Co-clustering of bilingual datasets as a mean for assisting the construction of thematic bilingual comparable corpora

Guiyao Ke, Pierre-Francois Marteau
UMR CNRS 6074 IRISA, Université de Bretagne Sud
Campus de Tohannic, 56000 Vannes, France
ke@univ-ubs.fr, marteau@univ-ubs.fr

Abstract

We address in this paper the assisted construction of bilingual thematic comparable corpora by means of co-clustering bilingual documents collected from raw sources such as the Web. The proposed approach is based on a quantitative comparability measure and a co-clustering approach which allow to mix similarity measures existing in each of the two linguistic spaces with a "thematic" comparability measure that defines a mapping between these two spaces. With the improvement of the co-clustering (k-medoids) performance we get, we use a comparability threshold and a manual verification to ensure the good and robust alignment of co-clusters (co-medoids). Finally, from any available raw corpus, we enrich the aligned clusters in order to provide "thematic" comparable corpora of good quality and controlled size. On a case study that exploit raw web data, we show that this approach scales reasonably well and is quite suited for the construction of thematic comparable corpora of good quality.

Keywords: Thematic comparable corpora, Comparability measure, Co-clustering, Cluster alignment

1. Introduction

The rapid growth of information sources available on the Internet provides a real and appealing opportunity for the construction of comparable corpus. In particular the news pages produced by news agencies in various languages, or Wikipedia articles constitute rich exploitable multilingual resources and generally free of copyright. With the increasing needs for comparable corpora, their quality become critical. The main issue in the construction of a bilingual "thematic" comparable corpus is the alignment between source language documents (or clusters of documents) and target language documents (or clusters of documents). The more similar or comparable the aligned documents or clusters are, the more the produced comparable corpus will be usable by end applications (e.g. terminology extraction or for cross language information retrieval).

Much research has been conducted to build comparable corpora. At first, quite rudimentary approaches have been exploited. For example, (Sheridan and Ballerini, 1996) simply used the publication date and similar thesaurus (considering documents as the indexing features and terms as retrieval elements) to build the alignment relationship between Italian and German texts. On this basis, (Braschler and Scäuble, 1998) integrated an indicator in the construction of comparable English (published by AP: Associated Press) and German (published by the Swiss agency SDA) corpora. This indicator corresponds to the word that has the average frequency in all of the English texts. It is then translated using the English-German bilingual dictionary and used as a query in the German corpus. The obtained similarities and dates are used to organize the comparable corpus. Moreover, (Resnik, 1999) proposed an approach to search the comparable corpus according to the following hypothesis: if the content of documents (web pages) in different languages are similar, they must have a similar structure, such as titles, paragraphs, etc. We can see that initially, the construction of comparable corpora is relatively empirical and heterogeneous. These approaches do not take much account of the quality of the text alignments obtained at the end of the construction process.

More recently, (Tao, 2005) proposed an approach based on the correlation of the words frequencies in the context of a common theme expressed in different languages in the comparable corpora, under the assumption that the distributions of thematic word frequencies in different languages are often correlated. (Munteanu et al., 2004) is the first to use a bilingual dictionary to transform the source texts into the target language texts. The first five translations (top-5) is then used as a query to search into language target texts on a same period. Based on the obtained similarities, the first K documents in the target language are selected by grouping pairs of similar text from 1 to K. Moreover, (Talvensaari et al., 2007) used cross-language information retrieval technique to build an English-Swedish comparable corpus. To avoid the translation of the entire text, only the relevant information is extracted and translated, and then searched using the information retrieval system. To improve the alignment quality, the obtained results are then filtered. (Otero and López, 2009) collected comparable corpus from Wikipedia by defining a theme and two languages (the source language and the target language) to collect similar documents for the selected theme. In addition, (Vu et al., 2009) proposed an alignment approach based on document characteristics like TNC (title and content), LIU (language independent unit) and MTD (monolingual terms distribution).

In this paper we develop an approach dedicated to assist the construction of "thematic" comparable corpora. We first define what we mean by "thematic" comparable corpora, then briefly recall the comparability measures we propose to overcome some limitation (for the considered task) of the quantitative measure proposed by (Li and Gaussier, 2010) and our proposal for the co-clustering of bilingual documents. We then detail our procedure for assisting the construction of thematic bilingual comparable corpora, present

some results that we discuss and finally suggest some perspectives.

