Improving the clustering or categorization of bi-lingual data by means of comparability mapping

Guiyao Ke, Pierre-François Marteau, Gilbas Ménier

Improving the clustering or categorization of bi-lingual data by means of comparability mapping

Guiyao Ke, Pierre-François Marteau, and Gildas Menier

Abstract—We address in this paper the co-clustering and co-classification of bilingual data by mixing similarity measures existing in each of the two linguistic spaces with a comparability measure that defines a mapping between these two spaces. A new approach is proposed to combine comparability and similarities measures with the aim to improve jointly the accuracy of classification and clustering algorithms performed in each of the two linguistic spaces, as well as the mapping of comparable clusters that are obtained. In this paper, we propose two variants of the comparability measure defined by [1] and evaluate our co-classification and co-clustering strategy on a data set collected from Wikipedia categories. Our experiments show clear improvements in clustering and classification accuracy when mixing comparability with similarities, with a higher robustness obtained when using the two comparability variants we propose. We believe that this approach is well suited for the construction of thematic comparable corpora of good quality.

Index Terms—Comparable corpora, Comparability measures, Classification, Clustering, Cluster mapping

1 Introduction

Parallel corpora are sets of tuples of aligned documents that are formed with texts placed alongside with their translation(s). If such resources are of great utility in particular in the field of assisted translation or multilingual information retrieval, they are expensive to develop and often difficult to transpose from a specialty domain to another. The notion of comparable corpora has emerged in the nineties to palliate this lack of versatility and expensiveness and to offer avenues to a wider scope of applications such as multilingual terminology extraction, multilingual information retrieval or knowledge engineering [2], [3]. However, the notion of comparability between documents expressed in different languages is not easy to introduce: it is widely admitted that two documents in different languages are comparable when they share analogous criteria of composition, genre and topics. The term of comparable corpora was introduced by [4], [5] and remains quite subjective. [6] proposed a quantitative definition of the concept of comparability according to which "Two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called comparable if there is a significant sub-part of the vocabulary of the $\mathcal{L}_1$ language corpus, respectively $\mathcal{L}_2$ language corpus, whose translation is in the corpus of language $\mathcal{L}_2$, respectively $\mathcal{L}_1$." [1] have then derived a quantitative measure that is based on a bilingual translation dictionary. This measure consists primarily in counting the presence of the translations of dictionary entries that occur in the paired documents. It depends on a non-explicit way upon jointly the coverage of the bilingual translation dictionary and the studied corpora themselves.

This comparability measure defined for bilingual corpora indeed applies when dealing with monolingual documents that partition in two distinct linguistic spaces, as far as a bilingual dictionary connecting the two spaces is available. At a document level we thus face a situation where monolingual similarities exist in each linguistic space that are potentially linked by a comparability measure. In the scope of the construction of thematic comparable corpora, this leads to address the co-classification or co-clustering of data since we are targeting the mapping of highly *comparable* clusters of documents that are furthermore thematically coherent in each linguistic space, i.e. characterized by a high *intra-similarity*. We confront such situation when harvesting multilingual data from the web for instance. With the need for comparable resources getting pressing, approaches that exploit consistently similarities and comparability are becoming particularly useful.

In this paper we introduce, study and evaluate the impact of three comparability measures (one referent measure and two variants), on what we call the co-categorization and co-clustering of bilingual data. To this end, we develop a new dedicated approach for combining comparability and similarities to provide the identification and mapping of comparable clusters that are thematically highly coherent. After recalling the concept of comparability we are using, and defining the two alternative variants we propose to overcome some limitation of the original measure, we detail our strategy to combine similarities and comparability in an efficient way that allows for the development of consistent co-clustering and co-classification of bilingual data sets. We then propose a quite exhaustive experimentation on a data

• *M. Ke, Marteau and Menier are with IRISA (UMR 6074), Universite de Bretagne Sud, 56000 Vannes, France.*
E-mail: firstname DOT name AT univ-ubs DOT fr

set collected from some Wikipedia categories. Basically, we evaluate jointly the three tested comparability measures and the similarities-comparability mixing strategy we propose in the scope of co-classification and co-clustering tasks. Finally we discuss our results and draw some perspectives.

