CSCE 5222: Feature Engineering

**TITLE: CROSS LANGUAGE INFORMATION RETRIEVAL.**

**Project Increment - 1**

**TEAM MEMBERS:**

MAYUR VORA – 11550893 – [mayuriswarbhaivora@my.unt.edu](mailto:mayuriswarbhaivora@my.unt.edu)

SHRIYA KANNOJ – 11507709 – [shriyakannoj@my.unt.edu](mailto:shriyakannoj@my.unt.edu)

**GITHUB LINK:**

<https://github.com/Mayur-vora/Feature_Engineering_fall_2022>

**VIDEO LINK:**

<https://drive.google.com/file/d/1mC-bLH-GF8uftSltfRsApTQONTX8OAPR/view?usp=share_link>

**IDEA DISCRIPTION:**

**OVERVIEW**

The main purpose of Cross-Language Information Retrieval (CLIR) is to build search engines that use a query described in one particular language (e.g., English) to ﬁnd the content that is expressed in any other language (e.g., German). Machine Translation (MT, it’ll automatically render the content of a document in different language) and CLIR are closely combined in another way; Machine Translation is the use of translation technology to render documents readable, whereas CLIR is the use of translation technology to render documents searchable. So, MT and CLIR are both built on the same foundation: TT Translation Technology. The focus should be on retrieval of documents in one language using queries in another language, in practical applications the collection to be searched can potentially contain documents in many diﬀerent languages, or even documents that themselves contain mixed languages, and the query too may contain terms from diﬀerent languages.

The purpose of this project is to develop a cross-language information retrieval system (CLIR) that can search text materials published in English and display the results in German in response to a German query. To evaluate the system's performance, we'll use machine translation, information retrieval with a vector space model, and lastly IR assessment tools. Many components of the project are detailed as we go along.

Information is increasingly being sought for not just in the user's native tongue but also in other languages. Cross-linguistic information retrieval (CLIR), which looks for pertinent material published in a language other than the one supplied in a query, is thus made possible. Translation is the main challenge in CLIR, along with the limitations of monolingual information retrieval (IR): either the query or the documents must be translated from one language to another. This translation challenge, however, differs from full-text machine translation (MT) in that it aims to offer a translation adequate for finding pertinent information rather than a translation that can be read by humans. As a result, certain translation techniques are required.

**GOALS AND OBJECTIVES:**

The aim of this project is to build a cross language information retrieval system (CLIR) which, given a query in German, will be capable of searching text documents written in English and displaying the results in German. We're going to use machine translation, information retrieval using a vector space model, and then assess the performance of the system using IR evaluation techniques. We are not sure of the techniques we use yet, but this is the conclusion we made after going through some research papers in this domain.

The goal of this project is to create a cross-language information retrieval system (CLIR) that can search text documents published in English and show the results in German given a German query. We'll employ machine translation, information retrieval with a vector space model, and finally IR assessment tools to rate the system's performance.

**MOTIVATION:**

The Motivation for this project is divided in two cases also we can use CLIR systems to solve those problems so, The following are the cases:

* cases in which the user doesn’t speak the language of the documents
* cases in which the user understands the documents but prefers to use a diﬀerent language for their query.

The given use cases, together with other factors such as frequency of their interaction, relatedness of the languages, collection size and format, or the expected number of queries then inﬂuence the design choices which need to be considered when building the CLIR system.

The motivation for this project is to gain a better grasp of the German language by utilizing the CROSS-LANGUAGE INFORMATION RETRIEVAL (CLIR) TRACK. The fundamental goal of the TREC CLIR activities has been to create a multilingual test collection that may be utilized for a range of assessment trials. As a result, the precision of relevance assessments is critical. The Twenty-One group conducted an insightful analysis of the accuracy of the evaluations and how this influenced the pool.

**SIGNIFICANCE:**

The CLIR track participants were told to gather resources from a multilingual pool that included papers in four different languages. Although they may choose the theme language, they were still required to search the pool for pertinent items regardless of the languages in which the works were produced. Most teams approached this problem by doing separate bilingual retrieval runs, analyzing the results, and combining them. These businesses faced an extra hurdle with the integration of their retrieval findings.