2. "Thematic" comparable corpora versus general comparable corpora

We define a theme as a subset of documents featuring a shared vocabulary. A theme covers an idea, a topic which is developed in a text or a collection of texts. According to different possible presentations depending on the audience, different types exist for a same theme. A type is a production format that possesses formatting features and typed lexicogrammatical choices (Swales, 1990): for example a research paper, a vulgarization article, a newspaper article belong to different types of documents. Therefore a domain includes all specialized (or typed) themes.
Similarly to the definition of a "translational" bilingual comparable corpus proposed in (Déjean and Gaussier, 2002), we define a "thematic" bilingual comparable corpus as a set of multilingual documents that deal with the same theme. In particular, the (discriminative) terms characterizing the domain are expected to be frequent and lowly ambiguous into the corpus. An operational definition of the concept of "thematic" comparable bilingual corpora is thus expressed as follows: two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called "thematically" comparable if:

- on the one hand there is a significant subset 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 $\mathcal{L}_1$.
- on the other hand, the concerned terms of the vocabulary subset must be such that the ratio between their frequency of occurrence and their number of translations is the largest as possible (frequent and lowly ambiguous terms are expected to be characteristic of the theme).

Due to this new requirement on the shared vocabulary subset, the quality of the alignment of "thematic" comparable documents (or document clusters) seems to be more important an issue than the size of the comparable corpora itself. Indeed, some works show that if the size of a comparable corpus is sufficient, its quality is not necessarily decisive. For example, in the context of the extraction of bilingual lexicon from specialized comparable corpora, (Morin et al., 2007) has showed that, if the quality is more important than the size of the corpus in the case of complex term alignments, it is not the case for simple term alignments. However, the question remains open. In (Prochasson, 2009), the author has showed that a properly constituted comparable corpus is at least as effective as a comparable corpus less well constituted but larger in size, and that the frequencies of term co-occurrences are unstable, even in the case of highly comparable corpora (this phenomenon seems to be aggravated in the case of lowly comparable corpora). Furthermore, several works like (Talvensaari, 2008), (McNamee et al., 2009), (Li, 2012), tend to show that the alignment quality of comparable corpora is more important than their volume. Particularly in (Rahimi and Shakery, 2011), the authors showed that the quality of comparable corpora (they build two comparable corpora: a first corpus constructed from an alignment based on the similarity of the concepts present in the documents and the publication date, and a second corpus built from an alignment based on the similarities of theme and concepts with different publication dates in order to treat long duration events) significantly improves the performance of the extraction of word translations and cross-language information retrieval from the translated queries.
We therefore believe that the building of comparable corpora with a strong thematic coherence while maintaining the alignment quality as high as possible is particularly relevant. This is the main motivation behind the constructive approach that we detail hereinafter.

3. Assisted construction of thematic bilingual comparable corpora

We develop our approach for the construction of "thematic" comparable corpus from two earlier and complementary works: (Ke et al., 2014) in which we proposed the concept of "thematic" comparability measures and (Ke et al., 2013) in which we developed a co-clustering approach for bilingual data by mixing *native* (or *thematic*) similarities and similarities that are *induced* by a comparability measure.

3.1. background

The "thematic" comparability measure developed in (Ke et al., 2014) takes the following form:

$$C_{VA_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2} \tag{1}$$

with

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

where $W(w_i, \mathcal{C}_i)$ is a weight coefficient (basically the term frequency weighting is used); $\tau(w_i, WD_i)$ is the number of translations of the lexical entry w_i of the corpus $\mathcal{C}_i$ into the translation dictionary WD_i. $\sigma(w_i) = 1$ if at least one translation of the lexical entry $w_i \in WC_i$ in language $\mathcal{L}_i$ exists in the vocabulary associated with the other corpus, 0 otherwise.