2 Variations around of a quantitative comparability measure

2.1 Comparability measure of Li and Gaussier (C_{LG})

The first quantitative comparability measure proposed by [1] is based on the simple counting of *word translation connections* that exist between two corpora in different languages according to a translation lexicon. Formally, let C_1 and C_2 be two corpora expressed respectively in language $\mathcal{L}_1$ and $\mathcal{L}_2$. This comparability measure is formally defined as:

$$C_{LG}(C_1, C_2) = \frac{\sum_{w_1 \in WC_1 \cap WD_1} \sigma(w_1) + \sum_{w_2 \in WC_2 \cap WD_2} \sigma(w_2)}{|WC_1 \cap WD_1| + |WC_2 \cap WD_2|} \quad (1)$$

where: WC_i, $i \in \{1, 2\}$ is the lexicon in language $\mathcal{L}_i$ associated with the corpus C_i ; WD_i is the set of entries in language $\mathcal{L}_i$ into the bilingual dictionary that occur in WC_i; $\sigma(w_i)$ is an indicator function that takes the value 1 if at least one potential translation of the term $w_i \in WC_i$ in language $\mathcal{L}_i$ exists in the vocabulary associated with the corpus of the other language, 0 otherwise.

2.2 Enrichment of the C_{LG} measure

The C_{LG} measure proposed by Li and Gaussier (eq.1) takes account of neither the number of occurrences of the lexical entries in the documents nor their number of translations into the paired documents. The binary presence or absence of joint translation entries that is modeled by the indicator function $\sigma(w_i)$ is a strong feature that may affect the average comparability between pairs of documents. This could be the case when addressing corpora for which frequency of lexical entries helps discriminating between genres and topics. We propose the following two similar variants of the C_{LG} measure that explicitly propose to go beyond the presence or absence of joint translations, conjecturing that this improvement will produce a positive effect in certain situations and tasks.

2.2.1 First variant : C_{VA_1}

The first variant symmetrically exploits (from the stand point of $\mathcal{L}_1$ and $\mathcal{L}_2$ languages) the following three elements: the number of occurrences of entries w taken into the vocabulary of the first language corpus, the number of their translations in the bilingual dictionary and the presence of at least one of their translations in the vocabulary of the second language corpus.

Let $A_{1|2}$, A_1, $A_{2|1}$, A_2 be defined as follows:

$$A_{1|2} = \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, \mathcal{C}_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right)$$
$$A_1 = \sum_{w_1 \in WC_1 \cap WD_1} \left(\frac{tf(w_1, \mathcal{C}_1)}{\tau(w_1, WD_1)} \right)$$
$$A_{2|1} = \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, \mathcal{C}_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right)$$
$$A_2 = \sum_{w_2 \in WC_2 \cap WD_2} \left(\frac{tf(w_2, \mathcal{C}_2)}{\tau(w_2, WD_2)} \right)$$

where $tf(w_i, \mathcal{C}_i)$ is the number of occurrences of entry w_i in the corpus $\mathcal{C}_i$ expressed in language $\mathcal{L}_i$, $i \in \{1, 2\}$; $\tau(w_i, WD_i)$ is the number of translations of entry w_i of the corpus $\mathcal{C}_i$ in the dictionary WD_i; $\sigma(w_i)$ is defined as above.

$$C_{VA_1} = \frac{1}{2} \cdot \left(\frac{A_{1|2}}{A_1} + \frac{A_{2|1}}{A_2} \right) \quad (2)$$

2.2.2 Second variant : C_{VA_2}

This second variant is very similar to the previous one. It distinguishes mainly on the way the measure is symmetrized. Basically the first variant relates to a geometric mean while the second variant relates to an arithmetic mean.

$$C_{VA_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2} \quad (3)$$

3 Combining Similarities and comparability

There is no existing direct measure or method to map comparable clusters of documents that partition in two different linguistic spaces. Indeed, there exists some work which is somehow correlated to our needs, like biclustering, co-clustering, or two-mode clustering introduced by [7] and [8]. However, these works are mainly relevant to the clustering of the rows and columns (objects and features axes) of a given matrix.

