International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 10 (2018) pp. 314-322
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Received: 21 Jun, 2018, Accepted: 27 Oct, 2018, Publish: 7 Nov, 2018

Cross-language Information Retrieval by Reduced *k*-means

Jasminka Dobša[1], Dunja Mladenić[2], Jan Rupnik[2] , Danijel Radošević[1] and Ivan Magdalenić[1]

[1] Faculty of Organization and Informatics, University of Zagreb,
Pavlinska 2, 42000 Varaždin, Croatia
jasminka.dobsa@foi.hr, danijel.radosevic@foi.hr, ivan.magdalenic@foi.hr

[2] Jožef Stefan Institute,
Jamova 39, 1000 Ljubljana, Slovenia
dunja.mladenic@ijs.si, jan.rupnik@ijs.si

***Abstract*: Cross-language information retrieval aims at retrieving relevant documents in one language for a query set in another language. Here we propose a new approach to the problem of cross-language information retrieval based on factorization of a term-document matrix by an iterative method of Reduced *k*-means clustering. Method of Reduced *k*-means intended at simultaneous reduction of objects (documents) and variables (index terms). Proposed method is compared to standard machine learning techniques of cross-language information retrieval by usage of latent semantic indexing and canonical correlation analysis. Motivation for usage of Reduced *k*-means method for a task of cross-language information retrieval comes from an observation that documents in a semantic space obtained by method of latent semantic indexing are clustered by their language and not by their topics in the first place. As Reduced *k*-means aims at preserving clustering structure of data, the idea is that the proposed method could address the mentioned problem.**

***Keywords*:** cross-language information retrieval, dimensionality reduction, latent semantic indexing, canonical correlation analysis, Reduced *k*-means

I. Introduction

Cross-language information retrieval is a text mining task of retrieving relevant documents in one language initiated by a query set in another language. The aim of the paper is to present a novel algorithm for cross-language information retrieval by usage of the dimensionality reduction method of Reduced *k*-means clustering [5]. The task of cross-language information retrieval is chosen for testing because the use of dimensionality reduction methods is essential for this task contrary to general task of information retrieval. The proposed algorithm is compared to the most commonly used algorithms for cross-language information retrieval based on factorization of a term-document matrix proposed by Dumas et al. [7] using latent semantic indexing (LSI) and by Vinokourov et al. [26] using canonical correlation analysis (CCA). Our motivation for the application of the Reduced *k*-means algorithm comes from the criticism of the method based on LSI regarding the fact that documents in lower-dimensional or semantic space are primarily clustered in the lower-dimensional space by language and not by their topic [3]. As the Reduced *k*-means method aims at preserving the clustering structure of data present in the semantic space, we presupposed it would be an adequate means of addressing the aforementioned problem. The iterative method of Reduced *k*-means (RKM) simultaneously reduces objects (by clustering) and variables (by extraction of factors). In this paper, we test the proposed method for cross-language information retrieval on the data from the Aligned Hansards data of the 36th Parliament of Canada [9], a parallel corpus of English and French documents by conducting mate retrieval, that is, retrieval of translated document in one of the languages as the most similar for a given document in another language as a query. Also, an academic example on a small data set in English and French shows how documents are projected by used algorithms in the semantic space.

The rest of the paper is organized as follows: The second section describes related work on cross-language information retrieval and classification as well as methods for reduction of dimension by simultaneous reduction of objects and variables. The third section describes the methods for cross-language information retrieval by LSI (CL-LSI) and by CCA (CL-CCA), while in the fourth section we introduce the proposed method of cross-language information retrieval by RKM (CL-RKM). In the fifth section, an example of the application of the method on a small data set shows visualization of documents in semantic space for observed methods, the sixth section describes our experiment and the final section concludes the paper.

II. Related work

A. Cross-language information retrieval and classification

There are a few approaches to cross-language information retrieval and classification [20]. They are based on several methods including translation and dictionaries, probabilistic topic models, matrix factorization, monolingual models and neural network word embeddings. The most basic way to find similar documents written in different languages is to perform information retrieval on previously machine-translated documents [15], [16]. In [16] and [17] the similarity of documents in different languages is inspected by using dictionaries, which in both cases are EuroVoc dictionaries.

The generality of such approaches is limited by the quality of available linguistic resources, which may be scarce or non-existent for certain language pairs. Another intuitive approach is to ignore the fact that documents come from different languages and to perform simple monolingual information retrieval. Such an approach is applied by representation of documents as bags of character n-grams in [16]. The intuition behind this approach is that the named entities and cognate words are written in the same or similar fashion in many languages. Monolingual information retrieval is also used as a baseline for comparison of methods of cross-language information retrieval using matrix factorizations, which are tested in this paper.

