NLA Project - Interim 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.*

Deliverables :

1. Crawl and collect documents.
2. Machine Translation Model.
3. Search and Retrieve.
4. Display top N fused documents.

Task 1: Crawl and collect documents.

This task involved crawling and scraping out domain centred documents in both L1 and L2 languages which are semantically similar. Here, L1 and L2 are English and Hindi respectively. The Domain we decided to choose 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.

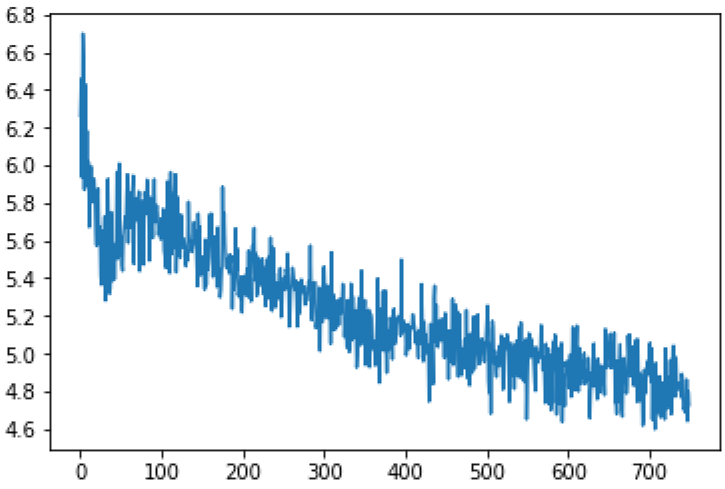
Above two links contain links to the individual news articles, which were retrieved using ***retrieve_links.py*** module within our source code. The retrieved links were then saved and were used by ***retrieve_articles.py***, which uses ***NewsPlease*** python library to retrieve the following data about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

Task 2: Machine Translation Model.

The model is trained using an encoder-decoder model with attention modelling. We tried translating using a pre-trained model and also tried a model that we trained on our own. For now, the model is not performing or giving satisfactory results in both cases. We will also be adding NER to it later on (as per the Final Deliverables), which may solve some issues. Another issue is that the model is not trained on the domain on which we are querying it. This leads to the problem of missing tokens, hence causing errors at the query translation stage. The poor performance of the attention based model can be visualised in the following Loss vs. Epochs graph. Loss up to amounts of 4.8 can be seen after several epochs of training (up to 3 Hrs on GTX 1050 on 50,000 pairs of English-Hindi Sentences.)



Task 3: Search and Retrieve.

Now, the queries in Hindi and English can be used separately for retrieval through the saved articles as JSON objects. For this task, Document Ranking is done using **cosine similarity** between the query and the articles. To do this the **TF-IDF values** of the **Text and the Title** were calculated first and then these values were later on converted to vectors. Higher weights are assigned to the Titles over the Text of the articles. Similarly, queries were also converted to that of vectors. The cosine similarity between each article and the query were then calculated and sorted in descending order. Top K articles can be retrieved in this matter. Following is an example of one such retrieval in English.

Query: *Salman Khan.*
Outputs:

Top Ranked Article is as follows:

*Title: Amitabh Bachchan:Did You Know Amitabh Bachchan Was Once Mistaken For **Salman Khan**? His Response Was Epic!*

Did You Know Amitabh Bachchan Was Once Mistaken For Salman Khan? His Response To It Was Quite Cool!

Did You Know Amitabh Bachchan Was Once Mistaken For Salman Khan? His Response To It Was Quite Cool!

Have you ever been mistaken for somebody else? For us, it might be a usual thing. But can you imagine celebrities being in a situation where they are confused for other stars?

In the west, it has happened quite a lot of times. For instance, once a fan asked This Is Us star Justin Hartley about his wife Blake. If you didn’t get the drift, he was mistaken for Ryan Reynolds. Can you imagine?

Recently, Amitabh Bachchan also revealed about one such incident that happened with him. Big B was once confused as Salman Khan. I know, even I want to know who that person was?

“We were shooting on the streets in Glasgow and then I had to walk on the footpath. As I was walking, a car went by and our desi bhai was sitting in the car, who waved and said: 'Hey, Salman Khan, how you doing?' So, I waved back at him and walked on. What to do," the Badla actor told Shah Rukh Khan,” Amitabh Bachchan recently revealed once again at the promotions of Badla.

Don't Miss 529 SHARES 8.9 K SHARES 3.1 K SHARES 248 SHARES 20.7 K SHARES

T 2850 - I walk the street of Glasgow by myself .. until a car drives by and occupant yells out .. " hey Salman Khan how you doin' .. " pic.twitter.com/RJ5neJXBaj — Amitabh Bachchan (@SrBachchan) June 27, 2018

If you remember, he had also tweeted about it last year writing, “I walk the street of Glasgow by myself .. until a car drives by and occupant yells out .. " hey Salman Khan how you doin' ..”

<u>10th ranked Article is as follows:</u>

*Title: Kareena Kapoor Has A Dating Advice For Sara Ali **<u>Khan</u>** And It Holds True For All The Newcomers*

Kareena Kapoor Has A Dating Advice For Sara Ali Khan And It Holds True For All The Newcomers
Kareena Kapoor Has A Dating Advice For Sara Ali Khan And It Holds True For All The Newcomers
Bollywood actress Kareena Kapoor is known to be the most fearless stars we have in the industry. Whether it's openly talking about her Whatsapp group with her girls which gives her all the gossip or admitting that she's a fashionista even in her gym clothes, Bebo is unapologetic about her views.
Twitter
She has completed 19 years in Bollywood and done over 50 films. Married to the Prince of Pataudi, Saif Ali Khan and the mother of Taimur, everyone wants a sneak peek of Kareena and her daily routine.
Right from her gym, to her make up, airport spotting, attending birthday parties, Kareena is one of the most searched celebrities on Instagram. And now her step-daughter Sara Ali Khan has also made her debut in Bollywood.
Twitter
Don't Miss 554 SHARES 3.7 K SHARES 8.9 K SHARES 1.6 K SHARES
And unlike a typical scenario of a stepmom household, Sara and Kareena are extremely fond of each other. Infact, in an episode of Koffee With Karan, Sara admitted she's extremely happy that her father Saif Ali Khan got married to Kareena Kapoor as she was too happy to have Pooh in the house.
Sara's friends feel she must have conspired this marriage as she is a big fan of the actress. Well, feelings are mutual as Kareena also absolutely loved Sara's performance in Kedarnath and Simmba.
Instagram
In one of the celebrity chat shows, Kareena was asked what is that one dating advice she's d like to give Sara. And like always, the stylish mom and celebrity had an on fleek answer. She advised Sara not to date her first hero.
Well, quite an apt one considering most of the newcomers end up dating their first co-stars. This advice is not only good for Sara but for all the upcoming actors too. What do you think?
Kareena Kapoor will next be seen opposite Akshay Kumar in Good News.

Here we see the top-ranked article was able to find an article on Salman Khan (even the next few articles along the ranking were able to find such articles.) But the 10th most ranked article was matched with ***‘Khan’*** as opposed to ***‘Salman Khan’*** as a whole, which is expected as only 180 articles were scraped (more can be scraped by changing parameters in ***retrieve_articles.py***). Similar results were obtained for the Hindi Language as well.

Task 4: Display top N fused documents

Since we have the cosine similarity scores for both English and Hindi documents, we can easily fuse them and segregate out top N documents. This can be done post we find a working MT model for our task.

The entire code with Data can be found at this GitHub Repository.