Recently, [9], [10] have developed quite successfully a supervised method that learns interlingual representations from aligned training documents. They exploit word association measures and bilingual dictionary to remove noisy pairs of aligned documents. [11] have proposed a solution for clustering bilingual corpora by using the comparability measure only. However, our approach is quite different since it seeks the joint clustering or classification of data in two distinct spaces, in which *native* similarity matrices exist (a native similarity has to be understood as any quantitative intra-language similarity measure, such as a cosine similarity measure). Our aim is to exploit the comparability measure that maps the two linguistic spaces to provide new similarity measures that combine *native* similarities with a similarity measure that is *induced* by the comparability mapping.

Our approach only rely on a bilingual dictionary and does not assume that any aligned data preexist as learning

data. Indeed, it could be enriched using feature-extraction technique, such as the one proposed in [12] for instance, to align bilingual documents that have a similar content.

3.1 Similarity measure *induced* by a comparability mapping

In [13] the authors proposed an algorithm, $Hit-ComSim$, to iteratively construct the concept of similarity *induced* by a comparability bipartite graph. Unfortunately, this algorithm does not scale well due to its high algorithmic complexity in $O(N^4)$. We propose here a much more straightforward approach consisting in exploiting directly the comparability matrix constructed from the two bilingual collections of documents.

Let us consider $\mathcal{C}_1$ and $\mathcal{C}_2$ two collections of documents belonging to two distinct linguistic spaces ($\mathcal{L}_1$ and$\mathcal{L}_2$ respectively) in which two *native* similarity measures $S_{\mathcal{C}_1}$ and $S_{\mathcal{C}_{\mathcal{L}_2}}$ are defined. Let $C(.,.) : S_{\mathcal{C}_1} \times S_{\mathcal{C}_2} \rightarrow \mathcal{R}$ be the comparability matrix that maps the two finite collections.

We define the similarity measure *induced* by the comparability mapping C as the following normalized (in $[0, 1]$) measures respectively noted $S_{\mathcal{C}_1,C}$ and $S_{\mathcal{C}_2,C}$:

$$\begin{aligned} &\forall (d_i, d_j) \in \mathcal{C}_1^2 \text{ and } \forall (d'_i, d'_j) \in \mathcal{C}_2^2 \\ &S_{\mathcal{C}_1,C}(d_i, d_j) = \frac{CC^T(i,j)}{\sqrt{CC^T(i,i)CC^T(j,j)}} \\ &S_{\mathcal{C}_2,C}(d'_i, d'_j) = \frac{C^TC(i,j)}{\sqrt{C^TC(i,i)C^TC(j,j)}} \end{aligned} \quad (4)$$

The interpretation of the similarities *induced* by a comparability mapping is straightforward. First, considering each row i of the C matrix as a feature vector that characterizes document $d_i \in \mathcal{C}_1$, for any $(d_i, d_j) \in \mathcal{C}_1$, $CC^T(i,j)$ can be interpreted as an inner product between the two feature vectors representing d_i and d_j respectively. Then, $S_{\mathcal{C}_1,C}(d_i, d_j)$ is nothing but a cosine similarity between documents d_i and d_j based on the comparability mapping only. Similarly, considering each column i of the C matrix as a feature vector that characterizes document $d'_i \in \mathcal{C}_2$, $S_{\mathcal{C}_2,C}(d'_i, d'_j)$ is nothing but a cosine similarity between documents d'_i and $d'_j \in \mathcal{C}_2$ based on the comparability mapping only.

