Multi-Lingual LSA with Serbian and Croatian: An Investigative Case Study

Colin Layfield[1(✉)], Dragan Ivanović[2], and Joel Azzopardi[1]

[1] University of Malta, Msida, Malta
{colin.layfield,joel.azzopardi}@um.edu.mt
[2] University of Novi Sad, Novi Sad, Serbia
dragan.ivanovic@uns.ac.rs

Abstract. One of the challenges in information retrieval is attempting to search a corpus of documents that may contain multiple languages. This exploratory study expands upon earlier research employing Latent Semantic Analysis (so called Multi-Lingual Latent Semantic Indexing, or ML-LSI/LSA). We experiment using this approach, and a new one, in a multi-lingual context utilising two similar languages, namely Serbian and Croatian. Traditionally, with an LSA approach, a parallel corpus would be needed in order to train the system by combining identical documents in two languages into one document. We repeat that approach and also experiment with creating a semantic space using the parallel corpus on its own without merging the documents together to test the hypothesis that, with very similar languages, the merging of documents may not be required for good results.

1 Introduction

The classic Information Retrieval (IR) scenario generally consists of a corpus of documents in the same language that can be searched via queries in the same language. With the far-reaching implications of the Internet, specifically the World Wide Web (WWW), encountering documents in many different languages is no longer unusual. The same reasoning applies to repositories of information as with easier distribution of new information the need naturally arises to extend IR technologies to cater for this. Much research has already been carried out regarding the task of Cross Language Information Retrieval (CLIR - querying in one language and possibly retrieving results in several languages) and Multi-Language Information Retrieval (MLIR - asking questions in more than one language and retrieving results from documents that are in potentially different languages) [5].

The scenario we are examining seeks to exploit the strong similarity between languages in the Balkan region; in this case specifically Croatian and Serbian. Some researchers are actually engaged in investigating the hypothesis that these languages are actually the same. The main intention of the project Languages and Nationalism (http://jezicinacionalizmi.com/) is to start an open dialogue between linguists and other experts about the existence of four "political" languages in Bosnia and Herzegovina, Montenegro, Croatia and Serbia which have

J. Szymański and Y. Velegrakis (Eds.): IKC 2017, LNCS 10546, pp. 155–164, 2018.
https://doi.org/10.1007/978-3-319-74497-1_15

grown from the one language of the former Yugoslavia (the Serbo-Croatian language). The whole question of the status of Serbian and Croatian languages is very sensitive, because of cultural and political implications [8]. However, the outside linguists suggest it is the same language with multiple dialects taking into account the numerous shared features and easy mutual comprehension [3]. Due to the similarity of these languages the research question being investigated is whether or not LSA can be utilised in a multi-linguistic capacity to enable IR from a mixed language corpus of documents without the requirement of merging parallel documents - specifically those which are very similar.

2 Background

This section will first take a brief look at what methods are utilised already in the field of multi-lingual IR. Next a description of Latent Semantic Analysis (LSA) will be presented followed by a description of how this has been applied to the MLIR problem in the past.

2.1 Methods of Multi-Lingual Information Retrieval

As alluded to previously there has already been a considerable amount of research investigating multi-lingual IR approaches. Several reviews on methods used for tackling these issues have been published in the last few years that give an overview on the techniques used for this problem [5,7,11]. The approaches utilised generally fall into the categories of query and document translation. The idea behind these approaches is to employ various techniques (such as dictionary, machine translation or parallel corpora [5]) to translate either the query or the documents in the corpora themselves in order to allow IR to take place. [11] highlights challenges in this area to include translation disambiguation, quality/comprehensiveness of dictionaries employed and usage of technical terms.

2.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a technique whose roots are found in factor analysis. The idea behind LSA is to generate a classic $m \times n$ term by document matrix (which we call A) from a corpus of documents. This matrix is then reduced into three factors via *singular value decomposition* such that:

$$A = TSD^T \tag{1}$$

where T is a $m \times r$ matrix, D^T is a $r \times n$ matrix where each of these matricies has its vectors comprised of orthonormal eigenvectors, and a $r \times r$ diagonal matrix consisting of the singular values of the decomposition (r also represents the *rank* of the matrix A) [4]. The rows of matrix T represent the terms in the newly created *semantic space* and the columns of D^T represent the documents. These document vectors can now be used for similarity comparisons between one another or with additional documents (or queries) that are projected into the semantic space by multiplying the new document vector by TS^{-1}; this operation is also referred to as *folding-in* a document into the preexisting semantic space). Comparisons between documents (vectors) in this study will be done using the popular cosine operator (further details on LSA and the mechanics can be found in [1,4,9]).

