Semantics-based, Multilingual Scientific Text and Book Recommendation System

Naumov Vitalii

Eötvös Lorand University, Budapest, Hungary
fy46in@gmail.com

Abstract. Recommending books for users who speak more languages than one, faces several challenges such that, for example, a certain book might not be translated in some languages, or might be freely available in only one language, etc. For example, a journalist or investigator, governmental institution, market analyst might want to find documents similar to a given one but in different language or companies that are constantly watching what people are writing about them or about their new product. Widespread information retrieval systems are generally monolingual and therefore incapable of satisfying these users information needs.

Keywords: Information retrieval, Cross-lingual information retrieval, NLP

1 Introduction

In the modern world the most valuable resource is information. Unsurprisingly some of the first electronic machines that humans built was about processing, acquiring and transmitting information. And this has not changed ever since.
People produce a lot of information every day — they publish their researches with publications, share their experience in books, etc. Of the millions of textual documents generated daily only a limited amount is available in a given language and this language diversity of the documents poses a great challenge in information retrieval. This presents a problem for users who are in the need of information in multiple languages.
Recommending books for users who speak more languages than one, faces several challenges such that a certain book might not be translated in some languages, or might be freely available in only one language, etc. [1] For example, a journalist or investigator, governmental institution, market analyst might want to find documents similar to a given one but in different language or companies that are constantly watching what people are writing about them or about their new product. Widespread information retrieval systems are generally monolingual and therefore incapable of satisfying these users information needs.
Another aspect of the problem of information retrieval that most of existing algorithms that are capable of multilingual document search is that they mostly look into the translated to target language topic. Therefore, there exists a problem and to resolve it we have to take into account the actual text inside these documents.
The aim of this work is to have a look deeper into this field and try to implement a recommendation system for English, German, Dutch, Italian, Portuguese, Polish,

Spanish and French languages which might help to ease the difficulty during information retrieval for scientists from all fields. To do this I first overview the more general field of information retrieval and learn how the concepts from those are adapted to the multilingual scope.

2 Information Retrieval

2.1 Overview of Information Retrieval

Information retrieval (IR) is the process of finding information in an unstructured source of data where unstructured means that it is hard or impossible to translate it for a computer. The data that is mostly the subject of information retrieval are text documents, but other types such as images, videos or audio are also possible targets. The most prominent usage of IR consists of finding documents or document fragments relevant to a query. A query is a short (or not so short) piece of text that describes the users, information need. The goal of an IR system is to return the most relevant documents to the query from a collection of documents. The system achieves its purpose if the returned documents satisfy the information need of the user. Another application of information retrieval is recommendation based on the content of the documents. In this use-case the document in question constitutes the query and the goal is to find relevant documents to it in the collection.

The task of information retrieval can be defined as follows. Given a set of documents $D = \{d_1, d_2, ..., d_n\}$ and a query document q we want to find and retrieve the most relevant documents from D to query q.

We say two documents are relevant to each other if their content is similar. In text if documents are about the same topic, in images if it is about the same object, etc. So, in order to return relevant documents to a query we have to be able to compare them. The challenge in this is the unstructured nature of the data. It is very easy to compare structured objects such as vectors, but there is no easy way for a computer to compare two ordered set of words with meaning that constitutes a document.

2.2 Cross-Lingual Information Retrieval

The task of Cross-Lingual Information Retrieval can be defined as follows. Given a language set $L = \{l_1, ..., l_p\}$, a set of documents in each language $D_{l_i} = \{d_{l_i,1}, ..., d_{l_i,n_i}\}$ and query document q written in one of the languages; find the most relevant documents to the query from all the D_{l_i} (i = 1, ..., p) document sets.

The task of CL-IR takes the challenge of information retrieval to another level. Not only we have to find relevant documents in the same language, but we also have to deal with the language diversity of the documents. [2] Three main approaches have emerged: bilingual dictionary based, machine translation based, and inter-lingual representation based models. In the current research I will be using pre-trained inter-lingual representation based model from Google. It was developed back in 2019, consists of 16 languages and shows a decent performance on cross-lingual retrieval.

3 The model

3.1 Model description

This chapter describes the model used in the experiments. As it has been previously mentioned because of the difficulties and drawbacks of other methods this work focuses on model based on some kind of inter-lingual representations of the documents. Two models will be implemented: one is for information retrieval within one language for both topic and text search and the other one is for information retrieval within all presented languages for both topic and text search.

