NLA Project - Final Report
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.*

Abstract

Crosslingual information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. This is opposite to that of **Monolingual information retrieval**, where both the query and the output are in the same language. CLIR can be both **Dictionary based** or based on a **Parallel Corpora** of sentences. As a part of this project, we chose L1 and L2 languages as English and Hindi respectively. Two major approaches are **Query Translation Based Retrieval** and **Document Translation Based Retrieval**. For our task, we opted for the *Query Translation Based Retrieval* as translating a Query and then carrying out the retrieval task is much easier as compared to translating an entire document. **Query Boosting** and **Query Expansion** techniques were further added to the CLIR system for further improvement in the results attained from the system.

Literature Survey

The CLIR task heavily depends on two major subtasks: **Query Translation** and **Document Ranking.** For the *Query Translation* phase, we tried using Hindi-English Models available on various open source GitHub repositories and we also tried using the models we trained as a part of Assignment-4 for the course based on [1], [2], [3]. These availed us of very poor translations which were completely unusable for our task as it would have had led to absurdly incorrect results. To avoid poor translations we decided to use a **Google Translate API** based on [4], [5] for our *Query Translation task.*

Coming to the backbone of *Information Retrieval,* we considered two models one based on **HMM** [6] and the other a **Vectorised TF-IDF based system** [7], [8] and a decision to use the latter was then made as the HMM Parameters would be very sparse due to a smaller dataset.

Dataset Created

The Domain chosen for the CLIR system is **Indian Entertainment (Bollywood) News Articles**. To achieve maximum possible semantic similarity we chose English and Hindi News websites owned by the same News House: India Times and Navbharat Times (Both owned by India Times). Following are the links to their Indian Entertainment section :

- India Times.
- Navbharat Times.

Above two links contain links to the individual news articles, which were retrieved using *retrieve_links.py* module within our source code. The retrieved links were then saved and were used by *retrieve_articles.py* , which uses **NewsPlease** python library to retrieve the following data about the article :

- URL of the Article.
- Title.
- Short Description of the Article.
- The Text of the Article.
- Authors.
- Publishing Date.
- URL of the Image in the Article.

These retrieved values were then saved as JSON objects with their filenames retrieved from their unique URLs. These objects could be displayed as HTML file outputs using *make_html.py* .

Approach

After the above dataset is created and their respective HTML pages are made, the data from both English and Hindi languages is converted in terms of *TF-IDF Vectors*. The language L1 in which the input query (out of the English-Hindi language pair) is given is detected and then translated into L2 language. The translation step is carried out by *Google Translate API* which also helps in detecting *Named Entities* within the query and transliterates it into the script used by language L2. Post translating the queries (Orignal and Translated Query) are also converted in terms of TF-IDF vectors and based on the cosine similarity values between the query and the document in the respective language, the documents are ranked. Top K documents in both English and Hindi are then displayed using an HTML page *output.html* (consisting links to the top K documents) for user convenience. The value of K is specified by the user.

The user can choose to **boost** certain terms within the query (see *Experimental Setting and Usage* section). To carry out query boosting, the boosted terms are considered as separate queries and documents are retrieved for using the above document ranking method for both L1 and L2. The cosine similarity values calculated for the top K documents for these separate

queries are then multiplied by the **boosting multiplier** (in our case, we chose it to be 1.5). The new boosted similarity values are then compared with the unboosted similarity values of the original query and are then they are ranked. The final top K ranked documents are then retrieved. (Please refer to **Fig 2.**)

In addition to this, the user can choose to expand the queries as well. *Query Expansion* in this CLIR system is based on synonyms. For English queries, thesaurus.com is scraped and top 5 single worded synonyms are retrieved and the query is expanded by replacing the words with their synonyms and the new synonym-replaced queries are then used for search and retrieval. A similar step is carried out for Hindi queries as well where synonyms are taken from a saved text file with synonyms for up to 800+ words. The word net was avoided for synonym finding process as it tends to also return multi expressions with the same meaning as the word. Similar to above, the documents are then combined and ranked to retrieve the top K documents. Query boosting is applied to the synonyms as well if the original word to the synonym is specified to be boosted by the user. (See **Fig. 1.**)