The power of LSA is derived from the ability to perform *dimensionality reduction*. To do this we pick a value, say k, which is less than r and reduce the dimensionality of the singular value matrix such that it now only has k dimensions. The end result is that both term and document vectors are now represented by k dimensions. This reduced vector space is then used in comparison operations. This reduced dimensionality reduces noise and brings to the foreground any underlying "latent semantics" present in the corpus of documents [4]. The selection of k is somewhat arbitrary but a value of 300 was used in this work (as typically used in research). A good introduction to LSA can be found in [9].

2.3 Multi-Lingual Latent Semantic Analysis

Typically if a multi-lingual LSA solution is going to be employed it is generally required that a parallel training corpus is used. The parallel corpus will contain copies of the same document in the target languages - so in our case it will contain a set of identical documents in both Serbian and Croatian. The next step will be to combine these identical documents together such that each document contains the text for both languages. A term by document matrix would be generated from this combined corpus and SVD would be applied to it as described above in Sect. 2.2.

[6] carried out a study utilising ML-LSA with this approach. The parallel corpus they utilised consisted of the Hansard Collection which is a collection of Canadian parliament interactions. As Canada is a bilingual country the collection of the exchanges exists in both official languages, namely English and French. The evaluation metric they used was to perform a 'mate-retrieval' test. A mate-retrieval test involves selecting the same document in both languages (that does not exist in the semantic space) and folding one into the LSA sematnic space and using the remaining document to 'search' for its mate with the document itself acting as the query. This dataset consisted of 2,482 documents. 1,500 were selected randomly for training the system (French/English documents combined and a semantic space generated). The remaining 982 were used as a test set for performing the evaluation. The test set documents were folded into the semantic space and the English document was used to search for its French mate and vice versa. The average success rate of this test was 98.4%.

A similar study was carried out by [13] using the Bible as a parallel corpus (many researchers use this as it is one of the most translated works in the world and is split up nicely to utilise as a parallel corpus). They try two approaches, one involves creating a common base of terms to generate a cross language semantic space, the second combines small portions of each document across the entire database (of the Bible). The languages used were English and Greek.

It is difficult to compare results directly as it could be the case that the document being searched for (in the other language) was already included in the semantic space and, thus, in the SVD computation which we are not doing. One conclusion reached, however, is that a good variety of documents included in the semantic space did seem to increase their precision and recall scores.

Research was carried out by [2] regarding the effects of language relatedness on multilingual information retrieval (with a focus on Indo-European and Semitic languages). In essence they experimented, using LSA in a similar fashion to the above (including usage of the Bible to create the semantic space but the Koran text was used for search), the effect of combining parallel corpus documents from multiple languages and whether or not the language "type" had any type of positive influence. They concluded that adding additional languages from a parallel corpus can be beneficial and the types of languages added can have an impact on the quality of results.

3 Experiment

The hypothesis we are testing in our experiment is that due to the similarity of Croatian and Serbian we may be able to use a semantic space that is comprised of documents from each instead of using the typical ML-LSA approach of concatenating documents from a parallel corpus. We will be experimenting with both scenarios using documents from the parallel corpus.

3.1 Dataset Used

For this experiment we needed a parallel corpus of Serbian and Croatian in order to compare both scenarios. The SETIMES dataset is used which is a collection of parallel news articles in the Balkan languages extracted from the http://www.setimes.com website [12] which consists of 9 languages overall. The dataset we employed was Croatian and Serbian parallel corpora. This corpora consisted of 170,466 aligned sentences which made up the 29,391 news articles which we will treat as documents in this study.

As with any exercise in language processing there were several steps that had to be performed in order to transform the data into a usable format. The data itself was in several large XML files which had to be processed to strip out the text[1]. Next the text was tokenized, transformed into lower case and Cyrillic characters (these only existed in the Serbian language documents) were transformed to their Latin equivalent by mapping between Cyrillic and Latin codes of Unicode character set. A small Serbian and Croatian stop word list was utilised and a Snowball stemmer[2] applied to all the text; the last step entails

[1] As a side effect, the XML turned out to be badly formed in places and needed to be fixed by hand.

