German & English Multi Lingual Information Retrieval

Vijaya Teja Rayavarapu
Keerthi Bhavani Gosika
Shreeyash Amit Yende
Ankur Yadav

December 2021

1 Section I

1.1 Overview of Multi Lingual Information Retrieval

Multi Lingual Information Retrieval (MLIR) refers to the ability to process a query in any given language and retrieve all the relevant documents available in any language. Unlike Monolingual Information Retrieval systems where both the documents available to the systems and the queries posed are in the same language, MLIR systems need to deal with documents which can be of two types: (i) Multiple monolingual documents belonging to various languages, (ii) individually multilingual documents, where more than one language may be present in a single document, and queries which can be posed in any of the available languages. As the importance of MLIR research became prevalent a separate track for Cross Language Information Retrieval was started in the TREC (Text REtrieval Conference) organized by the National Institute of Standards and Technology (NIST) which was continued by Cross Language Evaluation Forum (CLEF) focused on European languages.

1.2 Scope of the task

In the domain of MLIR the languages focused on for our task are restricted to German and English. The domain of the documents is restricted to English and German Wikipedia articles focusing on Science, Technology and their history. The ideal expectation from the system developed for this task is that, given a query in either English or German, the system retrieves all the relevant documents from the aforementioned collection containing both English and German documents in a single ranked list sorted based on their relevance to the query.

1.3 Importance to the users

With it's growing collection of millions of articles about versatile topics in various languages Wikipedia provides users with information to get started with a general idea of what a topic is about and cross references to related resources for further reading. As with the web, even though information related to multitude of topics across languages is available on Wikipedia it is not guaranteed that all the relevant information related to a topic can be found in a single language alone or that the translated documents of every language are present in every other language. In this scenario the language restriction of monolingual information retrieval systems might result in information gaps as it is not necessary that, for any given topic a single language would contain all the relevant information.

Even though the origins and advancements in Science and Technology are spread across the world, significant amount of this information is stored in English and German. Out of the 50 millions articles on Wikipedia, English and German constitute almost 16%. Also, Germany has been a key contributor to many innovations, so the original documents written in German might likely contain more detailed information and insights into these topics. Even though individual Wikipedia pages could be translated to the required language in order to read an article written in a foreign language, manually finding the right set of relevant documents in the foreign language (German or English in our case) which are worth translating and reading, across millions of available documents is a challenging task.

The system for our proposed task would break the language barrier between English and German and provide the most relevant documents across both the languages to the broad spectrum of users including English and German speakers who want to gain knowledge of various topics in Science, Technology, and their histories. This system ensures that users need not waste their time to conduct independent searches on each of these languages to find all the related information and then spend time to read through and figure out which documents are of value and which aren't. Instead they could solely rely on the list of documents that the system returns. This system not only saves the users' precious time but also, by combining useful documents from both the languages, provides a much broader general idea about a given topic.

1.4 Overview of queries

The queries for the defined task are informational in nature and can be either in English or German. As users begin their search to gain information about a topic, they often expect a set of relevant documents as compared to a single document. Our queries and information needs are formulated in such a way that the topic is not too narrow and is broad enough to get multiple documents which could be categorized as relevant. The queries for our system would be ideally 1 to 6 words in length. This query - **"types of particles in physics"** is an example of how a query in English would look like. This query - **"arten von chemischen Bindungen"** translating to "types of chemical bonds" is an example of a German query. The queries are case insensitive and any special character present in the query will be omitted.

1.5 Relevance judgement of queries

The relevance of results related to a given query is manually scored on a scale of 3.
(a.) **2** - Highly relevant
(b.) **1** - Moderately/Somewhat relevant
(c.) **0** - Irrelevant
For the first query mentioned in the previous section i.e **"types of particles in physics"**, if the resultant documents retrieved from across English and German articles contain information related to the particles in physics such as Fermions, Bosons, Hadrons etc. then the documents are highly relevant and are given a score of **2**. If the documents contain explanations about what particles mean in physics without indulging into the information about the types of particles or describes about the branch particle physics then they are considered somewhat relevant and given a score of **1**, and if the documents contain no information

related to these particles or what they mean then they are deemed irrelevant and given a score of **0**. Similar approach would be followed for the relevance judgement of the second example query in the previous section as well as for the queries we use for the evaluation.