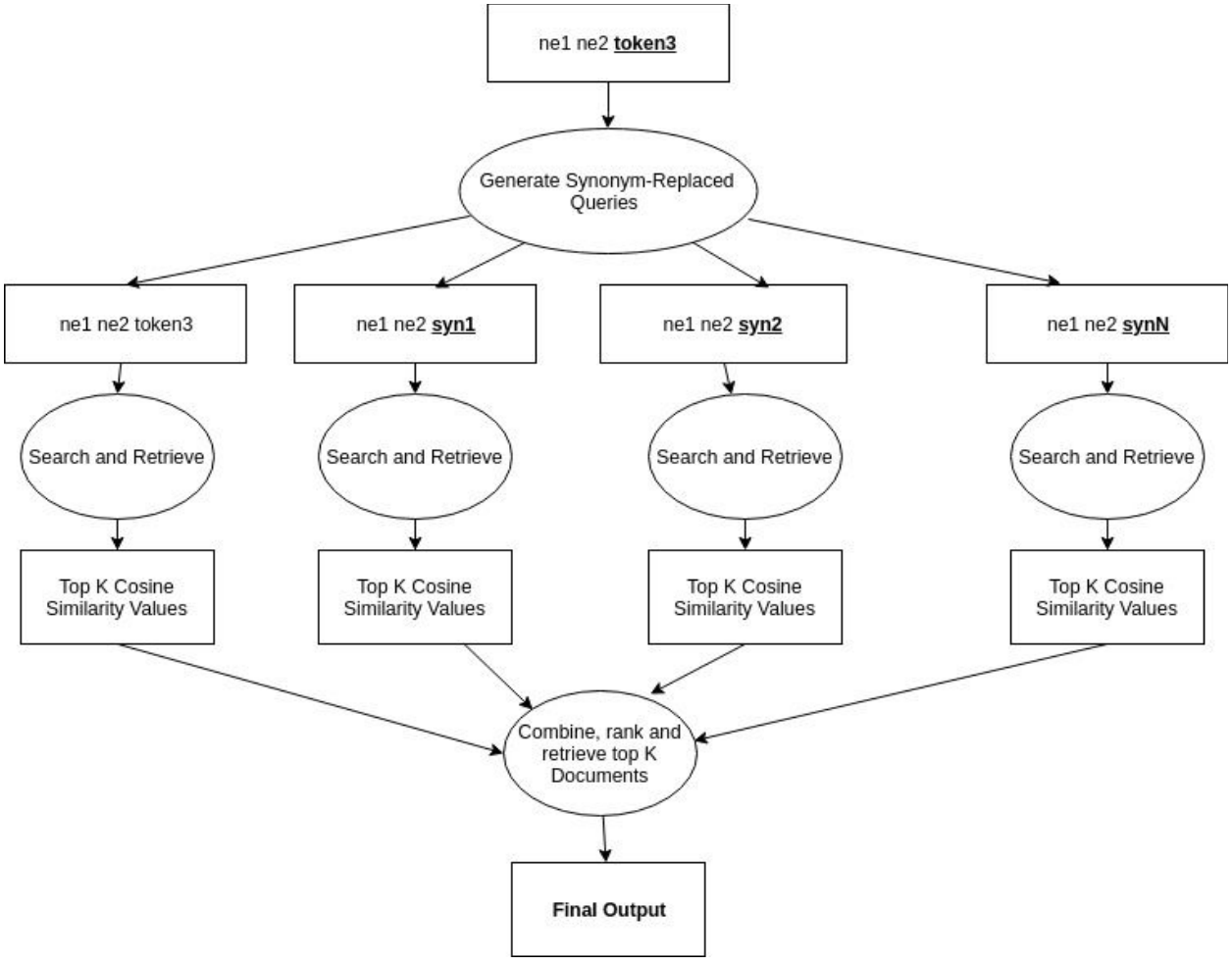
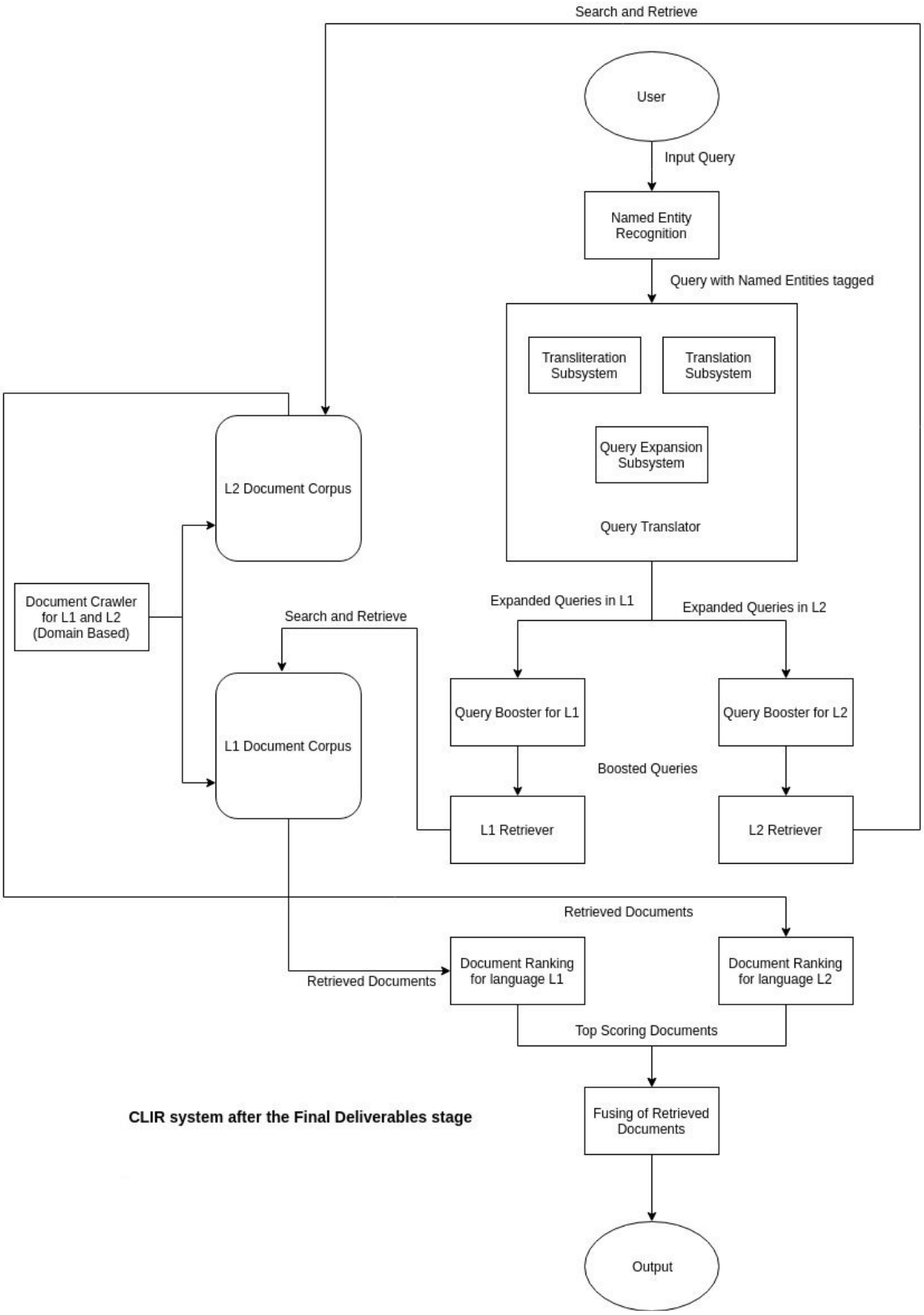