TREC-8 documentation was available in English, German, French, and Italian. Every single one of the 28 themes was accessible in all four languages. To attract new players, monolingual non-English runs were permitted; nevertheless, when they were unable to finish the whole challenge, individuals chose to do bilingual cross-language runs.

The TREC-8 job description additionally contained a vertical domain subtask that involved working with a second data collection of documents from a structured database in the topic of social science (the "GIRT" collection). Most of the items in this collection have English titles, and there is a multilingual thesaurus included. The University of California, Berkeley, conducted a series of highly extensive examinations on this collection.

The CLIR system will:

* **translate queries** from German into English (because our searcheable corpus is in English), using word-based translation, a rather simplistic approach as opposed to the sophistication you might see in, say, *Google Translate*.
* **search over the document corpus** using the Okapi BM25 IR ranking model, a variation of the traditional TF-IDF model.
* **evaluate the quality** of ranked retrieval results using the query relevance judgements.

**LITERATURE SURVEY:**

I have seen various pages who worked on the cross-language information retrieval who only worked till the translation of the project. Our team is planning to work till the noise cancelation and add more features for the project. In the projects that I have referred, Corpora-based approaches are also often used. The University of Montreal, IBM, and Eurospider are just a few of the companies using comparable methods in current or past CLIR tracks. Most cross-language retrieval systems draw on already-existing linguistic materials, notably bilingual dictionaries that can be read by computers. Many solutions have been proposed to address some of the problems associated with dictionary-based translations, such as ambiguity and vocabulary coverage. One group that has examined the use of these dictionaries is the Twenty-One consortium.

**OBJECTIVES:**

The difficulty that information retrieval solves is finding important information, or pertinent documents conveying that information, for a given information demand within a huge collection of documents (e.g., a large database of documents or the Web). For example, a customer may want to use the Internet to learn more about a product (such as the iPod's memory capacity). This material is useful if it appears in a paper. The inability to directly access and change some bits of information makes this problem incredibly difficult. It is defined by what we access and modify. This is not an original description. A document may include many descriptions of the same information, such as "there is a major earthquake in China in 2008," using a range of media (text, picture, video, audio, etc.) and languages (natural or formal). The user's information demands can also be described in several other ways. The main objective of this project is to clear these issues with no erroneous errors.

**FEATURES:**

* Housekeeping: File encodings and tokenization[¶](http://localhost:8888/notebooks/Downloads/Cross%20Language%20Information%20Retrieval.ipynb#Housekeeping:-File-encodings-and-tokenisation)
* Test the tokenize function.
* build a CLIR engine consisting of information retrieval and translation components, and then evaluate its accuracy.
* Information Retrieval using [Okapi BM25]
* BM25 TF-IDF representation
* we'll create the td-idf matrix for terms-documents, first without the query component.
* Query translation.
* Calculating the unigram, bigram and trigram counts.
* Creating Translation model.
* Combining and evaluation.
* Noisy channel translation.

**RELATED WORK (BACKGROUND):**

There are three main ways in which cross-language information retrieval approaches attempt to "cross the language barrier" – through query translation, or document translation, or both.

CLIR research started out with experiments using controlled vocabularies and associated dictionaries and thesauri, but nowadays free text approaches are most common.

These approaches also dominate experiments in past and present CLIR tracks. Free text methods can be further classified according to the resources used to cross the language boundary: machine translation, machine-readable dictionaries, or corpus-based resources.

Machine translation (MT) seems an obvious choice for cross-language information retrieval systems. It also played a large role in the TREC-8 experiments of a number of groups.

However, CLIR is a difficult problem to solve on the basis of MT alone: queries that users typically enter into a retrieval system are rarely complete sentences and provide little context for sense disambiguation.

Corpus-based approaches are also popular. Groups experimenting with such approaches during this or former CLIR tracks include Euro spider, IBM and the University of Montreal.