3.2 Mixing *native* similarities and *induced* similarities

The comparability/similarity mixing model we propose is a simple linear combination of the *native* and *induced* similarities defined in each linguistic space. Basically we use a single parameter $\alpha \in [0, 1]$ to combine linearly the two measures as follows

$$\begin{aligned} S'_{\mathcal{C}_1}(d_i, d_j) &= \alpha S_{\mathcal{C}_1,C}(d_i, d_j) + (1-\alpha) S_{\mathcal{C}_1}(d_i, d_j) \\ S'_{\mathcal{C}_2}(d'_i, d'_j) &= \alpha S_{\mathcal{C}_2,C}(d'_i, d'_j) + (1-\alpha) S_{\mathcal{C}_2}(d'_i, d'_j) \end{aligned}$$

Since the *induced* similarities are normalized into the unit interval $[0, 1]$, we advocate using cosine similarities as *native* similarities in the two connected linguistic spaces such that the mixed similarities defined by equation 5 are consistent.

4 Corpora and preprocessing

We collected the corpora from 21 Wikipedia categories, from English (EN) and French (FR) languages. It originally consists of 154828 documents in total with 87793 English documents and 67035 French documents categorized in 21 classes, taken from existing Wikipedia categories. Since such corpus is thematically very large, corresponding similarity and comparability matrices are basically very sparse. To avoid the algorithmic complexity behind the calculation of the induces similarity matrices ($O(N^3)$), we proceeded as follows which drastically reduces the sparsity of our matrices:

1) For each class and each language, we evaluate firstly the intra-lingual similarity matrices,

2) secondly, we prune these intra-lingual similarity matrices using a threshold (typically 0.5) and order the documents according to their number of remaining neighbors (with whom they share a similarity above the threshold).

3) by keeping for each language the best hundred documents, we get a refined corpus.

4) Finally, to complexify the experiment, we enrich this corpus by adding, for each language, and for each class, 50% of the initial number of documents. These added documents are randomly drawn from the initial 21 Wikipedia categories. Our Wikipedia corpus[1] contains 5822 documents in total, and is composed with 2745 French documents and 3077 English documents distributed into the 21 categories as listed in Table 1.

EN classes	# doc	FR classes	# doc	EN classes	# doc	FR classes	# doc
Astronomy	151	Astronomie	123	Movie	151	Film	151
Biology	151	Biologie	115	Music	151	Musique	151
Economy	144	Economie	151	Skating	151	Patinage	151
Food	147	Nourriture	4	Heritage	151	Patrimoine	151
Football	151	Football	151	Politics	151	Politique	151
Genetics	82	Génétique	151	Religion	150	Religion	133
Geograpphy	139	Geographie	151	Rugby	151	Rugby	151
Computer	151	Ordinateur	151	Health	151	Santé	63
Literature	150	Littérature	151	Sculpture	151	Sculpture	151
Mathematics	151	Mathématique	63	Tennis	151	Tennis	151
Medicine	151	Médecine	130				

TABLE 1
Composition of the corpus extracted from Wikipedia (EN: English, FR: French)

This corpus has been finally lemmatized using the TreeTagger [14] [15] and the term frequencies (*tf*) for each vocabulary entry/document pair has been evaluated, as well as the *idf* [16] that was estimated on the corpus.

4.1 Bilingual dictionary

To evaluate the quantitative comparability between a pair of Enlish/French documents we have used the bilingual dictionary available at ELRA under reference ELRA-M0033. This dictionary contains 243,580 pairs of lexical entries in French and in English, which decompose into 110,541

1. The Wikipedia corpus is available at http://people.irisa.fr/Pierre-Francois.Marteau/Corpora/Wikipedia_21classes.zip

lexical entries in English and 109,196 lexical entries in French.

4.2 Evaluation measures

The performance of the 1-NN classifier is evaluated using the error rate measure. The performance of the tested clustering algorithms are also evaluated by comparing the predicted label for each document with its *true*. The accuracy (AC) and normalized mutual information (NMI) measures are used to evaluate the clustering performance [17]. As an internal evaluation scheme for estimating the quality of the clustering obtained in each each linguistic space, we also use the Davies–Bouldin index (DB) [18] which roughly measures the quotient of intra and inter cluster average similarities.