In probabilistic topic models a representation of documents is modeled in a language independent way by using probabilistic graphical models. The models include: joint probabilistic latent semantic analysis (JPLSA) [14], coupled probabilistic LSA (CPLSA) [14], and polylingual topic models (PLTM) [13]. Several matrix factorization-based approaches exist in the literature. These are cross-language latent semantic indexing CL-LSI [7], cross-language canonical correlation analysis (CL-CCA) [26], and oriented principal component analysis (OPCA) [14]. CL-LSI and CL-CCA methods have gained greater popularity in comparison to OPCA due to the computation complexity of OPCA and the fact that it forces the vocabulary sizes for all languages to be the same, which makes this approach less intuitive. In [18] the authors investigated how to optimally select an LSI training dataset for specific classification tasks and thus built domain specific common document representations. Many approaches in the recent natural language processing literature involve neural network models which construct distributed index term representations [11], [27]. Such approaches typically involve many tuning parameters and the training is usually computationally intensive.

A problem of cross-language classification is similar task as cross-language information retrieval and it is interesting in the context of approach we are investigating because of ability of Reduced *k*-means method to preserve clustering (or classification) structure of the collection of documents. Bel and coworkers [1] were among the first to consider that problem. The major limitation in cross-language information retrieval and classification methods involving matrix transformations is their dependence on a parallel corpus which is sometimes hard to obtain. A possible solution to that issue is so-called domain adaptation. Domain adaptation is a problem of adaptation of statistical classifier trained on one domain for application on another. In [10] a method is proposed for using multilingual domain models from comparable corpora in an unsupervised way by introduction of a generalized similarity kernel function inside support vector classification. A method using nonlinear bilingual translation model trained on comparable corpora that provides cross-lingual estimates that correlate well with human judgments is described in [22]. Blitzer and coworkers [2] propose structural correspondence learning, domain adaptation method relaying on singular value decomposition, matrix transformation used in latent semantic indexing approach. In [19] a method built on structural correspondence learning which induces correspondences among words in languages by means of small number of suggestive words with high discriminative power called pivots is proposed and tested for cross-language sentiment analysis.

Another solution for overcoming the problem of a parallel corpus needed for training in cross-language tasks is using machine translation tools for that purpose. Performance of this approach is similar as on the collections where human generated parallel corpus is available, as reported in [8]. Furthermore, it is shown that performance is better compared to training of a classifier on a parallel corpus in a different domain. In our research, the machine translation approach is applied on an example provided in the fifth section of this paper.

B. Related methods of dimensionality reduction

A standard procedure for clustering of objects in a lower-dimensional space is the so-called tandem analysis [25]. It includes the projection of data on a lower-dimensional space by principal components analysis and clustering of data in that space. Hence, the extraction of new variables and clustering is applied sequentially. De Sarbo et al. in [5] and DeSoete and Carroll [6] warn against that approach since principal components may extract dimensions which do not necessarily significantly contribute to identify clustering structure of the data. They propose a method that simultaneously clusters data and extracts the factors of variables by reconstructing the original data with only centroids of clusters in a lower-dimensional space. They also propose an altering least squares algorithm for the proposed method that gives at least local minimum. Vichi and Kiers [25] propose Factorial *k*-means, an algorithm with the same aim of simultaneous reduction of objects and variables by reconstruction of data in a lower-dimensional space by its centroids in the same space. However, the application of this method is limited to cases in which the number of variables is less then number of cases. This usually is not satisfied in the case of text mining applications since the number of documents (objects) in collection usually is lower than the number of index terms (variables). In [23] the Reduced *k*-means and Factorial *k*-means methods are compared based on simulations and theoretically in order to identify cases for application of these methods. In addition, Timmerman et al. propose the method of Subspace *k*-means which gives an insight in cluster characteristics in terms of relative positions of clusters given by centroids and shape of the cluster given by with-in cluster residuals [24]. In [21] the principal cluster axes method is proposed which aims to find directions of maximal clusterability in multivariate data.

III. Methods for cross-language information retrieval using matrix transformations

A. Cross-language information retrieval by latent semantic indexing (CL-LSI)

Latent semantic indexing was introduced as a semantic information retrieval technique in 1990 [4]. The purpose of the technique is to capture semantic relations between index terms used for representation of documents in a bag of words representation and to overcome problems of synonyms and polysemy in information retrieval. It is based on the singular value decomposition (SVD) of a term-document matrix $\mathbf{X} = \mathbf{U\Sigma V}^T$ where **U** and **V** are orthonormal matrices and **Σ** is a diagonal matrix on whose diagonal are singular values of matrix **X** in descending order. Let the term-document matrix **X** be of type *m* x *n* where *m* is the number of index terms and *n* is the number of documents. Latent semantic representation of documents in the semantic *k*-dimensional space for $k < m$ is obtained by truncated SVD of the term-document matrix $\mathbf{X}_k = \mathbf{U_k \Sigma_k V}_k^T$ where $\mathbf{U}_k$ is the matrix formed by the first *k* columns of matrix **U**, $\mathbf{\Sigma_k}$ is the diagonal matrix on whose diagonal are the *k* greatest singular values of matrix **X** and $\mathbf{V_k}$ the matrix formed by the first *k* columns of matrix **V**. The representation of documents in the semantic space is given by $\mathbf{\Sigma_k V}_k^T$. The representation of a new document (query **q**) in semantic space is given by projection of the query onto the first *k* columns of matrix $\mathbf{U_k}$ given by $\mathbf{U}_k^T\mathbf{q}$.