As was mentioned earlier, the pre-trained model which will be used to find embeddings of the sentences is already exists, but to actually navigate between this ambiguous corpus of them, we have to perform nearest-neighbor lookups, since nearest-neighbor lookups are a quick way to find the items in the data set that are closest (or most similar to) any other item in your data. Simple Neighbors package was selected since it uses one of a handful of libraries behind the scenes to provide approximate nearest-neighbor lookups, which are ultimately a little less accurate than pairwise calculations but much faster.

The next step is to create the first model which will propose results within one language. All I had to do is to add calculated embeddings to the index file and then, build index trees for each language with a given number of those trees.

After that, the second multilingual model will be defined. The steps will be the same as for the previous model, except for the step, where we will add annotated sentence with a language code to the combined index.

The final step is set to verify the proposed results by the model. In this step I will analyze how the proposed results for each language will be similar to the passed query. To do that, cosine similarities will be calculated, and using that data will be handy to have a look, how different languages are similar to each other within the dataset which will be described in the next chapter.

3.2 Packages used

The models and the data processing were implemented in Python 3.8. Python is the most popular programming language of scientific computing and machine learning because of it easy to use and abundant number of packages for the task. Libraries such as Scikit-Learn, NumPy, ElementTree, or Pandas offer open source implementations of many tasks and algorithms related to these areas and the large community around them.

NumPy and SciPy are by far the most fundamental packages used in scientific Python. NumPy is the implementation of everything that is matrix or vector related. It offers an interface similar to MATLAB which makes it easy and intuitive to use while most of the functions are implemented externally in C/C++ which makes it very fast and efficient.

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data

structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Scikit-Learn is an open-source machine learning (ML) library maintained by the booming ML community of Python. It is built upon NumPy and SciPy and implements hundreds of algorithms for the whole range of tasks of a Machine Learning engineer. From preprocessing to cross validation to the most popular classification and clustering algorithms everything can be found in it.

XML is an inherently hierarchical data format, and the most natural way to represent it is with a tree. ET has two classes for this purpose - ElementTree represents the whole XML document as a tree, and Element represents a single node in this tree. Interactions with the whole document (reading and writing to/from files) are usually done on the ElementTree level. Interactions with a single XML element and its sub-elements are done on the Element level.

3.3 The dataset

To conduct experiments, first we need something to run our models on data. Data in our case has to be parallel documents in multiple languages. This is much harder to come by than regular documents for other text mining applications such as text categorization or clustering.

The largest source of parallel documents are international organizations and governments that are obliged to translate their documents to multiple languages. One of these organizations is the European Union where every piece of legislation and official document has to be translated to 27 languages. These professional translations are often used for cross-lingual and multilingual tasks such as machine translation or in our case cross-lingual document linking because of their accessibility, quantity and quality.

The JRC-Acquis corpus consists of the body of the European Union’s law and legislation papers. It was put together by an international group of researchers for the purpose of making the largest multilingual corpora freely available at the time. It contains documents from 1950 until the present and it is updated every few years as new legislations are passed in the EU. The documents are not only parallel but also sentence aligned. Therefore, the corpus is mainly used for machine translation projects but because of its size and number of languages contained it is a popular dataset for every kind of text mining application.

3.4 The preprocessing

After loading in the documents for the task the preprocessing stage begins. The dataset mentioned earlier does exist in xml format, so to load the data in an appropriate format into model, it is essential to get rid of all XML format specific

symbols and to leave only a title, a year of publishing and a text for each document. For this purpose, I wrote a script which takes all necessary data I mentioned before and puts it in a pandas dataframe.

To get rid of complications in the code, the next step was set to split the dataframe into sub-dataframes for each language, which I split into separate data frames for texts and titles with respect to the language.

Then to each dataframe above will be added one more column which will contain an embeddings array for each sentence. This data is essential since it will be necessary to compute cosine similarities and mate retrieval rate.