The accuracy (AC) measure is defined as follows: it measures the fraction of documents that are correctly labels, assuming a one-to-one correspondence between true classes and assigned clusters. Let p denotes any possible permutation of index set of clusters and *true* classes. The Accuracy is thus defined as

$$AC = \frac{1}{N} MAX_p \sum_{i=1\cdots K} n_{i,p(i)} \quad (5)$$

where $n_{i,p(i)}$ denotes the number of documents shared by class i and cluster $p(i)$, K is the number of classes and clusters, and N is the total number of documents.

The NMI measure between the *true* clustering $\mathcal{C}$ and the predicted one $\tilde{\mathcal{C}}$ is defined as follows:

$$NMI(\tilde{\mathcal{C}}, \mathcal{C}) = \frac{I(\tilde{\mathcal{C}}, \mathcal{C})}{(H(\tilde{\mathcal{C}}) + H(\mathcal{C}))/2} \quad (6)$$

with

$$I(\tilde{\mathcal{C}}, \mathcal{C}) = \sum_k \sum_j P(\tilde{c}_k \cap c_j) \log \frac{P(\tilde{c}_k \cap c_j)}{P(\tilde{c}_k) P(c_j)}$$

and

$$H(\tilde{\mathcal{C}}) = -\sum_k P(\tilde{c}_k) \log P(\tilde{c}_k)$$
$$H(\mathcal{C}) = -\sum_k P(c_k) \log P(c_k)$$

The Davies-Boulding index DB is a data intrinsic evaluation measure, which is defined as follows

$$DB = \frac{1}{K} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \quad (7)$$

where K is the number of clusters, C_k is the centroid of cluster, σ_k is the average distance of all elements in cluster k to centroid c_k, and $d(c_i, c_j)$ is the distance between centroids i and j. The lower is this DB index value, the better is the clustering since this corresponds to low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity).

Finally, the quality of the cluster mapping or cluster mapping is evaluated through the calculation of the inter cluster average comparability matrix and the resulting cluster mapping (i.e. a bipartite graph).

5 Experiments

On the basis of the previous categorized comparable corpora, we assess the benefit of mixing native similarities with comparability on a 1-NN classification task and on k-medoid clustering [19] [20] task.

5.1 1-NN classification task

We first study the effect of mixing similarity and comparability on the 1-NN classification error rate while varying the parameter $\alpha \in [0, 1]$.

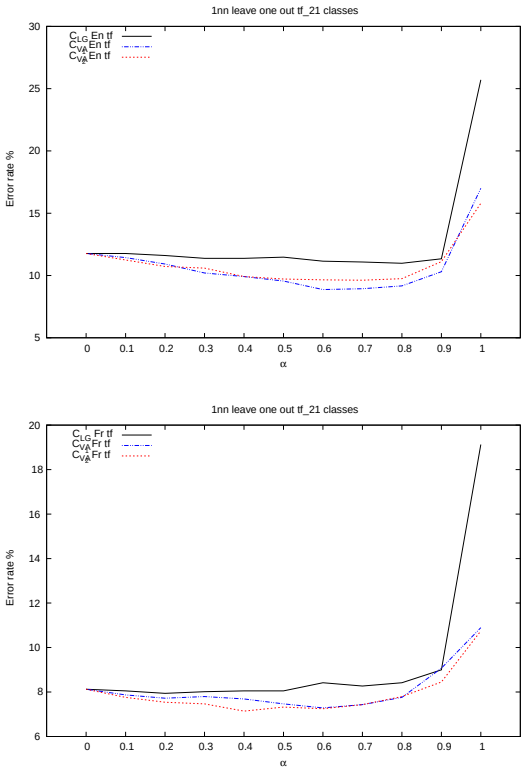

Fig. 1. Comparability/similarity mixing effect on the 1-NN classification task, according to the leave one out error rate (top EN documents, bottom FR documents)

Figures 1 and 2 show that the similarity/comparability mixing has a significant impact for the two variants C_{VA_1} and C_{VA_2} since it allows reducing by 3% the error rate of the classification for the English language documents and 1.5% for the French language documents. However, comparatively, the C_{LG} measure improves poorly for both languages the classification accuracy, and is less stable when α varies.