The co-clustering of bilingual document proposed in (Ke et al., 2013) is defined as follows:
If we 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 that is *induced* by the comparability mapping C as the following normalized (in

$[0,1]$) measures respectively noted $S_{\mathcal{C}_1,C}$ and $S_{\mathcal{C}_2,C}$:

$\forall(d_i,d_j) \in \mathcal{C}_1^2$ and $\forall(d'_i,d'_j) \in \mathcal{C}_2^2$

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

From the *native* and *induced* similarities, a simple mixing model is derived as a 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 *mixed* measures as follows

$$\begin{aligned} S'_{\mathcal{C}_1}(d_i,d_j) &= \alpha S_{\mathcal{C}_1,C}(d_i,d_j) + (1-\alpha)S_{\mathcal{C}_1}(d_i,d_j) \\ S'_{\mathcal{C}_2}(d'_i,d'_j) &= \alpha S_{\mathcal{C}_2,C}(d'_i,d'_j) + (1-\alpha)S_{\mathcal{C}_2}(d'_i,d'_j) \end{aligned} \quad (3)$$

Then the clustering in the two linguistic spaces can be applied to obtain a co-clustering of bilingual data.

3.2. Assisted construction approach

Starting from a raw bilingual corpora $\mathcal{BC}_0$ collected from the web for instance, the proposed approach consists of the following six steps:

STEP-1: Calculation and construction of the comparability matrix for the English and French documents of the initial *raw* corpus $\mathcal{BC}_0$.
The "thematic" comparability measure C_{VA_2} (Eq.1) is used to calculate comparability between pairs of documents in two different languages. The complexity to evaluate the comparability matrix is quadratic ($O(|\mathcal{BC}_0|^2)$) with the size of the initial corpus.

STEP-2: Filtering of the initial corpus $\mathcal{BC}_0$ and production of a bilingual corpus $\mathcal{BC}_1$ with higher comparability.
This step aims at extracting the most comparable pairs of aligned documents. It ensures also that an acceptable computational cost is maintained. The mixture model (Eq.3) that will be used in the third step is characterized by a $O(n^3)$ complexity, where n is the number of documents in the processed corpus. Hence, we assume that the size of the corpus $\mathcal{BC}_1$ produced at the end of this step is substantially smaller than the size of the initial corpus $\mathcal{BC}_0$ ($|\mathcal{BC}_1| \ll |\mathcal{BC}_0|$).
This step involves sorting and filtering documents from the comparability matrix calculated during the previous step, using a minimum comparability threshold, β, a threshold γ that characterizes the minimum nodes degree of the comparability bipartite graph obtained after pruning the links associated with a comparability below the threshold β. A third parameter $\sigma = |\mathcal{BC}_1|/2$ defines the desired filtered corpus size in each language.
We present below the sorting method and the document filtering process that we propose. The sorting is simultaneously performed on the rows and columns of the comparability matrix (see in Figure 1):
a) We calculate for each row i of the comparability matrix, the number nl_i of comparability values which are above the threshold β and we keep only the lines i for which $nl_i > \gamma$.

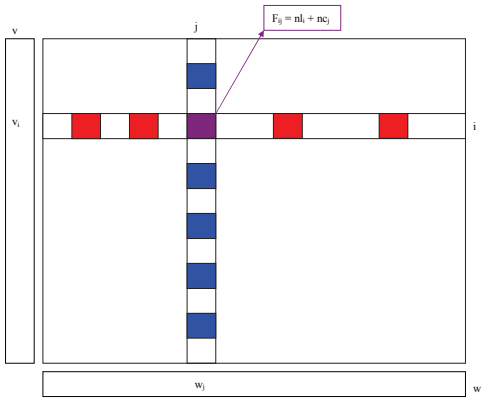

Figure 1: Sorting of pairs of documents based on the calculation of the matrix $F_{ij} = nl_i + nc_j$ and vectors $v_i = Max_j\{F_{ij}\}$ and $w_j = Max_i\{F_{ij}\}$

b) Similarly, we calculate for each column j of the comparability matrix, the number nc_j of comparability values which are bigger than the threshold β and we keep only the columns j for which $nc_j > \gamma$.
c) We perform the sum of the number of values of each row i and of each column j and build the matrix $F_{ij} = nl_i + nc_j$ if the row i and column j is stored, 0 otherwise.
d) We then calculate the vectors v and w ($v_i = Max_j\{F_{ij}\}$, $w_j = Max_i\{F_{ij}\}$)
The v and w vectors are then sorted in descending order (which is done in $O(n \times log(n))$ complexity) and we then select only the first σ values for each of them.
e) Finally, the corpus $\mathcal{BC}_1$ consists of bilingual documents that correspond to the raw and column indexes (i and j) that are retained.

