chosen manually.)

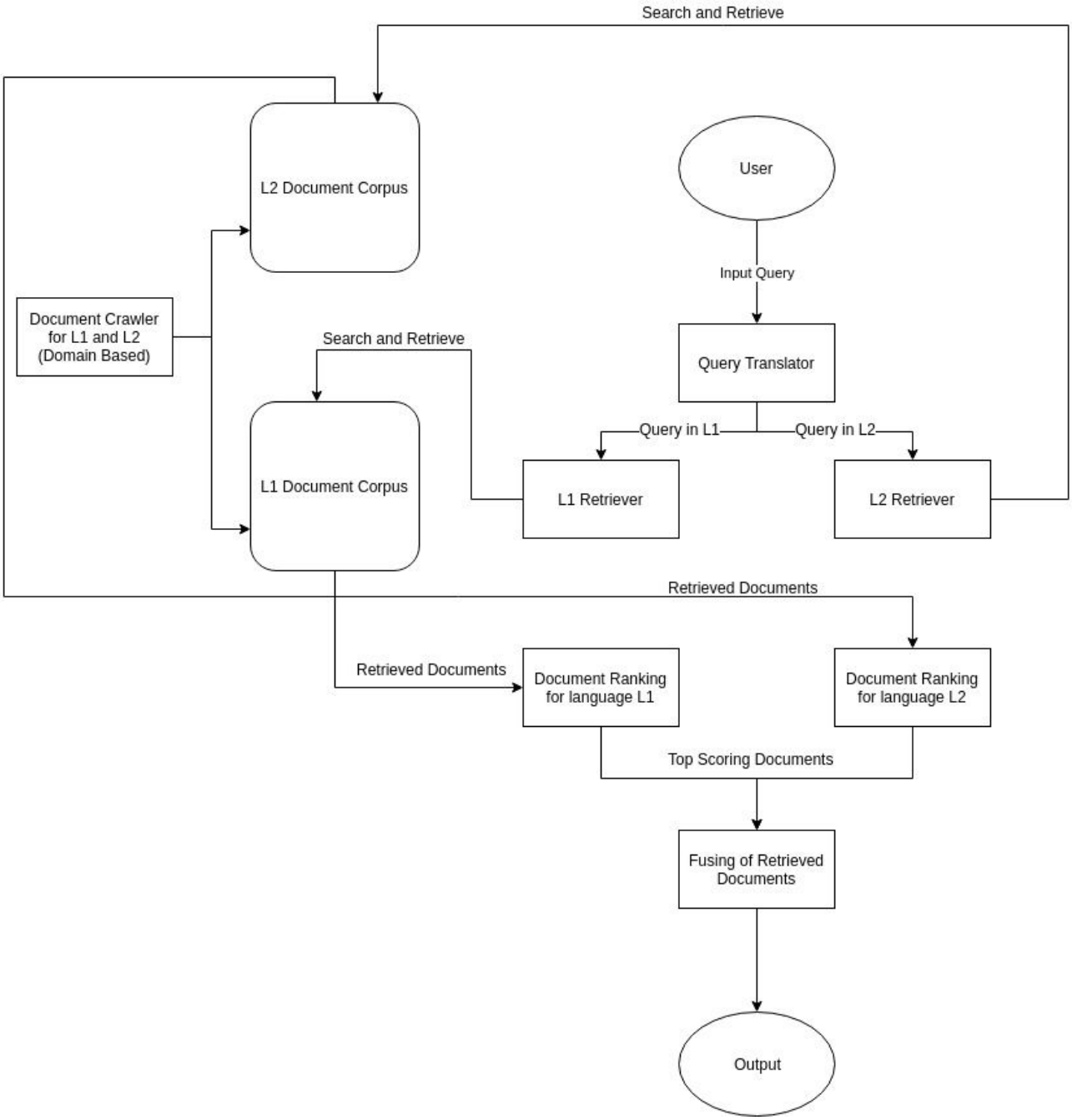

Fig. 1 CLIR system in the Interim Deliverables stage

Final Tasks/Deliverables: The final deliverables involve adding modules that improve this CLIR system. Following are the deliverables :