While performing cross-language information retrieval we assume that we have two different representations (or two different views) of the collection of documents corresponding to two different languages, or two term-document matrices, related to matched (i.e., translated) documents. Cross-language representation of documents by CL-LSI is given by the SVD applied on the concatenation of bag of words representations for two different languages. Let $\mathbf{R_1}$ and $\mathbf{R_2}$ be representations of matched documents for two different languages. Then the representation of documents in a language independent space is given by the truncated singular value decomposition obtained from the concatenated matrix of two different representations

$$\begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix} = \mathbf{U\Sigma V}^T. \qquad (1)$$

B. Cross-language information retrieval by canonical correlation analysis (CL-CCA)

Canonical correlation analysis is a technique proposed in [12] which aims to find basic vectors for two sets of variables such that the correlation of projections of variables on basic vectors is mutually maximized. The goal of application of CCA on representations of collections of documents $\mathbf{R_1}$ and $\mathbf{R_2}$ is to find two linear mappings from $\mathbf{R_1}$ and $\mathbf{R_2}$ onto a common language-independent space such that the correlation between these mappings is maximized. The problem is reduced to a generalized eigenvalue problem whose eigenvectors give the desired mappings. CCA can be kernelized in a way that CCA is not applied directly onto the representation of documents, but onto their scalar products, or the so-called kernel functions.

IV. Method based on Reduced *k*-means clustering

The method of Reduced *k*-means was introduced by DeSoete and Carroll [6]. The goal of the method is simultaneous projection of objects on a lower-dimensional subspace, and clustering of the objects within that subspace. Let **X** again be an *m* x *n* term-document matrix. Further, we define the following notation:

- **A** is an *m* x *k* columnwise orthonormal loading matrix of extracted factors
- **M** is an *n* x *c* membership matrix, where *c* is a predefined number of clusters; m_{ic}=1 if document *i* belongs to cluster *c* and 0 otherwise
- **Y** is a *c* x *k* matrix which gives centroids of clusters in the lower-dimensional space.

By definition, we suppose that every document in the collection belongs to exactly one cluster. The Reduced *k*-means method minimizes the loss function $F(\mathbf{M},\mathbf{A}) = \left\| \mathbf{X} - \mathbf{AY^TM}^T \right\|$ in the least squares sense. The projections of objects in the lower-dimensional space are given by $\mathbf{A}^T\mathbf{X}$. The dimension of the lower-dimension space must be less or equal to the number of clusters.

De Soete and Caroll [6] also propose alternating least squares (ALS) algorithm which alternates between corrections of loading matrix **A** in one step and membership matrix **M** in another. As each of the steps in the ALS algorithm improves the loss function, the algorithm converges to at least the local minimum. By starting the procedure from a large number of random initial estimates and coosing the best solution, the chances of obtaining global minimum are increased.

If $\mathbf{R}_1$ and $\mathbf{R}_2$ are representations of documents in a semantic language independent space, Reduced *k*-means is performed by solving the following least squares problem

$$F_{CL\_RKM}(\mathbf{M},\mathbf{A}) = \left\| \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix} - \mathbf{AY^TM}^T \right\| \to \min \qquad (2)$$

where $\mathbf{R_1}$ and $\mathbf{R}_2$ are m_1 x *n* and m_2 x *n* term-document matrices respectively that represent a collection of document in two different languages; m_1 and m_2 are numbers of index terms in the first and the second language.

V. Example: visualization of documents

We created a simple data set consisting of 15 documents written in English and French. The documents are titles of the books offered for sale on the Amazon website and their French translations. The books were selected from three different fields: art, computer science and medicine. From each of the fields we chose 5 titles and, creating in such a way three well separated classes of documents. The aim of this example was to project representations of documents onto the plane to get the impression of how the documents would be projected by aforementioned methods. Particularly, we were interested if the representations of documents would be clustered according to the defined classes. The documents in English and in French are shown in Table 1.

The documents were preprocessed by the ejection of punctuation and accents in the French documents as well as conversion to lowercase and ejection of the stop words in both languages. Stemming or lemmatization were not applied. The dictionary for both languages was made up of terms that appeared in at least two documents resulting in dictionary consisting of 12 English and 9 French terms. The English terms were *art*, *drawing* and *music* (from the field of art), *computer*, *problems*, *software* and *virus* (related to the field of computer science) and *clinical*, *medical*, *symptoms* and *virus* (in the field of medicine). Besides, the term *guide* was present in documents from all the three classes. Also, the index term *virus* is homonym present in document in the class of computer science and that of medicine. Index terms in French were: *art*, *dessin*, *guide*, *informatique*, *logiciels*, *maladies*, *problems*, *symptoms* and *virus*. This set of index terms is a subset of English set of index terms. The differences between the dictionaries appeared because of greater inflection in the French language. Projections of documents onto the plane are shown in Figures 1a to 1e. Figure 1a shows projections in a two-dimensional space obtained by LSI method. Figures 1b and 1c show projections in a two-dimensional semantic space obtained by the Reduced *k*-means method, the first one with arbitrary initial clustering and the second with initial clustering close to the given classification. In the case of arbitrary initial clustering, the final clustering is semantic, but differs from the given classes. In the class of art, the documents related to drawing (1,2 and 5) are clustered together, while the documents related to music are slightly separated. Also, the documents containing the word *guide* are clustered together. From Figures 1b and 1c it is evident that for obtaining representation that supports the given classification it is recommendable that the initial clustering close to the given classes is chosen.