1.6 Organization of results

Once the candidate set of relevant documents from each of the languages is retrieved then then they are merged and the final results comprising of the top-k relevant documents across the English and German collections are returned in a single ranked list sorted in descending order of relevance scores computed by the system. Each result would be presented in a single row on the web page and the composition of a single result would be as follows:

<url>Url of the page</url>
<snippet>Snippet of text from the document in the query language with query elements **highlighted** </snippet>

Example:
query : **"types of particles in physics"**

result 1: English article.
https://en.wikipedia.org/wiki/Fermion
In **particle physics**, a fermion is a **particle** that follows Fermi–Dirac statistics and has half odd integer spin.

result 2: German article with snippet translated to English.
https://de.wikipedia.org/wiki/Boson
Bosons (after the Indian physicist Satyendranath Bose) are all **particles** in **particle physics** that behave according to the Bose-Einstein statistics , in which a.o. several indistinguishable **particles** can assume the same state.

Since all the topics do not necessarily co-exist in both the languages, some queries might have results which are predominantly German Wikipedia articles and some might have results which have a majority of English articles.

1.7 Evaluation metrics

Similar to other Information Retrieval tasks dealing with millions of documents and with the web, the usage of traditional metrics such as Precision, Recall and their averages is not possible in our case as well, as for computing these metrics we would need to know the relevant documents across millions of documents for each of the queries or at least the number of relevant documents for the average precision. In order to overcome this limitation and evaluate our system efficiently we manually annotate the relevance judgement values for the retrieved results from our system for the test queries. Using these relevance judgements for the top-k results we compute **Mean Average Precision@10** and **Average of Normalized Discounted Cumulative Gain at rank 10**, both of which do not require knowledge of all the relevant documents in the collections or their count.

1.8 Proposed Modelling Approach

For this proposed task the data required would be the German and English Wikipedia articles which belong to the categories of Science, Technology and their histories. There are two ways for procuring the data:
1.) Build a dynamic web-crawler that updates these collections by periodically visiting the web pages.
2.) Procure static dump of Wikipedia articles for English and German.
Even though both of these approaches have their pros and cons we opted static dump as we wanted our focus in the modelling to be on the retrieval model rather than the gathering process of data. Moreover, using a static dump will allow the results to be deterministic and hence easier for evaluating the performance of the model as compared to when the articles are constantly updated and results keep changing.

This leads us to the next part of our proposed system which is cleansing of data and generating inverted indices. The steps for each of the languages for this process would be as follows.

For each document in the language specific collection:
1.) Set a unique id for the document.
2.) Exclude title section from the document. (This is done to promote the model to focus on content rather than the header or title.)
3.) Convert the entire document body to lowercase.
4.) Remove all special characters excluding apostrophe (').
5.) Split the document into tokens based on space.
6.) Lemmatize each token in the document.
7.) If a token is not present in the language specific vocabulary add it to the vocabulary and provide it with a unique index.
8.) Convert the document into term ids based on the mapping in the vocabulary.
9.) Compute unigram term frequencies for the document.
10.) Store these term frequencies with the unique id of document as the key.

Once all the documents in a language are processed and term frequencies are computed, this would be the input for generating inverted indices. All the document term frequencies for a language collection are traversed and for each unique term the following information is stored in the list associated with it: (document id, term frequency, document size). The third attribute stored i.e **document size** is for computing the smoothing value. This list for each term would be sorted in ascending order and the inverted indices would be stored to the disk.

Once the above steps are finished we would have two sets of inverted indices, one for English and the other for German. From here on the goal would be to measure the closeness of a document to a given query using these inverted indices. Since the tokens for the inverted indices are stored in the same language as the collection, the challenge of translating and making the query available in both the languages for using these inverted indices comes up. For tackling this process we intend to use Neural Machine Translation (a pre-trained Transformer) for converting a query given in English to German and vice-versa. This way once the given query passes through the NMT layer it would be available in both the languages. The queries in both the languages are then processed in the exact same way as the documents

in their respective languages. For the tokens in the query, if the token isn't present in the vocabulary of the language, we substitute it with a synonym which is present in the vocabulary and if no such value exists, we use dynamic programming to decompose the token into smaller tokens which are present in the vocabulary while trying to minimize the count of sub-words and maximizing the overall number of characters utilized. These additional steps are performed in order to overcome the unintended side effects that might be caused because of NMT.