[2] http://snowballstem.org.

the removal of diacritics from the text[3]. It should be noted that this process was applied to both Croatian and Serbian text in an identical fashion as the rules for these are virtually the same including usage of one common stemmer for Serbian and Croatian languages due to the similar morphology of those two languages [3].

The paragraphs were further processed to ensure that they contained a minimum number of words. The average number of words per document in the Hansard Collection used by [6] was 84 words in English and 86 words in French so we opted for a minimum word count of 60. This reduced the number of suitable documents from 29,391 to 19,842. It should be also be noted that this dataset contained a number of blank sentences/documents as well as documents that were entirely in English which were removed via testing documents to see if any passed a threshold of having more than 25% English stopwords[4]. The term by document matrix, before SVD is calculated, is subjected to Term Frequency - Inverse Document Frequency or Log-Entropy weighting (as described in [1,10]). In addition to this any term in the matrix that did not have a document frequency of at least 2 was discarded.

3.2 Method

Two sets of experiments were run with the dataset. The first experiment replicates the mate-retrieval test done in [6] and the second experiment tests our hypothesis. These are:

1. Perform the mate-retrieval test using the classic ML-LSA approach.
2. Perform the mate-retrieval test where each document in the semantic space consists only of a single language.

Each test was run 10 times. We varied the number of documents used in the semantic space 3 times using values of 1,000, 2,500 and 5,000 for a total of 6 sets of experiments. For each experiment, from the set of documents not used in the semantic space, 5,000 were chosen at random to perform the mate-retrieval test as in [6]. This set of experiments was run twice: once for Term Frequency-Inverse Document Frequency weighting and once for the application of the Log-Entropy weighting scheme as outlined earlier.

4 Results

The results from the experiment can be found in Table 1. These results are all of the averages from 10 runs as outlined above in the experiment description

[3] Diacritics are added to the top or bottom of a letter to indicate appropriate stress, special pronunciation, or unusual sounds not common in the Roman alphabet. In Serbian and Croatian, these markings indicate special pronunciation, like the difference between the pronunciation of C compared to Ć.

[4] The stop word list is available at http://www.lextek.com/manuals/onix/stopwords1.html. Note that single character stop words were not included as it was found that many Serbian/Croatian documents were flagged as English when they were present in the list.

in Sect. 3.2. The first column represents the dataset size in the experiment run. Note that in the case where the parallel documents are *not* combined into one the actual size of the semantic space will be double the dataset size as each document is put into the space individually. So a dataset size of 2,500 that does not have the parallel documents combined will actually contain 5,000 documents. The next two-columns indicate whether or not the term by document matrix was weighted using Term Frequency-Inverse Document Frequency (tf-idf) or Log-Entropy (l-e)[5]. The following column indicates whether or not it's a combined dataset (Serbian/Croatian documents combined into one). The next two columns indicate the result from the mate retrieval exercise. The 'Srb → Cro' column indicates the percentage of the time that the Serbian document used as a search document for its mate found it as the closest match, the column after is the opposite case. The next two columns indicate the percentage of the time during the mate retrieval exercise the correct document appeared in the top 5 returned results. The last two columns contain the average number of words in the semantic space as well as the average document length.

Table 1. Results for experiment runs

Dataset Size	tf-idf	l-e	Comb	Srb → Cro	Cro → Srb	top5Srb	top5Cro	Words	Doc Length
1000	✓		✓	98.1%	98.1%	98.9%	98.9%	8248.8	266.2
1000		✓	✓	99.2%	99.1%	99.7%	99.6%	8248.8	266.2
1000	✓			82.3%	82.3%	89.6%	89.2%	11237.4	135.2
1000		✓		90.1%	90.2%	94.5%	94.5%	11237.4	135.2
2500	✓		✓	97.7%	97.8%	98.2%	98.2%	13427.2	271.0
2500		✓	✓	98.7%	98.8%	99.1%	99.1%	13427.2	271.0
2500	✓			78.9%	79.0%	86.3%	86.3%	17506.4	136.3
2500		✓		86.4%	86.7%	91.4%	91.6%	17506.4	136.3
5000	✓		✓	97.5%	97.6%	97.8%	97.9%	18802.3	276.6
5000		✓	✓	98.3%	98.3%	98.7%	98.7%	18802.3	276.6
5000	✓			75.1%	75.3%	83.3%	83.2%	24066.7	137.1
5000		✓		83.4%	83.7%	88.6%	89.5%	24066.7	137.1

In addition to this data we calculated the average same similarity[6] scores of the twinned documents used in the mate retrieval exercise as well as the average of the closest match in the semantic space excluding the target document. These values can be found below in Table 2.