4 Experiments

4.1 Metrics

The only document that we are absolutely certain is relevant is the pair of the document in the other language. Therefore, the mainly used metric in cross-lingual information retrieval tasks is the mate retrieval rate which is the proportion of test documents that are linked with their cross language pair. This can be formulated as the number of rows/columns in the test similarity matrix, where the diagonal element is the largest. [3]

Mean reciprocal rank was selected since this is a performance measure commonly used in tasks where there is only one relevant document to retrieve. Define the rank of document i denoted by r_i as the order in which its parallel pair is retrieved. The mean reciprocal rank is then:

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{1}{r_i}$$

4.2 Results of monolingual model

This section details the results of the conducted experiments. The model which is used for an information retrieval within one language has a parameter defining the number of index trees which is set to 40. The section will be relatively short since this model is not the main goal of the work. The function of the model takes three arguments: a query itself, number of results it has to show and the index language. The examples of results of the function for English and French queries can be seen on Figure 1 and Figure 2 respectfully.

```
#for english sentence
find_same('Rules of the Advisory Committee on Vocational Training', 20, 'en')
```

```
["'Council Directive 66/402/EEC of 14 June 1966 on the marketing of cereal seed  /* CODIFIED VERSION CF 374Y0608(03) */'",
 "'Regulation No 7/63/Euratom of the Council of 3 December 1963 on rules of procedure of the Arbitration Committee provided for in Article 18 of the Treaty establishing the European Ato
 "'Regulation No 136/66/EEC of the Council of 22 September 1966 on the establishment of a common organisation of the market in oils and fats'",
 "'Regulation No 423/67/EEC",
 "'Council Directive 68/221/EEC of 30 April 1968 on a common method for calculating the average rates provided for in Article 97 of the Treaty'",
 "'EEC Council: Regulation No 17: First Regulation implementing Articles 85 and 86 of the Treaty'",
 "'Council Directive 68/297/EEC of 19 July 1968 on the standardisation of provisions regarding the duty-free admission of fuel contained in the fuel tanks of commercial motor vehicles'"
 "'EEC: Council Directive on the approximation of the rules of the Member States concerning the colouring matters authorized for use in foodstuffs intended for human consumption'",
 "'Council Directive 64/221/EEC of 25 February 1964 on the co-ordination of special measures concerning the movement and residence of foreign nationals which are justified on grounds of
 "'Council Directive 66/401/EEC of 14 June 1966 on the marketing of fodder plant seed  /* CODIFIED VERSION CF 374Y0608(02) */'",
 "'Council Directive 64/433/EEC of 26 June 1964 on health problems affecting intra-Community trade in fresh meat  /* CONSOLIDATED VERSION SEE 375Y0820(02) */'",
 "'Regulation (EEC) No 1017/68 of the Council of 19 July 1968 applying rules of competition to transport by rail",
 "'Regulation (EEC) No 2264/69 of the Commission of 13 November 1969 on applications for reimbursement of aid granted by Member States to organisations of fruit and vegetable producers'
 "'Council Directive 68/193/EEC of 9 April 1968 on the marketing of material for the vegetative propagation of the vine'",
 "'First Council Directive 67/227/EEC of 11 April 1967 on the harmonisation of legislation of Member States concerning turnover taxes'",
 "'66/399/EEC: Council Decision of 14 June 1966 setting up a Standing Committee on Seeds and Propagating Material for Agriculture",
 "'63/266/EEC: Council Decision of 2 April 1963 laying down general principles for implementing a common vocational training policy'",
 "'Regulation (EEC) No 2146/68 of the Council of 20 December 1968 amending Regulation No 136/66/EEC on the establishment of a common organisation of the market in oils and fats'",
 "'Council Directive 69/169/EEC of 28 May 1969 on the harmonisation of provisions laid down by Law",
 "'EEC Council: Regulation No 1 determining the languages to be used by the European Economic Community'"]
```