As can be seen, there has been improvement in the clustering of documents for projections obtained by the Reduced *k*-means method with initial clustering close to the given classes in comparison to the LSI method. Figures 1d and 1e show representations obtained by the CCA method. Figures 1d and 1e show representations of documents in a language-independent space written in English and French, respectively. For the CCA method it can be seen that projections for English and French documents are close, which results in efficient cross-language information retrieval. Nevertheless, there is no clustering of documents according to topics.

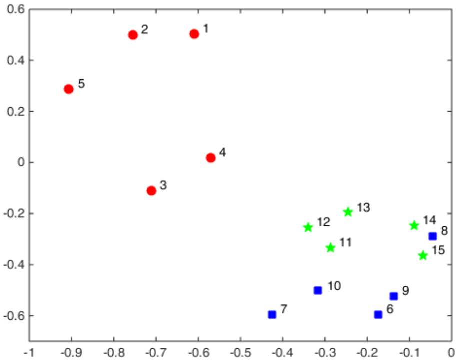

Figure 1a. Projections by LSI method

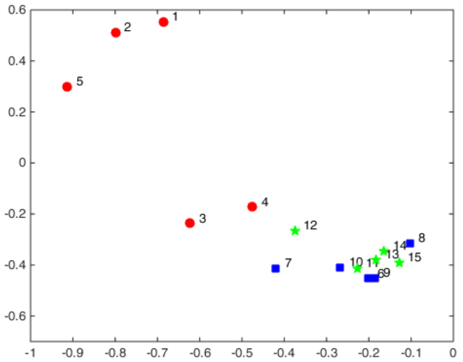

Figure 1b. Projections by Reduced *k*-means method (arbitrary initial clustering)

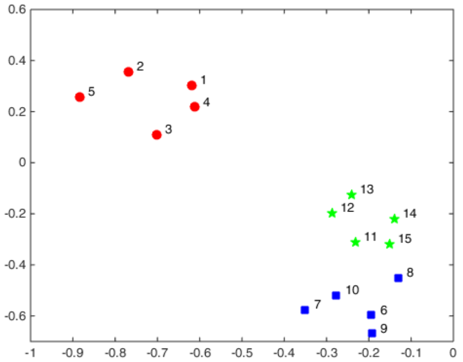

Figure 1c. Projections by Reduced *k*-means method (initial clustering close to given classification)

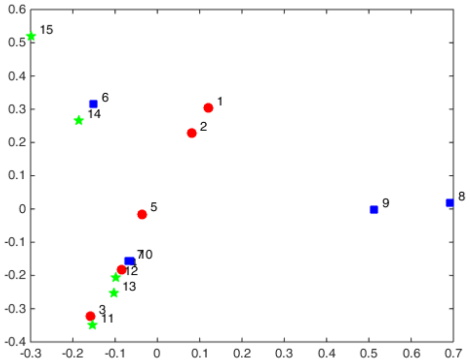

Figure 1d. Projections of documents by CCA method represented by English terms

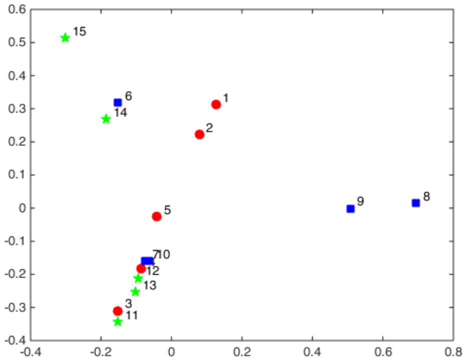

Figure 1e. Projections of documents by CCA method represented by French terms

V. Experiment

A. *Design of experiment*

Experimental analyses were conducted on the publicly available data Aligned Hansards of the 36th Parliament of Canada [9]. We used a House debates training set consisting of 13610 parallel training documents in each of the two used languages (English and French). The text consists of paragraphs delimited by “ *** “ that are treated as separate documents. As a testing data set used only “testing 1“ collection consisting of 963 test documents in each of the two languages. We divided the training data set in five parts in a systematic way and preprocessed it separately for each language by removing stop words, stemming (or reducing of words to their root form) and removal of words appearing