[5] We discovered, serendipitously, that the results of using tf-idf and l-e were actually superior when the folded-in search queries were only weighted using raw term-frequency. This was unexpected and will be a topic of future research. The results reported here use the commonly accepted approach of weighting the query appropriately with the weighting method used for the creation of the semantic space.

[6] The same similarity score is the cosine similarity between the two 'mate' documents.

Table 2. Average similarity values

Dataset Size	tf-idf	l-e	Combined	Avg SS	Average Best Srb Sim	Average Best Cro Sim
1000	✓		✓	0.951	0.614	0.613
1000		✓	✓	0.947	0.561	0.561
1000	✓			0.777	0.640	0.640
1000		✓		0.800	0.592	0.595
2500	✓		✓	0.956	0.686	0.686
2500		✓	✓	0.954	0.634	0.635
2500	✓			0.797	0.703	0.705
2500		✓		0.813	0.659	0.662
5000	✓		✓	0.959	0.720	0.720
5000		✓	✓	0.957	0.661	0.662
5000	✓			0.799	0.735	0.737
5000		✓		0.809	0.683	0.686

5 Discussion

First, it is interesting to note that the performance of the classic ML-LSA approach slightly out-performs the English/French mate retrieval experiment described in [6] so this acts as another confirmation of the ML-LSA approach (with l-e). A more detailed comparison on this test is difficult to realise as many details of how LSA was employed are left out. For example the number of dimensions used in the dimensional reduction is not specified nor was it mentioned if any weighting was applied to the generated term by document matrix which makes an exact replication of the scenario impossible.

Overall, it should also be noted, that Log-Entropy gives superior results compared with Term Frequency-Inverse Document Frequency every time. It should also be noted that the quality of results from the non-combined results outperformed most of the results reported in the study, [2], referred to in Sect. 2.3 although, to be fair, they were using much more challenging languages for their experiments; however this lends some support to our view that very closely related languages can be treated more like a 'regular' LSA implementation.

Next it should be observed that the performance of the non combined document test runs were somewhat worse than was expected. The accuracy, however, would still be reasonable for an information retrieval system so they can be viewed as positive overall. It is also surprising that as the dataset size increased the results started to worsen. The average similarity values found in Table 2 may offer some insight as to why this may be the case. In all the test scenarios explored there is a trend where as the data size increases so does the average best similarity for a search in the semantic space (including the same similarity average between the document mates with one close exception). These increases in average value are not inconsiderable - especially the average values representing the best similarity value in the semantic space with the Serbian or Croatian query document.

It is also worthy to note that the average number of words present in the semantic space for each increases as the number of documents used to form it increases. One of the co-authors of this paper is fluent in both Serbian and Croatian and spent some time going through a sample of several of the cases where the mate-retrieval test returned a different document than the targeted one. It was found that the news stories in question were actually related and it would have been viewed as a valid candidate for the search performed. For instance, there were some political issues in the Balkan region over a short timeframe and, as a result, there were several news articles addressing this topic in the corpus that contained similar information. It was not feasible to validate this for all the missed searches but it did seem to be a very strong theme throughout. The fact the average similarity values are increasing as the semantic spaces increases would seem to support this view; especially as the gap between the best same similarity and best similarity between the query document and the semantic space is narrowing considerably. As the variety and richness of words in the search space increases further underlying relationships of documents may be linked to potentially boost the similarity values of documents compared. Without a more exhaustive search of the scenarios where the mated document is not returned as the top choice, however, this cannot be said with certainty.

There are pros and cons to using an approach where a parallel corpus is not needed. One obvious advantage is all that may be required is a set of documents in the languages you wish to support; this alone may be sufficient to act as a useful training corpus for creating a semantic space. The downside to this is, of course, increased complexity as more documents may be required than what a parallel corpus could offer which would, therefore, increase the number of documents in our semantic space that we need to process. Performing the SVD computation is not cheap but on modern computers, in work unrelated to this, it was found that performing SVD on the Wikipedia corpus took a little under 2 weeks of computation time on a modestly powered server (the largest constraint was having enough RAM for the computation). The complexity barrier is not as inpeneratrable as it was 10 years ago so this could still be considered a viable option.