Lastly, a significant number of cross-language retrieval approaches make use of existing linguistic resources, mainly machine-readable bilingual dictionaries. Various ideas have been proposed to address some of the problems associated with dictionary-based translations, such as ambiguities and vocabulary coverage. One of the groups that have investigated the use of such dictionaries is the Twenty-One consortium.

Here, We're going to use machine translation, information retrieval using a vector space model, and then assess the performance of the system using IR evaluation techniques.

**DATASET:**

* bitext.(en,de): A sentence aligned, parallel German-English corpus, sourced from the Europarl corpus (which is a collection of debates held in the EU parliament over a number of years). We'll use this to develop word-alignment tools, and build a translation probability table.
* newstest.(en,de): A separate, smaller parallel corpus for evaluation of the translation system.
* devel.(docs,queries,qrel): A set of documents in English (sourced from Wikipedia), queries in German, and relevance judgement scores for each query-document pair.

The files are available to check out in the data/clir directory of the repository.

**DETAIL DESIGN OF FEATURES :**

As mentioned earlier, we're going to build a CLIR engine consisting of information retrieval and translation components, and then evaluate its accuracy.

The CLIR system will:

* Translate queries from German into English (because our searchable corpus is in English), using word-based translation, a rather simplistic approach as opposed to the sophistication you might see in, say, Google Translate.
* Search over the document corpus using the Okapi BM25 IR ranking model, a variation of the traditional TF-IDF model.
* Evaluate the quality of ranked retrieval results using the query relevance judgements.

So for our first task, we'll load the devel.docs file, extract and tokenize the terms, and store them in a python dictionary with the document ids as keys.

Now we'll build an inverted index for the documents, so that we can quickly access documents for the terms we need.

On to the BM25 TF-IDF representation, we'll create the td-idf matrix for terms-documents, first without the query component. The query component is dependent on the terms in our query. So we'll just calculate that, and multiply it with the overall score when we want to retrieve documents for a particular query.

We have designed several features for query translation and language model as a part of the project.

As we'll train the model on different files, it's obvious that we'll run into words (unigrams) and trigrams what we hadn't seen in the file we trained the model on. To account for these unknown information, we'll use add-k or Laplace smoothing for the unigram and Katz-Backoff smoothing for the trigram model.

Also we calculated perplexity of our language model.

**ANALYSYS:**

Since we are working on German data there are not much things to visualize so here’s what we have done for analysis part

We are performing information retrieval using Okapi BM25

Which is a ranking function used by various search engines to estimate the relevance in queries. We'll start by building an IR system, and give it a test run with some English queries.

Here's an overview of the tasks involved:

* Loading the data files, and tokenizing the input.
* Preprocessing the lexicon by stemming, removing stop words.
* Calculating the TF/IDF representation for all documents in our Wikipedia corpus.
* Storing an inverted index to efficiently documents, given a query term.
* Implementing querying with BM25.
* Test runs.

We performed tests on various models out of all trigram model performs better than unigram or bigram models.

**IMPLEMENTATION:**

As we mentioned all the tasks we have to go through, I am denoting the tasks that we have completed for this first evaluation.

Also we will attach the screenshots of the results of the implementation in the Preliminary results section of this report

As per the overview The aim of this project is to build a cross language information retrieval system (CLIR) which, given a query in German, will be capable of searching text documents written in English and displaying the results in German.

We're going to use machine translation, information retrieval using a vector space model, and then assess the performance of the system using IR evaluation techniques.

At first we did file encoding and tokenization and Since the data files we use is utf-8 encoded text, we need to convert the strings into ASCII by escaping the special symbols. We also import some libraries in this step as well.

Then we performed IR operation using Okapi BM25 function. Here’s the detailed steps for that operation

* Loading the data files, and tokenizing the input.
* Preprocessing the lexicon by stemming, removing stop words.
* Calculating the TF/IDF representation for all documents in our Wikipedia corpus.
* Storing an inverted index to efficiently documents, given a query term.
* Implementing querying with BM25.
* Test runs.