	Field	Language	Document
1	Art	English	Artist's Drawing Techniques
		French	Techniques de dessin d'artiste
2	Art	English	Art for Kids: Drawing: The Only Drawing Book You'll Ever Need to Be the Artist You've Always Wanted to Be
		French	Art pour les enfants: Dessin: le seul livre de dessin dont vous aurez toujours besoin pour être l'artiste que vous avez toujours voulu être
3	Art	English	The Art of Practicing: A Guide to Making Music from the Heart
		French	L'art de pratiquer: un guide pour faire de la musique du coeur
4	Art	English	The Art of Music Production: The Theory and Practice
		French	L'art de la production musicale: la théorie et la pratique
5	Art	English	The Art of Botanical Drawing: An Introductory Guide
		French	L'art du dessin botanique: un guide d'introduction
6	Computer science	English	What was the first computer virus?
		French	Quel fut le premier virus informatique ?
7	Computer science	English	The Beginner's Guide to Engineering: Computer Engineering
		French	Le guide du débutant en ingénierie: ingénierie informatique
8	Computer science	English	Formal approaches to software development
		French	Approches formelles pour le développement de logiciels
9	Computer science	English	Power and limitations of computer systems. Computer studies and software
		French	Puissance et limites des systèmes informatisés. Etudes et logiciels informatiques
10	Computer science	English	Computer Science Distilled: Learn the Art of Solving Computational Problems
		French	L'informatique distillée: apprendre l'art de résoudre les problèmes de calcul
11	Medicine	English	Medical Symptoms: A Visual Guide: The Easy Way to Identify Medical Problems
		French	Symptômes médicaux: un guide visuel: la façon facile d'identifier les problèmes médicaux
12	Medicine	English	Clinical Guide to Popular Diets
		French	Guide clinique des régimes populaires
13	Medicine	English	Medical Resonance Music 934: General Stress Symptoms
		French	Medical Resonance Music 934: Symptômes de stress général
14	Medicine	English	Essentials of Clinical Infectious Diseases
		French	Essentiels des maladies infectieuses cliniques
15	Medicine	English	Zika Virus and Diseases: From Molecular Biology to Epidemiology

	French	Le virus Zika et les maladies: de la biologie moléculaire à l'épidémiologie

Table1. Documents: titles of books from www.amazon.com and their French translation

	EN – FR			FR - EN		
k	CL-LSI	CL-CCA	CL-RKM	CL-LSI	CL-CCA	CL-RKM
10	0.5372 ±0.0268	0.5452 ±0.0094	0.4004 ±0.0094	0.5495 ±0.0157	0.5531 ±0.0165	0.4081 ±0.0123
20	0.7441 ± 0.0119	0.7500 ±0.0363	0.6592 ±0.0248	0.7731 ±0.0147	0.7687 ±0.0341	0.6816 ±0.0225
30	0.8033 ±0.0092	0.8359 ±0.0333	0.7657 ±0.0111	0.8272 ±0.0158	0.8507 ±0.0191	0.7890 ±0.0181
40	0.8417 ±0.0049	0.8887 ±0.0215	0.8160 ±0.0161	0.8675 ±0.0090	0.8959 ±0.0159	0.8322 ±0.0141
50	0.8688 ±0.0044	0.9146 ±0.0105	0.8557 ±0.0186	0.8882 ±0.0059	0.9211 ±0.0125	0.8675 ±0.0093
100	0.9319 ±0.0049	0.9572 ±0.0105	0.9292 ±0.047	0.9352 ±0.0075	0.9533 ±0.0059	0.9317 ±0.0031
150	0.9510 ±0.0052	0.9711 ±0.0029	0.9442 ±0.0028	0.9501 ±0.0043	0.9630 ±0.0036	0.9506 ±0.0035
200	0.9580 ±0.0054	0.9715 ±0.0017	0.9504 ±0.0052	0.9597 ±0.0029	0.9695 ±0.0021	0.9568 ±0.0034
300	0.9603 ±0.0019	0.9747 ±0.0023	0.9545 0.0056	0.9658 ±0.0052	0.9732 ±0.0038	0.9620 ±0.0019
500	**0.9618 ±0.0023**	**0.9753 ±0.0012**	**0.9560 0.0040**	**0.9701 ±0.0042**	**0.9769 ±0.0029**	**0.9652 ±0.0020**

Table 2. Proportion of mate retrieval (mean ±st. dev.) for English (columns 2 through 4) and French documents (columns 5 through 7)

in less than 5 documents. The division of the collection facilitated matrix transformations and gave us insight in the stability of the applied algorithms. In this way we obtained dictionaries consisting, on average, of 7093 English and 8540 French index terms for a part. After that, *TfIdf* weighting was used (frequency of index term in the document is used as a local weight of term in the document and function ln(*n/df*) as a global weight of term in collection of documents, where *n* is the total number of documents and *df* is the number of documents in which a specific term appears).

In addition, representations of documents were normalized to the unit norm. On each part of the training data set the documents were projected on a lower-dimensional space using SVD, CCA and RKM methods. The methods were always evaluated on the same test data set which was represented using index terms for English and French language and test documents were projected on a language-independent semantic space using the aforementioned methods. Evaluation was conducted for dimensions ranging from 10 to 500, wherein the number of clusters for the RKM method was 505 since it has to be greater or equal to the number of dimensions. We have used 10 random starts and chose representation with the least value of the loss function.

B. Results

As evaluation measures we used the following:

a) the proportion of documents in one language for which the first retrieved document among documents in the other language is its translation, or the so-called mate retrieval

b) the proportion of documents written in one language (English or French) for which the first retrieved document among all the other documents (English and French) is its translation

c) the proportion of English/French documents among the 100 first retrieved documents for an English/French document given as a query.