6 Conclusion and Future Work

The overall results from this exploratory early study shows some promising results. It does appear that, using the mate retrieval test, a mixed language semantic space can distinguish between the two separate languages being used. The fact that the languages are very closely related almost certainly plays a role in this result. As has been shown in other studies, even using ML-LSA, results between very different languages (in the morphological sense, say English and Arabic) are much more difficult to use in a multi-lingual search environment so this can be viewed as a special case.

This research leaves plenty of scope for future work. There are additional Balkan languages that could be added to the current Croatian and Serbian dataset to further test the applicability of this method with respect to similar languages. Besides Serbian and Croatian, there are the Bosnian and Montenegrin languages which appeared after the former Yugoslavia was broken-up. In addition, between Bosnian-Croatian-Montenegrin-Serbian (BCMS) there are the numerous shared features and easy mutual comprehension [3]. Almost 15 million people from the former Yugoslav region can have benefit of Multi-lingual IR tools and systems for BCMS languages.

The research paper discussed in Sect. 2.3 regarding the Hansard collection also utilised additional evaluation methodologies other than the mate retrieval test [6]. One such test attempted to simulate more realistic queries by finding the 5 nearest terms to each English test document and using these as a query. Presumably these would be expected to find the natural mate for the document the terms originated from but this could potentially involve a more customised search space to verify the results (they repeated a similar experiment using yellow pages categories and enhanced their semantic space by folding in a number of these too in order to see if the other language's version of these categories could be searched).

Additional experiments can be run using the same method discussed in this paper but, instead, with a semantic space generated randomly from the documents available to it rather than with matched pairs. It would also be useful to perform the same experiments again but with the creation of a semantic space that just contains roughly equal amounts of documents from each language that are not from a parallel corpus at all; this opens the possibility of trying this approach with more than 2 languages.

The corpus we are utilising is split down into the word/sentence level. A repeat of this experiment at the sentence level may also yield interesting results.

Acknowledgements. This article is based upon work from COST Action KEYSTONE IC1302, supported by COST (European Cooperation in Science and Technology).

References

1. Berry, M.W., Browne, M.: Understanding Search Engines: Mathematical Modeling and Text Retrieval, 2nd edn. SIAM, Philadelphia (2005)
2. Chew, P., Abdelali, A.: The effects of language relatedness on multilingual information retrieval: a case study with Indo-European and semitic languages. In: Proceedings of the 2nd International Workshop on "Cross Lingual Information Access" Addressing the Information Need of Multilingual Societies, pp. 1–9, January 2008. http://anthology.aclweb.org/I/I08/I08-6.pdf#page=10
3. Corbett, G.G., Browne, W.: Serbo-croat: Bosnian, Croatian, Montenegrin, Serbian. In: The World's Major Languages, pp. 330–346. Routledge, London (2009)
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
5. Dhavachelvan, P., Pothula, S.: A review on the cross and multilingual information retrieval. Int. J. Web Semantic Technol. **2**(4), 115–124 (2011)

6. Dumais, S.T., Letsche, T.A., Littman, M.L., Landauer, T.K.: Automatic cross-language retrieval using latent semantic indexing. AAAI Technical Report SS-97-05, pp. 18–24 (1997)
7. Dwivedi, S., Chandra, G.: A survey on cross language information retrieval. Int. J. Cybern. Inform. **5**(1), 127–142 (2016)
8. Greenberg, R.D.: Language politics in the federal republic of Yugoslavia: the crisis over the future of serbian. Slavic Rev. **59**(3), 625–640 (2008)
9. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Process. **25**, 259–284 (1998)
10. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
11. Sharma, M., Morwal, S.: A survey on cross language information retrieval. Int. J. Adv. Res. Comput. Commun. Eng. **4**(2), 384–387 (2015)
12. Tyers, F.M., Alperen, M.S.: South-East European Times: a parallel corpus of Balkan languages. In: Proceedings of the Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages, LREC 2010, pp. 49–53 (2010). http://xixona.dlsi.ua.es/~fran/publications/lrec2010.pdf
13. Young, P.G.: Cross-language information retrieval using latent semantic indexing. Master's thesis. University of Knoxville, Tennessee (1994)