Fig. 1. English monolingual results

```
#for dutch sentence
find_same("Voorzitter van het Economisch en Sociaal Comité", 20, 'nl')
```

```
["'Richtlijn 64/432/EEG van de Raad van 26 juni 1964 inzake veterinairrechtelijke vraagstukken op het gebied van het intracommunautaire handelsverkeer in runderen en varkens  /* GECODIF
 "'68/416/EEG: Beschikking van de Raad van 20 december 1968 betreffende het sluiten en uitvoeren van de speciale intergouvernementele overeenkomsten inzake de verplichting voor de Lid-S
 "'Richtlijn 64/225/EEG van de Raad van 25 februari 1964 ter opheffing van de beperkingen van de vrijheid van vestiging en van het vrij verrichten van diensten",
 "'Verordening nr. 423/67/EEG",
 "'Richtlijn 68/414/EEG van de Raad van 20 december 1968 houdende verplichting voor de Lid-Staten van de E.E.G. om minimumvoorraden ruwe aardolie en/of aardolieprodukten in opslag te ho
 "'Verordening (EEG) nr. 1108/70 van de Raad van 4 juni 1970 betreffende de invoering van een boekhouding van de uitgaven voor de wegen voor het vervoer per spoor",
 "'Verordening (EEG) nr. 448/69 van de Raad van 11 maart 1969 tot wijziging van Verordening (EEG) nr. 315/68 houdende vaststelling van kwaliteitsnormen voor bloembollen en bloemknollen'
 "'EEG Raad: Verordening Nr. 26 inzake de toepassing van bepaalde regels betreffende de mededinging op de voortbrenging van en de handel in landbouwprodukten'",
 "'Verordening (EEG) nr. 316/68 van de Raad van 12 maart 1968 houdende vaststelling van kwaliteitsnormen voor verse snijbloemen en vers snijgroen'",
 "'63/688/EEG: Statuut van het Raadgevend Comité voor de beroepsopleiding'",
 "'Verordening (EEG) nr. 2146/68 van de Raad van 20 december 1968 tot wijziging van Verordening nr. 136/66/EEG houdende de totstandbrenging van een gemeenschappelijke ordening der markt
 "'EGA Raad: Richtlijn inzake de vrije toegang tot gekwalificeerde arbeid op het gebied van de kernenergie'",
 "'Verordening (EEG) nr. 784/68 van de Commissie van 26 juni 1968 betreffende de wijze van berekening van de c.i.f. -prijzen voor witte suiker en ruwe suiker'",
 "'Richtlijn 69/60/EEG van de Raad van 18 februari 1969 houdende wijziging van de Richtlijn van de Raad van 14 juni 1966 betreffende het in de handel brengen van zaaigranen'",
 "'EGA Raad: Verordening N°3 ter toepassing van artikel 24 van het Verdrag tot oprichting van de Europese Gemeenschap voor Atoomenergie'",
 "'Verordening nr. 7/63/Euratom van de Raad van 3 december 1963 betreffende het reglement van de Arbitragecommissie bedoeld in artikel 18 van het Verdrag tot oprichting van de Europese
 "'Verordening (EEG) nr. 2264/69 van de Commissie van 13 november 1969 betreffende de verzoeken om vergoeding van de door de Lid-Staten aan de verenigingen van groenten- en fruittelers
 "'Richtlijn 68/193/EEG van de Raad van 9 april 1968 betreffende het in de handel brengen van vegetatief teeltmateriaal voor wijnstokken'",
 "'Richtlijn 68/419/EEG van de Raad van 20 december 1968 houdende derde wijziging van de Richtlijn van de Raad betreffende de aanpassing van de wettelijke voorschriften van de Lid-State
 "'Verordening nr. 741/67/EEG van de Raad van 24 oktober 1967 betreffende de bijstand door het Europees Oriëntatie- en Garantiefonds voor de Landbouw"]
```

Fig. 2. French monolingual results

4.3 Results of multilingual model

This model for CL-IR has the same parameter number of index trees, but the number of these is set to 60 since it is a more sophisticated method.

For this model the previous function was modified much. The function of the model still takes same three arguments: a query itself, number of results it has to show and the index language. But in the result, it shows not only the specified amount, of sentences, but also calculates cosine similarities and draws mean reciprocal rank for all other languages that are present in the given by the model query. An example of the output of the function is shown on the Figure 3.

Fig. 3. Result of multilingual search function

Now, let's have a closer look how that function works on all languages one by one. We will start with an English query as an input. As you can see on the picture above, with the given data for training, the highest score surprisingly got a Portuguese language. Here comes the rest:

– For a query in French, the highest score apart from French itself has received Polish language. A Spanish column is empty since in the results pushed by model there is no Spanish sentences. See Figure 4.