Table 2 shows the results of the first evaluation measure which is also used in [7] and [26] for the evaluation of CL-LSI and CL-CCA methods. The proportion of mate retrieval for documents represented in a high-dimensional space by bag of words representation was 0.7944 for English documents and 0.8199 for French documents. It can be seen that by using projections on a lower-dimensional space this is outperformed for the dimensions of 30 by the CL-LSI and CL-CCA method and 40 by the CL-RKM method. Standard deviations values indicate that the performance of all the used methods is stable.

Figure 2a shows the proportion of mate retrieval in the set of English and French documents for a given English document, while Figure 2b shows the proportion of mate retrieval in the set of English and French documents for a French document given as a query. This proportion is slightly lower than in the first evaluation measure. Moreover, on a higher dimension of projections the results for the CL-LSI and, in particular, CL-RKM methods deteriorate. Figure 3a shows the proportion of English documents in the set of first 100 retrieved documents for a given English document. Figure 3b shows the proportion of French documents in the set of first 100 retrieved documents for a French document given as a query. On Figure 3a we can see that the proportion is approximately 0.5 for the dimensions less than 100 and it remains at this level for the CL-CCA method, while for CL-LSI and CL-RKM methods it increases to the level of almost 0.7 for the dimension of 500. In case of French documents given as a query (Figure 3b) for the dimension of 100 and above the proportion of French documents increases for all the examined methods. Here we can observe some assymetry between languages, which could be due to differences in the dictionary (arising from different level of inflection in English and French). The difference is notable in the very size of the dictionary: namely, the dictionary for English documents consists of 7093 English index terms, while the dictionary for French documents consists of 8540 French index terms.

VI. Conclusion and discussion

In this paper we introduced a novel method of cross-language information retrieval based on the Reduced *k*-means algorithm which simultaneously reduces variables by extracting factors and objects (here, documents) by clustering.

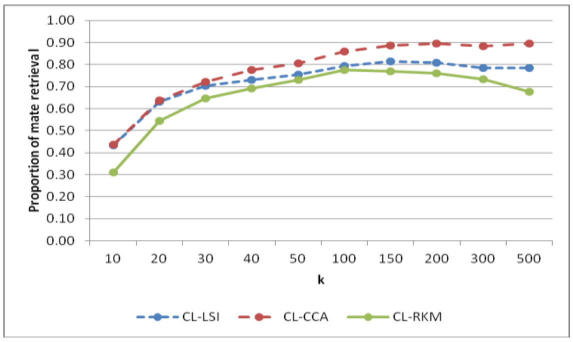

Figure 2a. Proportion of mate retrieval in the set of English and French documents for an English document query

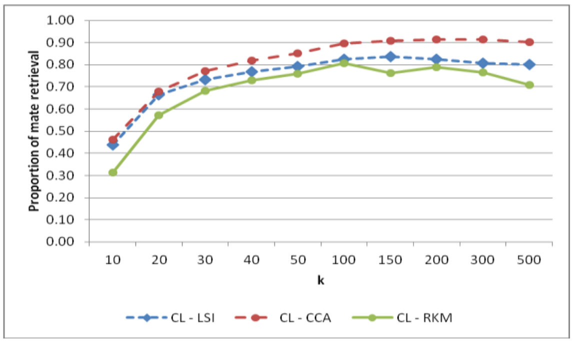

Figure 2b. Proportion of mate retrieval in the set of English and French documents for a French document query

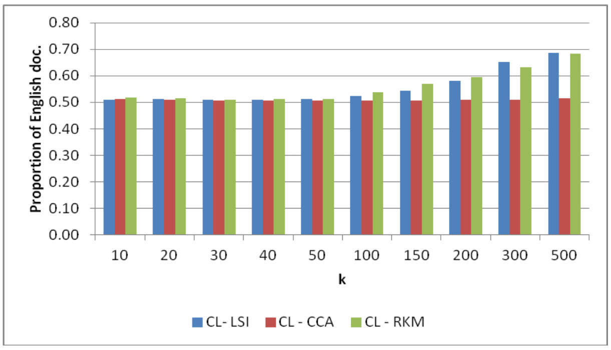

Figure 3a. Average proportion of English documents in the set of first 100 retrieved documents for English documents

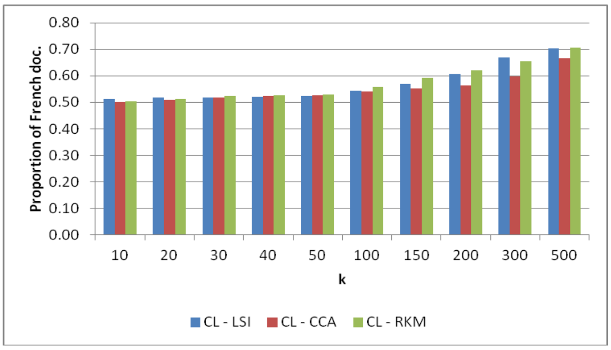

Figure 3b. Average proportion of French documents in the set of first 100 retrieved documents for French documents

Our motivation for this research came from the criticism of application of LSI for the task of cross-language information retrieval regarding the fact that documents are primarily clustered by their language and not by the topics.