We then extract the comparability matrix for the corpus $\mathcal{BC}_1$ and calculate the corresponding *native* similarity, that is nothing but a cosine similarity based on a vector model based on term frequency weights, and the *induced* similarity matrices (Eq.2) for each linguistic sub-corpus.

STEP-3: Selection of the initial number of clusters (K_0).
Exploiting, for each language, the outputs produced by a k-medoids clustering based on the comparability/similarity mixing model (Eq.3) while varying the number of clusters k, we perform the calculation of the average intra and inter clusters similarities (δ_{intra} and δ_{inter} resp., Eq.4) to determine empirically an initial number of clusters K_0. If the k-medoids clustering is performed independently for each language, the mixing model of *native* and *induced* similarities is exploited, which thus preserves the characteristic of a bilingual co-clustering. δ_{intra} and δ_{inter} for a clustering $\mathcal{C}$ containing Nc clusters are defined as follows:

$$\begin{aligned} \delta_{intra}(\mathcal{C}_l) &= \frac{1}{Nc}\sum_{i=1}^{Nc}\left(\frac{1}{|C_i|}\sum_{d,d' \in C_i} S'_{\mathcal{C}_l}(d,d')\right) \\ \delta_{inter}(\mathcal{C}_l) &= \frac{1}{Nc(Nc-1)}\sum_{i=1}^{Nc}\left(\sum_{j}^{Nc} S'_{\mathcal{C}_l}(m_i,m_j)\right) \end{aligned} \quad (4)$$

where $S'_{\mathcal{C}_l}(d,d')$ and $S'_{\mathcal{C}_l}(m_i,m_j)$, are similarities provided by the *native* and *induced* similarities mixing model

(Eq.3, with $l \in \{1, 2\}$ defining the language), as defined in (Ke et al., 2013). m_i is the medoid of cluster i, that is to say the element of cluster i for which the average similarity to all other elements of the cluster is maximal (or the most central element in the cluster). After this step, from the curves δ_{intra} and δ_{inter} considered as a function of k, K_0 is manually set (using an empirical criteria that is described in the experimental section).

STEP-4: Filtering the most highly comparable cluster pairs.

From a set of k-medoids clustering with $k = K_0$, the objective here is to identify the cluster pairs that will be retained as well as the documents that should be kept within each of the retained clusters. In practice, we set a second inter-cluster comparability threshold φ, then we calculate the average degree of the bipartite graph obtained from the aligned clusters after pruning the links associated with a comparability below the threshold value φ. The evolution of the average degree of the graph, when the comparability threshold φ varies, provides a decision criterion for the choice of clusters to be finally kept. Ideally, we target relatively *pure* cluster alignments, i.e. we seek a comparability relationship that tends to be a *1-to-1* mapping. This means that the average degree of the inter cluster comparability bipartite graph should tend towards 1, while limiting the number of *orphan* (non-aligned) clusters.

Because of the dependence of the k-medoids algorithm to initial conditions, we process not just one clustering, but a series of clustering as mentioned above.

The analysis of the impact of the comparability threshold φ on - 1) the number of retained clusters, 2) the number of retained documents and 3) the degree of the bipartite graph of aligned clusters - provides some useful information for the manual selection of a comparability threshold φ balancing the size of the corpus and the aligned clusters average comparability.

Once the comparability threshold φ has been selected, we align the obtained clusters by using the *native* and *induced* similarities mixing model. Running several times the k-medoids algorithm, in order to vary the initial conditions, leads to get a larger set of aligned cluster pairs.

By the end of this step, we obtain a corpus $\mathcal{BC}_2$ consisting of aligned bilingual cluster pairs.

STEP-5: Manually verification of the aligned clusters.

This step aims at manually validating (or invalidating) the aligned cluster pairs of corpus $\mathcal{BC}_2$. The medoids that are judged correctly aligned are kept. Duplicates of medoids are eliminated, possibly at this step some cluster pairs can be merged.