5.2 k-medoids clustering task

We study here the effect of mixing comparability and similarities on a k-medoids clustering task for all three comparability measures. We used the previously defined AC, NMI and DB measures for the assessment of this clustering task.

Figures 3 and 4 show that both AC and NMI measures can be improved up to 15% in the scope of the clustering

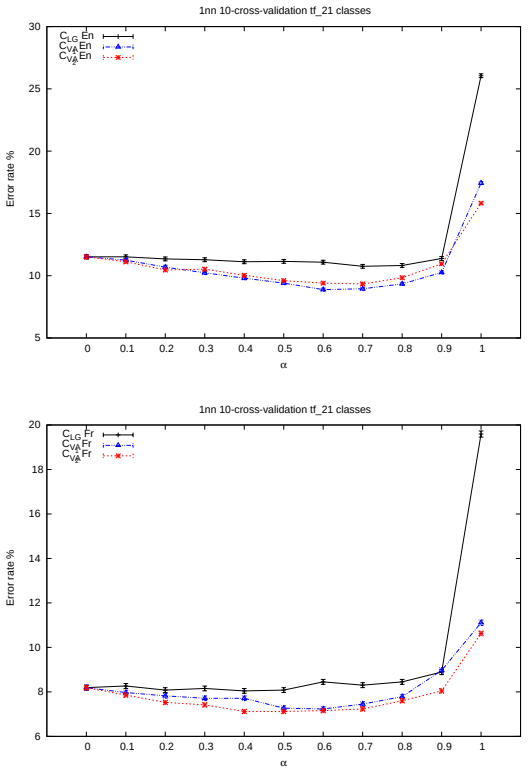

Fig. 2. Comparability/similarity mixing effect on the 1-NN classification task, according to 10^{th}-cross-validation error rate (top EN documents, bottom FR documents)

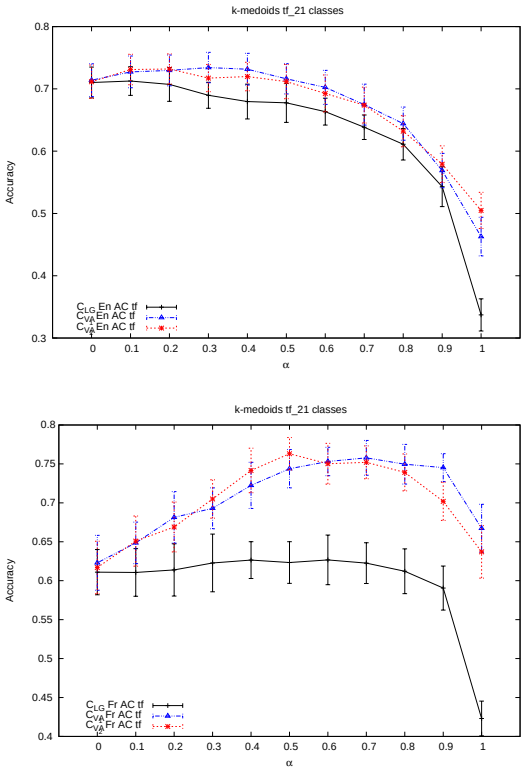

Fig. 3. Evaluation of the comparability/similarity mixing on the k-medoids clustering accuracy (AC) (top EN documents, bottom FR documents).

of French language documents and up to 3% in the scope of the clustering of English language documents for both C_{VA_1} and C_{VA_2} measures. However, once again, the C_{LG} brings comparatively few improvement for both languages.

Figure 5 depicts the DB measure as a function of parameter α, for all three comparability measures. It is shown that, for C_{VA_1} and C_{VA_2}, this ratio decreases for some *good* α values, especially for the French language, whereas for the measure C_{LG}, this value increases in general. A *good* mixing of the comparability and similarities has thus a positive impact when using C_{VA_1} and C_{VA_2} measures and a rather negative impact when using the C_{LG} measure.