According to the conducted experiments we can see that the performance of all the tested methods increases with the increasing number of dimensions of projection and for 500 dimensions achieves the high values ranging from 0.95 to 0.98. Regardless the number of dimensions, the best performance is achieved by CL-CCA. While on lower dimensions the CL-RKM method is outperformed by CL-LSI, their performance is comparable on higher dimensions of semantic space.

When we analyze the proportion of mate retrieval among all documents (English and French) for English documents as query (Fig 2a) or French documents as query (Fig 2b) we can see that these proportions are also high, reaching approximately 0.9 for the CL-CCA method and 0.8 for the CL-LSI and CL-RKM method. In case of the CL-RKM and CL-LSI method results deteriorate for higher dimensions of projection. The reason for such behavior may be the fact that for higher dimensions of projection documents are clustered according to language (Fig 3a and 3b). For dimensions lower than 100 the average proportion of the first 100 retrieved documents in the language of the query is approximately 0.5. However, for a higher number of dimensions that proportion rises both for CL-LSI and CL-RKM in the case of English documents and for all the three observed methods in the case of French documents given as a query.

The experimental results point to the conclusion that we did not succeed in addressing the problem regarding the CL-LSI method previously mentioned in the literature. An explanation for that may be that Reduced *k*-means groups documents in small clusters since the number of clusters is high (505) because of its limitation that the number of clusters must be equal or greater than the dimension of semantic space. The design of CL-LSI and CL-RKM methods forces learning of similarities among languages, which is achieved for smaller dimensions of projections. On the higher dimensions, the semantic within the languages is learned, which causes clustering of documents according to languages for higher dimensions of projections.

Based on a small example in fourth section we believe that the method of Reduced *k*-means could be useful for representation of textual documents because it has potential to cluster documents in previously defined classes. In the case of the experiment of information retrieval (or mate retrieval) conducted on a collection of Aligned Hansards of the 36th Parliament of Canada this advantage was not fully exploited since we do not have any information about clusters of documents in the collection. In the further work we plan to apply Reduced *k*-method in a supervised setting for the task of cross-language classification.

Acknowledgments

This work was supported by the Slovenian Research Agency and the Croatian Ministry of Science and Education. We thank Professor Henk A. L. Kiers for allowing us to utilize his implementations of used algorithms in our research and for very useful comments on the paper.

References

[1] N. Bel, C. H. A. Koster, & M. Villegas. „ Cross-lingual text categorization“ In *Proceedings of ECDL-03*, Tondheim, pp. 126-139, 2003.

[2] J. Blitzer, R. McDonald, & F. Pereira. „Domain adaptation with structural correspondence Learning“ In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Sidney, pp. 120-128, 2006.

[3] P. A. Chew, B. W. Baker, T. G. Kolda, & A. Abdelai. “Cross-language information retrieval using PARAFAC2”, In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Jose, California, USA, pp. 143-152, 2007.

[4] S. Deerwester, S. T. Dumas, G. W. Furnas, T. K. Landauer, & R. A. Harshman. “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, 41(6), pp. 381-407, 1990.

[5] W. S. De Sarbo, K. Jedidi, K. Cool, & D. Schendel. “Simultaneous multidimensional unfolding and cluster analysis: an investigation of strategic groups”, *Marketing Lett.*, 2, pp. 129-146, 1990.

[6] G. De Soete, ‚& J. D. Carroll. “K-means clustering in a low-dimensional Euclidean space”, in *New Approaches in Classification and Data Analysis*. E. Diday, et al. (eds) Springer, Heidelberg, pp. 212-219, 1994.

[7] S. Dumais, T. Letsche, M. Littman, & T. Landauer. “Automatic cross-language retrieval using latent semantic indexing”, in *Proceedings of the AAAI spring symposium on cross-language text and speech retrieval*, American Association for Artificial Intelligence, 16, pp. 15-21, 1997.

[8] B. Fortuna, & J. Shave-Taylor. “The use of machine translation tools for cross-lingual text mining”, In *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.

[9] U. Germann.” Aligned Hansards of the 36th Parliament of Canada”, Release 2001-1a, 2001. Available at http://www.isi.edu/natural-language/download/hansard/

[10] A. Gliozzo, & C. Strapparava. “Cross-language text categorization by acquiring multilingual domain models from comparable corpora, *in Proceedings of the ACL Workshop on Building and using Parallel Texts*, 2005.

[11] K. M. Hermann, & P. Blunsom. “Multilingual models for compositional distributional semantics”, in *Proceedings of ACL 2014*, Baltimore, Maryland, USA, 2014. Available at: http: //arxiv.org/abs/1404.4641.

[12] H. Hotelling, H. “Relations between two sets of variates”. *Biometrika*, 28, pp. 312-377, 1936.

[13] D. Mimno, H. M. Wallach , J. Naradowsky, D. A. Smith, & A, McCallum. “Polylingual topic models”. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP '09*, Association for Computational Linguistics, Singapore, pp. 880-889, 2009.

[14] J. C. Platt, & K. Toutanova, W.-t Yih. “Translingual document representations from discriminative projections”. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing EMNLP*, Association for Computational Linguistics, Massachusetts, USA, pp. 251-261, 2010.