So for our first task, we load the devel.docs file, extract and tokenize the terms, and store them in a python dictionary with the document ids as keys.

Then we build an inverted index for the documents, so that we can quickly access documents for the terms we need.

Project Management:

**WORK COMPLETED:**

We figured out what steps we need to go through to get the desired results for cross language information retrieval

Here we mentioning tasks which has been completed

We found a dataset which has all the requirements for our cross language queries, also we have gone through various research papers which showed us the path and tasks view what we should do step by step.

1. We performed data preprocessing and merging data to get desired dataset to work on and increase the size of the dataset
2. Gathered information about Okapi BM25 function through research papers and implemented queries with this function and run various tests
3. After getting many errors we created some simplified features which can make the process smooth in future use like inverting index for the documents so it can be easily accessed
4. We created tf-idf matrix for terms-documents
5. We ran a language model. As we train the model on different files, it's obvious that we run into words (unigrams) and trigrams what we hadn't seen in the file we trained the model on.
6. The sentences are also converted appropriately by adding sentinels at the start and end of sentences.
7. We calculate the perplexity for the model, as a measure of performance also used Using add-k smoothing.

**Responsibilities:**

* Shriya:
  + Documentation of reports
  + Researching about features and implementation
  + Worked on tokenization
  + Meeting every weekend and task assigning
* Mayur:
  + Work on Okapi BM25 Function
  + Building Language Model
  + Create TF-Idf matrix
  + Make functions to calculate perplexity of n-grams

**Contribution Percentage:**

This team worked together and hard for better performance of the code and the linear progress in the project in the short span of 3-4 weeks.

The project is about 75% finished and team members contributed almost equally.

CODE:

MAYUR VORA - 65%

SHRIYA KANNOJ – 35%

DOCUMENTATION:

MAYUR VORA – 50%

SHRIYA KANNOJ – 50%

**WORK TO BE COMPLETED:**

For the next iteration we have many tasks left.

Right now we know that we need to run the translation model which can be a huge task and after that combining all and perform evaluation for the project

We found out that for translation model, we'll estimate translation model probabilities. For this, we'll use IBM1 from the NLTK library. IBM1 learns word based translation probabilities using expectation maximization.

**Responsibilities:**

We have not yet decided the responsibilities for the second iterations yet but we just got the way what we need to implement so we will work on that.

**Issues and Concerns:**

The issues are we still not sure for translation model what other steps we might need to perform regarding the problem statement.

We might need to create translation dictionaries in both English to German, and German to English directions based on this model and which is pretty challenging. Also then dictionary for probabilities.

We just trying our best to get something from the data we have and we hope it will give better accuracy.

**REFERENCES:**

# Some example project ideas are follows:

<https://trec.nist.gov/pubs/trec8/papers/trec8ov.pdf>

<http://www.columbia.edu/~ph2046/docs/Hackett/trec97.pdf>

<https://terpconnect.umd.edu/~oard/pdf/elis09.pdf>

<https://aclanthology.org/W99-0605.pdf>

<https://www.semanticscholar.org/paper/Cross-Language-Information-Retrieval-Li-Xu/0a51ae1d9a972b9bd6fc5c7efdaf74a5eee5d801>

<https://www.semanticscholar.org/paper/Embedded-Fuzzy-Bilingual-Dictionary-model-for-cross-Onifade-Ibitoye/b0dbbb281f95e267e81511db9c495497d290df22>

<https://www.semanticscholar.org/paper/Opening-Machine-Translation-Black-Box-for-Retrieval-Ma-Nie/16e7da75e5a41c7b77c2611aa9d29abe98db28f6>

<https://www.semanticscholar.org/paper/Query-dependent-learning-to-rank-for-cross-lingual-Ghanbari-Shakery/fc97719caf308575c5e421262df073e44555ce4f>

<https://www.semanticscholar.org/paper/A-learning-to-rank-approach-for-cross-language-Azarbonyad-Shakery/ed079f51f05972fa7010875360de54b4cada42df>