6 Analysis and Conclusions

In this paper, we have proposed a new approach for the clustering and categorization of bi-lingual data. This approach is based on the concept of similarity *induced* by a comparability bipartite graph. It involves a quantitative comparability measure that is based on the exploitation of a bilingual dictionary. The implementation of our approach on semi-manually constructed comparable corpora collected from the Web (from Wikipedia) shows to be quite effective for the building of comparable corpora that are thematically coherent. Our detailed experimentation shows that the mixing of *native* similarities with quantitative comparability has a significant impact on the classification and clustering accuracies. Our approach works specifically well for the C_{VA_1} and C_{VA_2} comparability measures with stable and robust classification or clustering result improvements. It nevertheless has a poor positive impact on the C_{LG} measure, leading to conclude that taking into account of the frequency of occurrence of lexical entries and frequencies of their translations into the comparability measure design is of crucial importance for thematic classification or clustering of bilingual English/French documents. One potential explanation is that these frequencies of occurrence pair well with the *tf* heuristic that takes place in native *cosine* similarities. Moreover, according to our results, the choice of the value of combination parameter α is quite important. An α value relatively high (between 0.5 and 0.8), that slightly favor the *induced* similarities, will be a good choice in general. Finally our experimentation shows that the C_{VA_2}, whose symmetrization is homogeneous to an arithmetic mean, is more robust than C_{VA_1}, a result that need to be consolidated on other independent experiments.

In terms of perspective, ensuring the scalability and generalizing the approach and experimentation are major prospects to help constructing thematic comparable corpora on demand. Another objective is to expand it to various pairing of languages for which bilingual resources are available, in particular bilingual dictionaries.

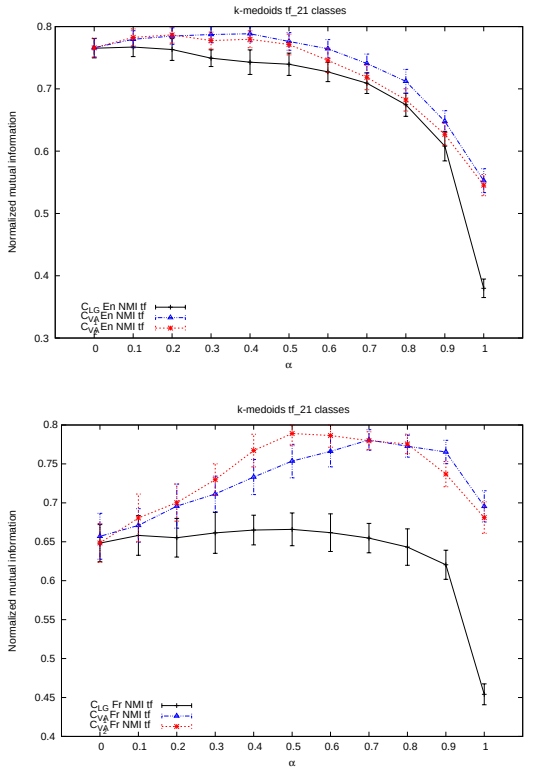

Fig. 4. Evaluation of the mixing of comparability and similarities on the k-medoids clustering according to the NMI measure (top EN documents, bottom FR documents)