[15] C. Peters, & M. Braschler. *Multilingual Information Retrieval*, Springer, Berlin, Heidelberg, 2012.

[16] M. Potthast, A. Barrón-Cedeño, B. Stein, B., & P. Rosso. “Cross-language plagiarism detection”, *Language Resources and Evaluation*, 45(1), pp. 45-62, 2011.

[17] B. Pouliquen, R. Steinberger, & O. Deguernel. “Story tracking: Linking similar news over time and across languages”, In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics*, Manchester, United Kingdom, pp. 49-56, 2008.

[18] J. Pozniak, & R. Bradford. “Optimization of cross-lingual LSI training data”, *Computer and Information Science*, pp. 57-73, 2015.

[19] P. Prrettenhofer, & B. Stein. “Cross-language text classification using structural correspondence learning”, In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL’10*. Uppsala, Sweden, pp. 1118-1127. 2010.

[20] J. Rupnik, “Multi-view canonical correlation analysis”, Doctoral Dissertation, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia. 2016.

[21] D. Steinley, M. J. Brusco, & R. Henson, “Principal Cluster Axes: A Projection Pursuit Index for the Preservation of Cluster Structures in the Presence of Data Reduction”, *Multivariate Behavioral Research*, 47(3), pp. 463-492, 2012.

[22] T. Šnajder, & D. Mladenić. “Cross-lingual document similarity estimation and dictionary generation with

comparable corpora, to appear in *Knowledge and Information Systems*, 2018.

[23] M. E. Timmerman, E. Ceulemans , H. A. L. Kiers, & M. Vichi. “Factorial and reduced K-means reconsidered”, *Computational Statistics and Data Analysis*, 54, pp. 1856-1871, 2010.

[24] M. E. Timmerman, E. Ceulemans, K. De Roover, & K. Van Leeuwen. “Subspace K-means clustering”. *Behavioural Resarch*, 45, pp. 1011-1023, 2013.

[25] M. Vichi, & H. A. L. Kiers. “Factorial k-means analysis for two way data”. *Computational Statistics & Data Analysis*, 37, pp. 49-64, 2001.

[26] A. Vinokourov, N. Cristianini, & J. S. Shawe-Taylor. “Inferring a semantic representation of text via cross-language correlation analysis”, In *Advances in Neural Information Processing Systems* 15, NIPS, Vancouver, British Columbia, Canada, pp. 1473-1480, 2002.

[27] I. Vulić, & M. F. Moens. “Bilingual Distributed Word Representations from Document-Aligned Comparable Data”, *Journal of Artificial Intelligence Research*, 55, pp. 953-994, 2016.

[28] M. Xiao, & Y. Guo. “A novel two-step method for cross language representation learning”, In *Advances in Neural Information Processing Systems* 26 NIPS, Sateline, NV, USA, pp. 1259-1267, 2013.

[29] D. Zhang, Q. Mei, & C. Zhai . “Cross-lingual latent topic extraction”, In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL*, Association for Computational Linguistics, Uppsala, Sweden, pp. 1128-1137, 2010.

Author Biographies

Jasminka Dobša was born in Čakovec, Croatia on 28th August 1971. She has obtained her BSc and MSc degrees at the Department of Mathematics, University of Zagreb, Croatia, in 1995 and 1999, respectively. She has obtained her PhD in the field of Computer Science at the Faculty of Electrical Engineering and Computing, University of Zagreb in 2006. She is currently working as an Associate Professor at Faculty of Organization and Informatics, University of Zagreb. Her major fields of study are data and text mining and applied statistics.

Dunja Mladenić was born in Pula, Croatia on 14th April 1967. She has obtained her BSc and MSc degrees at the University of Ljubljana, Slovenia, in the field of Computer Science in 1990 and 1995, respectively. She has obtained her PhD at University of Ljubljana with the thesis “Machine learning on non-homogeneous, distributed text data” in 1999. Her main research areas are artificial intelligence, machine learning, text and web mining, stream mining, semantic sensor networks, and social networks.

Jan Rupnik was born in Ljubljana, Slovenija on 6th December, 1982. Following graduation at the Faculty of Mathematics and Physics, University of Ljubljana in 2006, he was employed as a researcher at the Artificial Intelligence Laboratory at the Jožef Štefan Institute in Ljubljana. He defended his PhD thesis titled "Multiview Canonical Correlation Analysis" at the Jožef Stefan International Postgraduate School in June 2016. His main research area is machine learning, specifically statistical methods for cross-modal data analysis, sensor mining and anomaly detection.

Danijel Radošević was born in Zagreb, Croatia, on 27th March, 1969. He has obtained the PhD in Information and Communication Sciences at the University of Zagreb, and has been working as a Full Professor (since 2016) at the Faculty of Organization and Informatics in Varaždin, Croatia. His major fields of study are automatic programming, machine translation and teaching programming.

Ivan Magdalenić was born in Čakovec, Croatia on 17th April 1977. He received his BSc. and MSc degrees in Electrical Engineering from the University of Zagreb in 2000 and 2003, respectively, with a major in Telecommunications and Information Science. He has obtained a PhD in Computer Science from the University of Zagreb in 2009. His current research interests include automatic programming and semantic Web technology.