1. **Adding a Named Entity Recognition Subsystem:** Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, etc. This process is carried out for both L1 and L2 languages. The importance of this is described in the following task.
2. **Transliteration of Named Entities:** Lot of named entities like names or organizations may contain words which contain nouns meaning something entirely different. These entities must not be translated but rather must be transliterated within the Query Translator module to avoid irrelevant document retrievals. *For eg., The query ' रवि शास्त्री ' contains the word ' रवि ' which means 'sun' in English. We don't want this named entity to be translated into 'sun' but to be transliterated into 'Ravi Shastri'.*
3. **Query Boosting:** This task involves boosting the weights of certain important words in a query. This can be done to avoid and reduce the number of output documents that are relevant to only some certain words but not to the overall query. *For eg., A query regarding the Russian Prime Minister shouldn't generate output documents regarding Prime Minister of India. Here the weight of the word 'Russian' is boosted to avoid such a scenario.* What values are to be boosted are learned with data.
4. **Query Expansion:** The Query Translator generates multiple outputs such that all the outputs are semantically similar. *For eg., The word 'India' can be translated into ' भारत ', ' भारत देश ', ' हिन्दुस्तान ' without losing semantic value.* Now multiple such semantically similar queries can be considered and used for retrieval to *expand the query* so better outputs can be generated. This is the case for expanding the translated query. The original query can be expanded using a dictionary look-up based system that makes use of the synonyms of the words in the query to further expand it.

Please refer to the diagram of the complete CLIR system post the addition of the above modules in the final deliverables stage below.

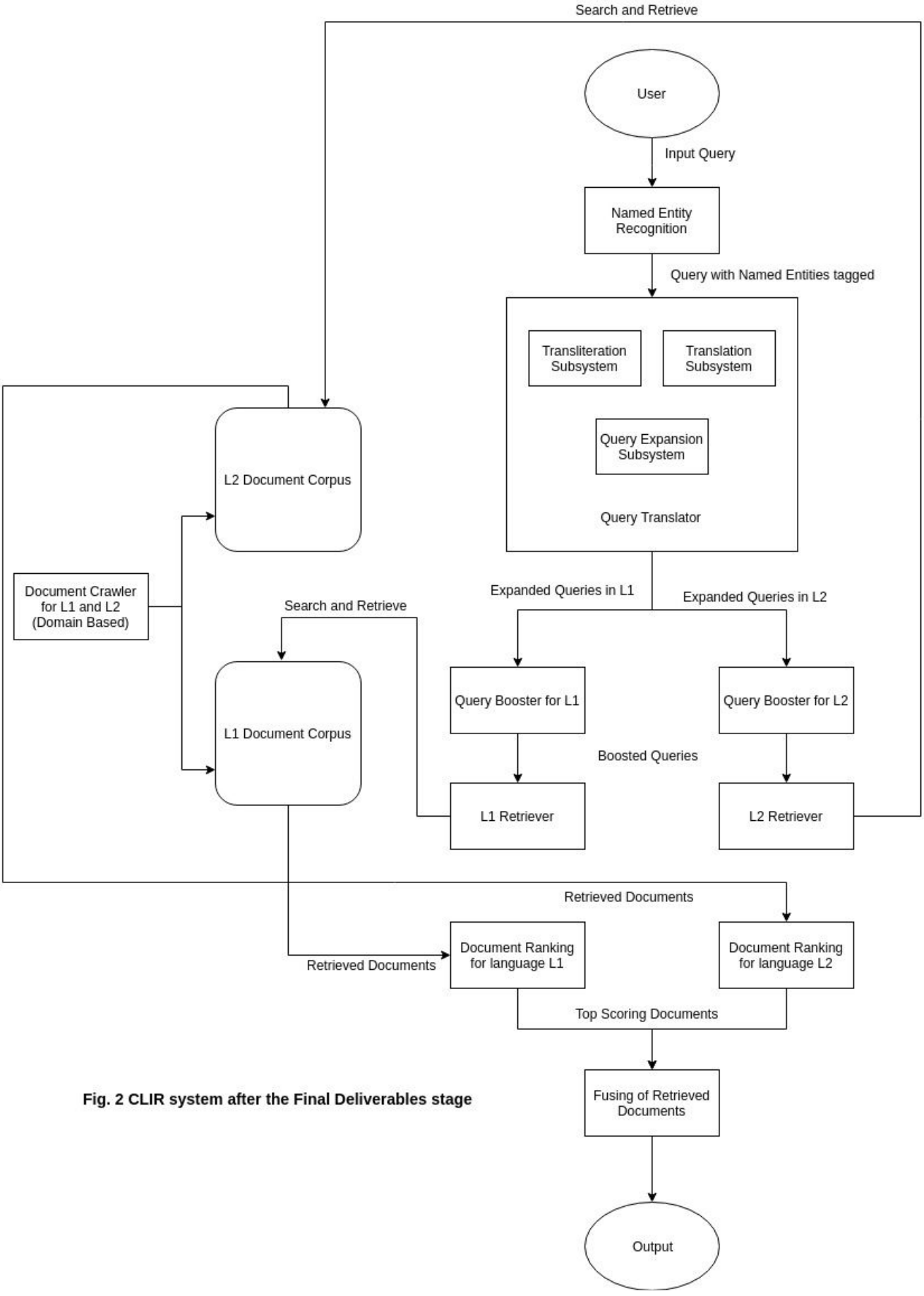

Fig. 2 CLIR system after the Final Deliverables stage