Fig. 1 Query Expansion using synonyms. ('ne' here represents Named Entity)

Fig 2. Query Boosting Architecture (Bold and Underlined tokens are Boosted)

Following is the **Final Architecture after the Final Deliverables stage (Fig. 3)** :



CLIR system after the Final Deliverables stage

Experimental Setting and Usage

The code is written in **Python3.**

Required Libraries:

- argparse
- webbrowser
- json
- os
- requests
- BeautifulSoup
- newsplease
- re
- datetime
- urllib.request
- nltk
- string
- collections
- Num2words
- string
- numpy
- math
- googletrans

Usage:
python3 CLIR.py "query" k e

Here, *'k'* represents the number of top-ranked documents that are to be retrieved. *'e'* is a 1/0 bit value which is to specify whether to expand the query or not (1 - Expand, 0 - Don't Expand).

Within the query, user can boost certain terms by encapsulating them within double square braces eg. *[[boosted query tokens]]*.

Example Input:
python3 CLIR.py "Farhan Akhtar's [[Marriage]]" 5 1

Here the word *'Marriage'* is to be boosted and top 5 documents are to be retrieved after carrying out query expansion.

Results

Following is the top ranked document for the above example query *“Farhan Akhtar’s [[Marriage]]”* in both English and Hindi languages :

English:

Title: Farhan Akhtar:In April, We May: Farhan Akhtar Drops A Major Hint About His And Shibani Dandekar's Wedding!