Once these tokens are available a Query Likelihood retrieval model which uses uni-gram language models of the documents and collections would be used to find the relevant documents. Since there are chances that the exact combination of all words might not be present in all the relevant documents we use Dirichlet smoothing to avoid assigning zero probabilities to relevant documents. The steps followed by the model would be as follows.
For each of the languages, the tokens of query in that specific language are used.
1.) All the candidate documents which contain at least one query token are found.
2.) For each of these documents, their uni-gram language models along with the language model of the collection are used to compute the sum of log probabilities of how likely it is to generate the given query from the document.
The following formula is used to compute this score for each of the documents:

$$logP(Q|D) = \sum_{i=1}^{n} log \frac{f_{q_i,D} + \mu \frac{c_{q_i}}{|C|}}{|D| + \mu}$$

The log probability computed as above using the respective language models is the document score after applying Dirichlet smoothing. $P(Q|D)$ is the probability of generating a query from a given document. $f_{q_i,D}$ is the frequency of i^{th} term of query in the given document and $|D|$ is the size of the document. μ is the parameter value set. c_{q_i} is the frequency of i^{th} term of query in the entire collection and $|C|$ is the size of the entire collection.
3.) Once the document scores are computed, all the candidate documents are sorted in descending order based on these scores.
4.) Top-K documents from each of the language are chosen.
5.) All these Top-k documents are merged together and sorted again based on their document scores, now irrespective of the language they are from. This is done with the intuition that irrespective of the language, the document score measures the closeness of a document to a given query.
6.) Snippets are generated, along with translations wherever required, for the Top-K documents in the merged and sorted list, These Top-k final results are organized as mentioned in section 1.1.6 and returned to the user.

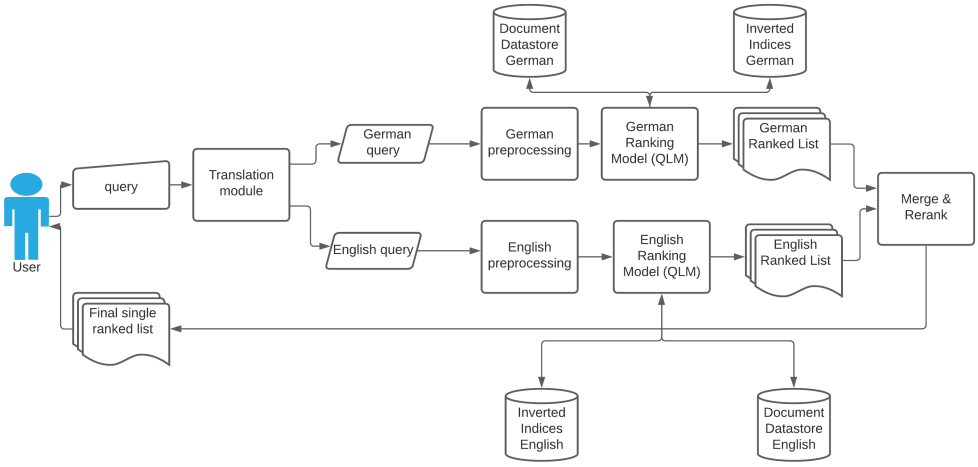

Figure: Schematic of the proposed model.

2 Section II

2.1 Approach for query formulation

The queries have been carefully curated keeping the domain and the preference of having more than one relevant document in mind. Since we were a team of four, two of us worked on formulating 10 German queries while the other two worked on 10 English queries. We worked in pairs where one person drafted the query and wrote down relevance criteria and the information needs while the other person searched both English and German Wikipedia to get relevant articles and noted down 5 urls for each language. Once these 10 urls were written down the person posing the query went through them and gave them relevance scores based on the defined relevance criteria. The roles were reversed after 5 queries. This gave us a total of 20 queries that could be used for evaluating the performance of our model.

2.2 Queries for evaluation

Below are the 20 queries formulated by us for the evaluation of our model. Each query is accompanied by information needs and relevance criteria. The relevance criteria is same as explained in section 1.1.5 i.e
(a.) **2** - Highly relevant
(b.) **1** - Moderately/Somewhat relevant
(c.) **0** - Irrelevant

Query	Information need	Relevance criteria
chemical compound types	Documents about various types of chemical compounds such as organic, inorganic, ionic etc.	2 - Explanations about a specific types of chemical compounds. 1 - Explanation about what a chemical compound is and examples of chemical compounds. 0 - Anything other than the above two.
life of einstein	Documents about Einstein's childhood, education, family members, wealth, inventions, discoveries, awards.	2 - Documents containing aspects mentioned in information needs. 1 - Documents about building, awards etc named after Einstein. 0 - Anything other than the above two.
submarines used in world wars	Expected about U-boat and different types of submarines manufacturer details and the countries that had submarine building technology.	2 - Documents containing aspects mentioned in information needs. 1 - Explanations about what submarines are. 0 - Anything other than the above two.
effects of technology on lifestyle	Expected the documents to give information on how it saves time and effort and also the differences in lifestyles in the past and now .	2 - Documents containing aspects mentioned in information needs. 1 - Explanation about what technology or lifestyle is. 0 - Anything other than the above two.
allopathy medicine	Documents about who invented, why people shifted to allopathy from ayurveda, homeopathy.	2 - Documents containing aspects mentioned in information needs. 1 - Explanation about what medicine is or other types of medicines. 0 - Anything other than the above two.