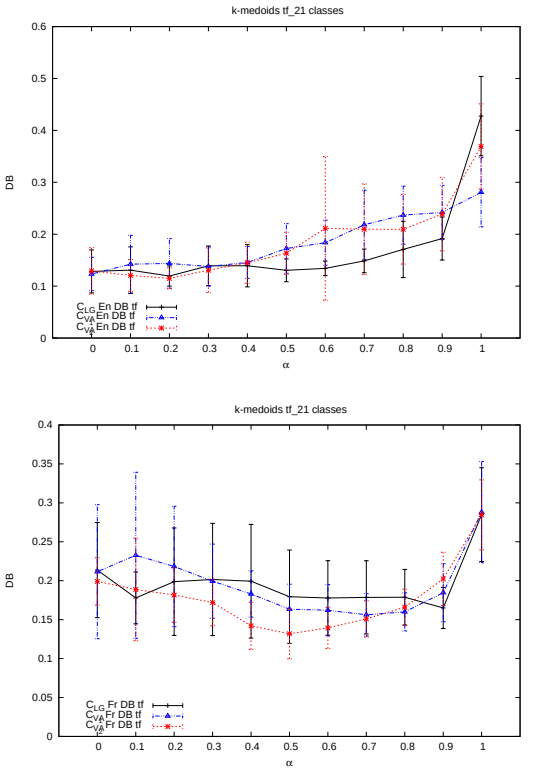

Fig. 5. Comparability/similarity mixing effect on a k-medoids clustering according to the DB measure (top EN documents, bottom FR documents)

REFERENCES

[1] B. Li and E. Gaussier, "Improving corpus comparability for bilingual lexicon extraction from comparable corpora," in *COLING*, 2010, pp. 644–652.

[2] M. Baker, "Corpus-based translation studies: The challenges that lie ahead." in *In Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, 1996.

[3] EAGLES, "Expert advisory group on language engineering standards guidelines: http://www.ilc.pi.cnr.it/eagles96/browse.html," EAGLES, Tech. Rep., 1996.

[4] P. Fung and L. Y. Yee, "An ir approach for translating new words from nonparallel, comparable texts," in *Proc. of the 36th ACL meeting, Vol. 1*, ser. ACL '98. Stroudsburg, PA, USA: ACL, 1998, pp. 414–420. [Online]. Available: http://dx.doi.org/10.3115/980845.980916

[5] D. S. Munteanu, A. Fraser, and D. Marcu, "Improved machine translation performance via parallel sentence extraction from comparable corpora," in *HLT-NAACL*, 2004, pp. 265–272.

[6] H. Déjean and E. Gaussier, "Une nouvelle approche a l'extraction de lexiques bilingues à partir de corpus comparables," *Lexicometrica*, vol. Numéro spécial, corpus alignés, pp. 1–22, 2002.

[7] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Publishers, 1996.

[8] D. B. P. Van Mechelen I, Bock HH, "Two-mode clustering methods:a structured overview," *Statistical Methods in Medical Research*, vol. 13(5), pp. 363–394, 2004.

[9] J. Jagarlamudi, H. Daumé, III, and R. Udupa, "From bilingual dictionaries to interlingual document representations," in *Proc. ACL-HLT - Vol. 2*, ser. HLT '11. Stroudsburg, PA, USA: ACL, 2011, pp. 147–152. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002768

[10] J. Jagarlamudi, R. Udupa, H. Daumé, III, and A. Bhole, "Improving bilingual projections via sparse covariance matrices," in *Proc.s of the Conf. on EMNLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 930–940. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145534

[11] B. Li, E. Gaussier, and A. Aizawa, "Clustering comparable corpora for bilingual lexicon extraction," in *Proc. of the 49th ACL-HLT- Vol. 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 473–478. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002831

[12] T. Vu, A. T. Aw, and M. Zhang, "Feature-based method for document alignment in comparable news corpora," in *Proceedings of the 12th EACL Conf.* Stroudsburg, PA, USA: ACL, 2009, pp. 843–851. [Online]. Available: http://dl.acm.org/citation.cfm?id=1609067.1609161

[13] P.-F. Marteau and G. Ménier, "Similarités induites par mesure de comparabilité : signification et utilité pour le clustering et l'alignement de textes comparables." in *TALN*, 2013, p. 515–522.

[14] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of the Int. Conf.e on New Methods in Language Processing*, 1994, pp. 44–49. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.1139

[15] ——, "TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/," 2009.

[16] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972. [Online]. Available: http://www.soi.city.ac.uk/~ser/idf.html

[17] X. L. Wei Xu and Y. Gong, "Document clustering based on non-negative matrix factorization," in *SIGIR'03*, 2003, pp. 267–273.

[18] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. PAMI-1(2), pp. 224–227, 1979.

[19] L. Kaufman and P. J. Rousseeuw, *Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods*. North-Holland, 1987.

[20] ——, *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley and Sons, 1990.