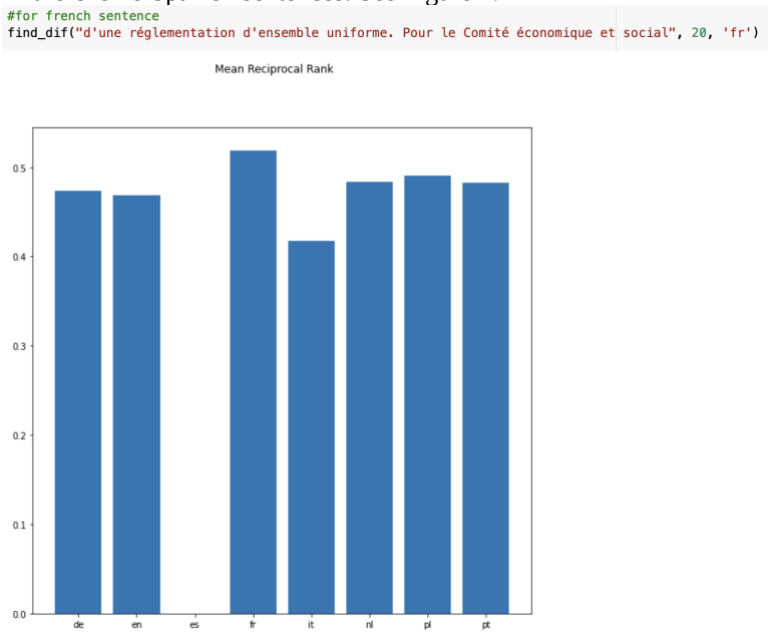

Fig. 4. Mean Reciprocal Rank for French language

– For a query in Portuguese, the highest score apart from Portuguese itself has received Dutch language. An Italian column is empty since in the results pushed by model there is no Italian sentences. See Figure 5.

```
#for portuguese sentence
find_dif('após parecer da Comissão Paritária e tomando em consideração a competência', 20, 'pt')
```

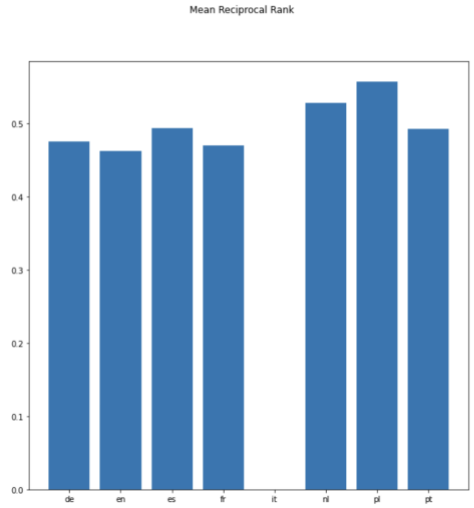

Fig. 5. Mean Reciprocal Rank for Portuguese language

– For a query in Italian, the highest score apart from Italian itself has received Dutch language. An English column is empty since in the results pushed by model there is no English sentences. See Figure 6.

```
#for italian sentence
find_dif('SU PROPOSTA DELLA COMMISSIONE', 20, 'it')
```

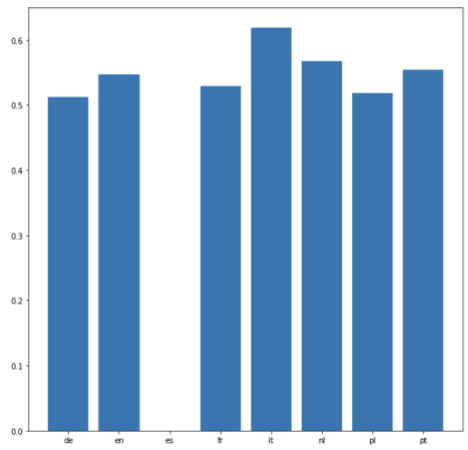

Fig. 6. Mean Reciprocal Rank for Italian language

– For a query in Spanish, the highest score apart from Spanish itself has received French language. A German column is empty since in the results pushed by model there is no German sentences. See Figure 7.

```
#for spanish sentence
find_dif(' visto el Reglamento n º 15 del Consejo', 20, 'es')
```