Text:
Farhan Akhtar and Shibani Dandekar might not have given any official confirmation on their relationship status but their social media PDA says it all. Be it their pictures together or his valentine's poem for his ladylove Shibani, they have confirmed it without actually saying the words. Love is in the air for Farhan and Shibani and their frequent vacays, public appearances and social media PDA have said it all! While the couple is spending most of their time together, several news reports claimed that they are planning to tie the knot soon. When a leading newspaper asked Shibani if she was in a relationship with Farhan, she chose to neither confirm or deny the news. She said, Don't Miss 3.8 K SHARES 24.3 K SHARES 17 K SHARES 11.2 K SHARES 9.5 K SHARES "I am not secretive, but I don't feel the need to say things out loud. I don't need to make an announcement about who I'm dating. It's up to me to decide when and what I want to share about my personal life, and it is up to the audience to decide how they look at it. How much information I want to put out in the public domain is my prerogative." A few months ago, a news report had quoted a source close to a couple saying, "They are extremely serious about each other and Farhan's kids have also warmed up to Shibani, so this looks like the most obvious step for the two." After all the cryptic interviews, looks like Shibani and Farhan are serious about taking the next step together! Instagram Farhan appeared on Film Companion's Tapecast along with Bhumi Pednekar, where he dropped a major hint on his wedding. Bhumi played a Do-Not-Play cassette for Farhan, where Shibani asked when they're getting married. Apart from laughing at the question, he added, Instagram "She's just having fun with all the rumours going around. I don't know it maybe April or April may be May." Reports also claim that the couple will opt for a private christian ceremony.

Hindi:

Title:
फरहान अख्तर अप्रैल में करेंगे गर्लफ्रेंड शिबानी दांडेकर से शादी!

Text:
X ऐक्टर फरहान अख्तर और उनकी कथित गर्लफ्रेंड शिबानी दांडेकर भले ही अपनी रिलेशनशिप को लेकर बात करने से कतराते हों, लेकिन उनकी हॉलिडे पिक्स, इंस्टाग्राम पोस्ट, इवेंट में साथ में मौजूदगी रिश्ते पर मुहर लगाती दिखती है। हाल ही में शिबानी ने ऐसी तस्वीर शेयर की थी जिसके बाद दोनों के सगाई कर लेने को लेकर चर्चा शुरू हो गई थी। अब दोनों की शादी की संभावना को लेकर फरहान ने बयान दिया है।एक चैट शो में भूमि

पेडनेकर के साथ पहुंचे फरहान अख्तर ने अपनी पर्सनल और लव लाइफ को लेकर खुलकर बात की। शिबानी के साथ शादी का सवाल किए जाने पर उन्होंने कहा कि वह अप्रैल में शादी कर सकते हैं 'इन अप्रैल, वी मे'। अब यह बात कितनी सच होती है यह तो अगले महीने ही पता चल सकेगा। वैसे अगर दोनों शादी के बंधन में बंधते हैं तो फरहान के लिए यह दूसरा विवाह होगा।शिबानी दांडेकर ने हाल ही में एक फोटो शेयर किया था, जिसमें वह फरहान अख्तर का हाथ पकड़े दिख रही हैं। दोनों के हाथों पर स्टाइलिश रिंग्स नजर आ रही थीं। शिबानी और फरहान के एक साथ रिंग्स पहने फोटो सामने आने के बाद से दोनों के सगाई कर लेने का अनुमान लगाया जा रहा है।इस स्टार कपल के रिश्ते में होने की बात तब सामने आई थी जब शिबानी ने सोशल मीडिया पर एक मिस्ट्री मैन के साथ तस्वीर शेयर की थी। फैन्स को यह पहचानने में देर नहीं लगी कि यह शख्स फरहान अख्तर हैं। इसके बाद से ही दोनों एक-दूसरे के साथ बिताए पलों के फोटोज शेयर करते दिखाई देते हैं।

References

[1] Ilya Sutskever, Oriol Vinyals, Quoc V. Le - Sequence to Sequence Learning with Neural Networks: https://arxiv.org/abs/1409.3215
[2] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio - Neural Machine Translation by Jointly Learning to Align and Translate: https://arxiv.org/abs/1409.0473
[3] Minh-Thang Luong, Hieu Pham, Christopher D. Manning - Effective Approaches to Attention-based Neural Machine Translation: https://arxiv.org/abs/1508.04025?context=cs
[4] Yonghui Wu, Mike Schuster, Zhifeng Chen, et. al - Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation: https://arxiv.org/abs/1609.08144
[5] Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team - A Neural Network for Machine Translation, at Production Scale: https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html
[6] Jinxi Xu, Ralph Weischedel - Cross-lingual Information Retrieval using Hidden Markov Models: https://www.aclweb.org/anthology/W00-1312
[7] Apra Mishra, Santosh Vishwakarma - Analysis of TF-IDF Model and its Variant for Document Retrieval: https://ieeexplore.ieee.org/document/7546200
[8] Shadi Saleh - Cross-lingual information retrieval systems: Methods and Challenges: http://ufal.mff.cuni.cz/~zabokrtsky/pgs/thesis_proposal/shadi-saleh-2-proposal.pdf
[9] CFILT - Cross lingual Information Retrieval: http://www.cfilt.iitb.ac.in/resources/surveys/Swapnil-Cross-lingual-Information-Retrieval.pdf
[10] William Scott - TF-IDF from scratch in python on real world dataset: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089