Query	Information need	Relevance criteria
physics principles used in automobiles	What physics principles are involved in automobile design, how performance of vehicles can be improved through physics and how the structure is designed for cars, how batteries work.	2 - Documents containing aspects mentioned in information needs. 1 - Explanation about what a what physics principles are or what automobiles are. 0 - Anything other than the above two.
technological advancements in agriculture	Expected list of technologies and practices that were implemented in the agricultural field like pest control, large scale farming, tractors and other heavy machinery, motorized irrigation etc.	2 - Documents containing aspects mentioned in information needs. 1 - Explanation about what a agriculture is or what technological achievements happened in the past. 0 - Anything other than the above two.
types of particles in physics	Explanation about the particles in physics such as Fermions, Bosons, Hadrons etc.	2 - Explanations about the particles in physics. 1 - Explanation about what a particle physics is and details related to it without discussing the details of the types of particles. 0 - Anything other than the above two.
famous physicists in the 1900s	Anything about famous physicists of 1900s for e.g. Frédéric Joliot-Curie Dennis Gabor Wolfgang Pauli Enrico Fermi etc.	2 - Details about famous physicists of 1900s and their inventions/discoveries. 1 - Details about physicists not from 1900s. 0 - Anything other than the above two.
heat transfer devices	Anything about the devices which could be used in heat exchange or transfer and their background.	2 - Explanations about the specific types of heat transfer devices. 1 - Working of heat transfer. 0 - Anything other than the above two.

Query	Information need	Relevance criteria
Typen Atommodelle translation: types of atomic models	Documents related to description of various types of atomic models e.g. Bohr's model, plum pudding model etc.	2 - Explanations about a specific types of atomic models. 1 - Explanation about what atomic model constitutes. 0 - Anything other than the above two.
Fortschritte im Bauwesen translation: advancements in construction	Documents containing details about techniques and advancements in construction field.	2 - Explanations satisfying the information needs. 1 - Explanation about construction. 0 - Anything other than the above two.
Geschichte der Grundchemikalien translation: history of commodity chemicals	Anything about various commodity chemicals that came into existence over time.	2 - Explanations about a specific types of commodity chemicals and their existence. 1 - Explanation about what a commodity chemicals are. 0 - Anything other than the above two.
Teilchenbeschleuniger translation: particle accelerators	Anything about various types of particle accelarators and how they function.	2 - Explanations that meet the information needs. 1 - Explanation about what accelerators in general mean and similar. 0 - Anything other than the above two.
Quatum-Mechanik translation: quantum mechanics	Description of quantum mechanics, principles associated with quantum mechanics.	2 - Explanations that meet the information needs. 1 - Explanation about what mechanics is and similar. 0 - Anything other than the above two.

Query	Information need	Relevance criteria
globale Erwärmung translation: global warming	Anything about what global warming is, its effects, conferences, news controversies etc.	2 - Documents satisfying the information needs. 1 - Explanation about what warming in general is or something similar. 0 - Anything other than the above two.
Maßeinheiten translation: units of measurement	Documents about any of the units of measurements.	2 - Explanations about a specific unit of measurement. 1 - Explanation about what measurement is or what unit is. 0 - Anything other than the above two.
arten von chemischen Bindungen translation: types of chemical bond	Anything about various types of chemical bonds such as ionic covalent etc.	2 - Explanations about specific types of chemical bonds. 1 - Examples of chemical bonds or explanation of what a chemical bond is without the types. 0 - Anything other than the above two.
Euklids Prinzipien translation: euclid's principles	Anything about principles postulated or proved by Euclid.	2 - Explanations about the principles put forward by Euclid. 1 - Explanation about Euclid or about principles in general. 0 - Anything other than the above two.
Atombombe translation: atomic bomb	Anything about atomic bomb and various examples of atomic bombs e.g. dirty harry, Navajo etc.	2 - Explanations about atomic bombs or documents with examples of atomic bombs. 1 - Explanation about bombs. 0 - Anything other than the above two.

A detailed excel sheet of these queries, information needs, relevance criteria and relevance judgements for the results provided by the team-mates mocking as the retrieval model by conducting search of both Wikipedia English and German is provided in the github repository: **The excel sheets with details and the python notebooks with all the code and evaluation computation can be found here** https://github.com/rvteja24/MLIR_Project.

