STSM VISIT REPORT

STSM COST-1302
Reference: COST-STSM-ECOST-STSM-IC1302-030417-084423

Visiting Researcher: Dr. Colin Layfield, University of Malta
Host Researcher: Prof. Dragan Ivanović, University of Novi Sad

Title: Latent Semantic Analysis (LSA) applied towards a corpus of academic documents in more than one language.

Purpose

The purpose of this STSM was to determine how feasible it would be to apply LSA to a set of similar languages – this could be categorized as a subfield of Information Retrieval (IR), the so-called Cross-Language Information Retrieval problem (CLIR).

The underlying hypothesis to this work is to try and leverage the inherent similarities between various languages such as Serbian, Croatian, Montenegrin and Bosnian; these similarities may open a unique opportunity to take a more aggressive approach, with techniques such as LSA [1], to enable CLIR with a training set comprised of documents from these languages in an almost interchangeable fashion.

Description of Work

Many activities were performed during the week of April 3 – April 7, 2017 at the University of Novi Sad.

A brief literature review was performed by Dr. Layfield before his arrival in Novi Sad so an approach could be formulated immediately [2] [1] [3] [4] [5]. Utilizing LSA for CLIR is not a new idea and it was important to understand the approaches that have been tried previously.

The most relevant piece of work was from [4] where experiments were performed using LSA to achieve CLIR between French and English utilizing a parallel corpus of each language (in this case Canadian Parliamentary proceedings which must be available in both official languages). This dataset, known as the Hansard collection, is parallel in the sense that each paragraph recorded from the parliamentary proceedings is available in both French and English, thus there are two versions of the same text in two languages.

The experiments performed in [4] entailed of the creation of "documents" to be folded into an LSA "semantic space" with the requirement that each document contains the text for the same passage in both Canadian official languages – thus each document contains the same information twice, once in English and once in French. The experiment then generated a LSA semantic space from this corpus of documents. A set of documents was withheld from

being placed into this sematic space and they were used as a test set. A “mate-retrieval” test was employed where 2 documents (each of a single language, one English and one French) that are not contained in the semantic space are “folded in” and, for example, the best match for the English document is found (which should be its French equivalent). Their mate finding rate reported was in the range of 98%. This test would be the objective of the research with the added constraint that parallel documents in other language(s) would not be merged into one and form the semantic space as such (for example, a Serbian/Croatian document would not be merged to formulate one document, they would be added as individual documents on their own).

The host provided a parallel corpus of documents in various languages. The languages that were to be focused on, at this point, were Serbian and Croatian although there are additional libraries to explore for future research.

Utilities and tools had to be developed to parse the parallel document sets and process them via a stemming/stop word removal tool that was supplied by the host. Various issues arose at this point as some of the XML was not well formed and this had to be fixed manually as the issues were discovered (given that the Croatian/Serbian parallel document sets consisted of over 170,000 parallel sentences this turned out to be somewhat time consuming). It was also discovered, later in the process, that there were numerous pieces of text in English which needed to be removed/ignored so additional development time was spent implementing a stop word approach to testing as to whether a particular piece of text was likely to be English such that they could be skipped and not included in the results (they were skewing the initial results noticeably).

Once the data was properly processed additional infrastructure had to be produced that would be able to deal with the processed text and, at the user’s discretion, create a semantic space based on a subset of the parallel corpus, read in the documents *not* in the semantic space and perform the mate matching experiment with a portion of that data.

Results Obtained

Although there is much more experimentation planned the early results are very promising. Using the “mate-retrieval” model it was found that the mate is correctly identified as the most similar document more than 95% of the time which is comparable to the earlier research by Dumais, et al. [4] with the added distinction that dual language documents were *not* used in the creation of the semantic space, rather different languages were mixed in together in the semantic space via their own individual document representations.

This opens the door for future research using other language pairs (or larger sets together) in similar experiments. The implication is that a parallel corpus, for certain types of languages that are very similar, may not be required in for CLIR-LSA to be utilized effectively.

Future Collaboration Opportunities

The visit, by both parties, was deemed a success. The long term potential application for this collaborative effort is to apply it to a large corpus of academic documents in several

languages such that they can be searched natively in the language of the searcher's choice (provided it is one of the languages in the corpus). The early results from this research visit are a positive step towards this goal.

Dr. Layfield also spent time with another faculty member at the University of Novi Sad, Professor Aleksandar Kovacevic, with whom he shares various research interests with; specifically LSA and Latent Dirichlet Allocation (LDA). It was proposed that there was a potential for future collaboration, perhaps doing comparative studies between the two technologies applied to different IR themed problems.

Future Publications/Articles

It is likely that a paper will be submitted with some of the preliminary results to the 3rd International KEYSTONE conference to be held in Gdansk Poland as part of the COST action which sponsored this STSM.

Dr. Layfield believes there is enough research and experimental potential in this work that it can be expanded upon further in the upcoming year and turned into a longer submission to a relevant high quality Journal in the field of IR.

Dr. Colin Layfield
18/4/2017 - Malta

Bibliography

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science,* vol. 41, no. 6, pp. 391-407, 1990.

[2] P. A. P. Chew, B. W. B. Bader and A. Abdelali, "Latent morpho-semantic analysis: multilingual information retrieval with character n-grams and mutual information," in *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, United Kingdom, 2008.

[3] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester and R. Harshman, "Using latent semantic analysis to improve access to textual informaiton," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1988.

[4] S. T. Dumais, T. Letsche, M. L. Littman and T. K. Landauer, *Automatic Cross-Language Retrieval Using Latent Semantic Indexing,* 1997, pp. 18-24.

[5] P. G. Young, *Cross-Language Information Retrieval Using Latent Semantic Indexing,* Tennessee: MSc. Thesis, Univesity of Knoxville, 1994.