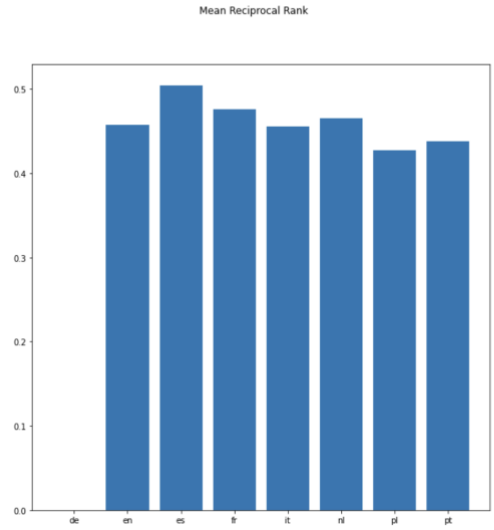

Fig. 7. Mean Reciprocal Rank for Spanish language

– For a query in German, the highest score apart from German itself has received Italian language. See Figure 8.

```
#for german sentence
find_dif(' Die zum Zeitpunkt des Inkrafttretens dieser Verordnung geltenden Geschäftsordnungen des Beratenden Ausschusses und des Fachausschusses werden weiter angewandt.', 20, 'de')
```

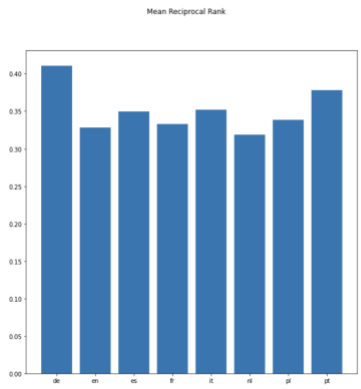

Fig. 8. Mean Reciprocal Rank for German language

– For a query in Polish, the highest score apart from Polish itself has received Dutch language. A Spanish column is empty since in the results pushed by model there is no Spanish sentences. See Figure 9.

```
#for polish sentence
find_dif('po zasięgnięciu opinii Wspólnego Komitetu', 20, 'pl')
```

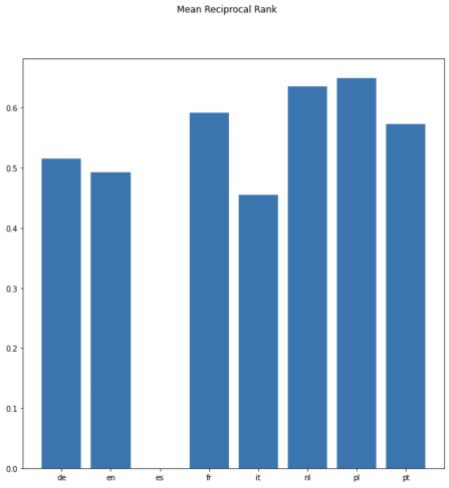

Fig. 9. Mean Reciprocal Rank for Polish language

– For a query in Dutch, the highest score apart from Dutch itself has received German language. Spanish and Italian columns are empty since in the results pushed by model there is no German, nor Italian sentences. See Figure 10.

```
#for dutch sentence
find_dif("Voorzitter van het Economisch en Sociaal Comité", 20, 'nl')
```

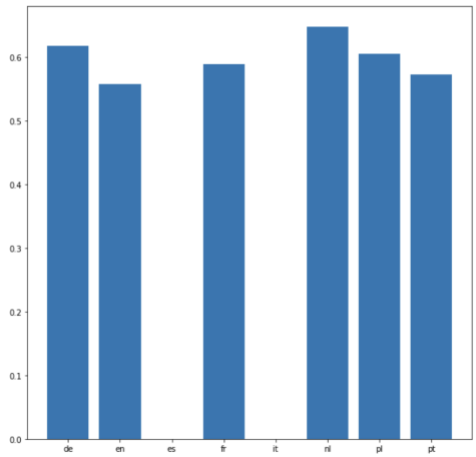

Fig. 10. Mean Reciprocal Rank for Dutch language

5 Summary

The aim of the thesis was to review the problem of cross-lingual information retrieval and document recommendation and to contribute to the field with novel methods that show promising results for further investigation.

As you could see in previous chapters, with the achieved results we can already make some hypotheses about similarity of groups of languages. Knowing the fact of how similar a pair of languages, we might adjust the parameters during model creation. Because these experiments were limited in resources for computation, further investigations are needed for the model, but the results are already promising. The future of the project involves experiments on more data and comparison with other baseline and state-of-the-art models.

References

1. Salamon V.T.: Content based recommendation in catalogues of multilingual documents (2018).
2. W. Cox and B. Pincombe. Cross - lingual latent semantic analysis. ANZIAM Journal, 48:1054–1074, dec 2006.
3. M. Yano, J. D. Penn, G. Konidaris, and A. T. Patera. Matrices and Least- Squares. 2012.
4. LNCS Homepage, http://www.springer.com/lncs, last accessed 2016/11/21.