These relevant judgements are used to compute the following retrieval metrics over the 20 queries for our manual retrieval exercise:
Average of Normalized Discounted Cumulative Gain at rank 10: 0.79915
Mean Average Precision @ 10: 0.59105

3 Section III Extra credit

3.1 Implemented Model for the task

Based on the modelling approach proposed in section 1.1.8, we have implemented the system which can be used by users to pose queries in either German or English and all the relevant articles from the categories of Science, Technology and their histories from both the languages are returned in a single ranked list.

For this implementation we have used the Monolingual Wikipedia dump of articles for German and English from linguatools website. These dumps contain one xml per language and there are 5,690,374 articles in English and 2,212,874 articles in German. Each of the article in the xml sheet is delimited by an <article> tag. Since our focus areas are Science, Technology, and their histories we leveraged a perl script to extract only the articles which belong to these categories. Once we extracted the relevant articles we ended up with **358,392** articles in English and **263,281** articles in German.

We developed a script to pre-process this data and generate:
1.) vocabularies for each language,
2.) term to index mappings,
3.) document url to index and vice versa mappings,
4.) term frequencies for each document,
5.) term frequencies for each language collection,
6.) inverted indices.
All these statistics and mapping were stored to the disk.

We then developed a program which would take a query from user as input and does the following:
1.) Translate the query from English to German and vice-versa by using the google translate API which uses Transformer models for the translation process.
2.) Pass both these English and German queries through the same pre-processing steps as the documents and generate tokens.
3.) For any token which is not present in the respective language's vocabulary find synonyms which are present in the vocabulary, if no such synonyms are found use the dynamic programming algorithm to find decomposed tokens from the given token which minimize the number of tokens while maximizing the number of characters from the original token are utilized.
4.) Pass the English query tokens to the English Query Likelihood (QL) model and German query tokens to the German Query Likelihood model which use the inverted indices, document level term frequencies and collection level term frequencies to assign scores to the candidate documents. The μ value chosen for the German QLM is 1100 and the μ value for the English QLM is chosen to be 1300. These values are chosen based on prior works.
5.) Get top-k ranked lists of documents and their QL scores from each of these models and merge the two lists of documents retrieved. Re-rank the documents based on their QL scores and return the urls associated with the top-k documents along with their scores.

In our implementation, our system returns only the urls of the top-k relevant documents

for the user to navigate to. https://github.com/rvteja24/MLIR_Project.

3.2 Evaluation

We tested our model using 10 queries, 5 from English and 5 from German. Below are the Precision@K and Normalized Discounted Cumulative Gain computed for the model for these 10 queries based on our manual relevance judgement.
Mean Average Precision at 10: 0.44865
Mean Normalized Discounted Cummulative Gain: 0.58184.
The retrieved results and computation of these metrics are in the Retrieval Model python notebook in the Git repository. One of the reasons for the low score is due to limitation we've put on the articles used for the system. Due to this for some of the queries the model returned results which were not at all relevant and manual search on the text file containing all the documents showed that there weren't any associated documents in either of the documents.

3.3 Scope for improvement and future work

Our system uses a μ value of 1100 for German and 1300 for English QLMs, further emperical analysis could be conducted to choose most suitable values for the collections chosen by us. Also, in our approach we haven't used pseudo-relevance feedback and also solely relied on the translation module for querying across languages by keeping the collections separate. The future scope would involve exploring the usage of pseudo-relevance feedback for this system as well as more sophisticated language models instead of the uni-gram language model which works with the assumption of binary independence, which isn't always true. Also, approaches to perform combined indexing and maintaining a single collection is left as future work. Adding on to these, future work can also include dealing with multiple languages instead of just two languages.

4 Contributions

All team members contributed equally on this project.

5 References

https://trec.nist.gov/pubs/trec8/papers/trec8ov.pdf - TREC overview of CLIR track.
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.1986rep=rep1type=pdf - Precision @ K for MLIR.
https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/ - data corpus.
https://medium.com/lily-lab/a-brief-introduction-to-cross-lingual-information-retrieval-eba767fa9af6 - Intro to CLIR.
https://nlp.stanford.edu/IR-book/html/htmledition/standard-test-collections-1.html - Test collections.
https://nlp.stanford.edu/IR-book - Reference book.
https://spacy.io/ - pre-processing tools.

6 Appendix

https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/ - data.
https://github.com/rvteja24/MLIR_Project - code repo.