This step must generally involve a thematic and linguistic expertise. We obtain a validated corpus $\mathcal{BC}_3$ by the end of this verification step.

STEP-6: Corpus enrichment.

The purpose of the enrichment phase is to increase the corpus size once the "thematic" clusters have been extracted. The enrichment phase is indeed relevant when the corpus $\mathcal{BC}_3$ is estimated to be too small for the targeted application. We exploit here two variants to estimate the degree of adequacy $S(d, m_i, m_j)$ between a document d, candidate for the enrichment, and an aligned medoid pair (m_i, m_j).

a) the first alternative simply merges, using the mixing parameter α, the *native* similarity between the candidate document and the medoid of same language with the comparability between the candidate document and the medoid of the other language. For d and m_i belonging to the same linguistic space (l), we thus have:

$$S_{v1}(d, m_i, m_j) = \alpha S_{\mathcal{C}_l}(d, m_i) + (1 - \alpha)C(d, m_j) \quad (5)$$

where C stands for the comparability measure.

b) the second alternative uses the mixing model obtained from the document d and the set of considered aligned medoids pairs. If C_d is the comparability matrix calculated on this basis, we have:

$$S_{v2}(d, m_i, m_j) = \alpha S_{\mathcal{C}_l}(d, m_i) + (1 - \alpha)S_{\mathcal{C}_l, C_d}(d, m_i) \quad (6)$$

For this second variant, the medoid m_j is taken into account through the matrix $C_d.C_d^T$ exploited to calculate the similarities *induced* by the comparability measure in the linguistic space l (Eq.2).

We use a reject threshold, τ, on S_{v1} or on S_{v2}, to decide if the document d will be finally retained to enrich a cluster or not. In practice, each document in the initial corpus $\mathcal{BC}_0$ is tested and will enrich the corpus if its S_{v1} or S_{v2} value is greater than the threshold τ. Any additional corpus can naturally be used to further enrich the corpus. The threshold τ can be adjusted according to the requirements expressed by the user in matter of average comparability and corpus size. If τ is low, we will get more documents in each cluster pair, but these documents will be less comparable in average. However, if τ is high, there will be less documents in each clusters pair, but these documents will be more comparable in average.

After the enrichment step is carried out, the final thematic bilingual comparable corpus, $\mathcal{BC}_F$ is produced.

This semi-supervised approach exploits 7 parameters that need to be setup carefully, depending on the user's need and the available initial resources. We recall these parameters synthetically hereinafter:

1. parameter α is used in our mixing model to merge *native* and *induced* similarities,
2. parameter β determines the minimum comparability value for filtering the raw corpus $\mathcal{BC}_0$ before extracting co-clusters,
3. parameter γ corresponds to the minimum degree of nodes (documents) in the comparability bipartite graph once a pruning conditioned by the threshold β (minimal comparability value) has been performed,
4. parameter σ determines the number of documents that are kept in the filtered corpus $\mathcal{BC}_1$ (in which comparability is maximized),
5. the parameter K_0 specifies the initial number of clusters extractable from the corpus $\mathcal{BC}_1$,

6. the parameter φ is a comparability threshold used for the extraction of the most similar cluster pairs that constitute the corpus $\mathcal{BC}_2$ (becoming $\mathcal{BC}_3$ after the manual verification step)

7. the parameter τ is used as a *fitness* threshold for adding suitable documents to enrich the corpus $\mathcal{BC}_3$ in order to produce the final bilingual corpus $\mathcal{BC}_F$ consisting, in principle, of highly comparable and "thematic" aligned clusters.

4. Case study

4.1. Experimental initial corpus

To test our semi-supervised approach in a real situation exploiting the Web, we used a crawler to collect, on a six-month period (from December 2012 to May 2013), documents from 23 RSS feeds listed in Table 1. The collected initial corpus $\mathcal{BC}_0$ is composed of 39,729 documents (18,168 English documents and 21,561 French documents). For each document, the *non stop* words are lemmatized by exploiting the TreeTagger (Schmid, 1994) (Schmid, 2009) and weighted according to the term-frequency weighting scheme (*tf*).

RSS feed	Language
www.globaltimes.cn/...	EN
www.shanghaidaily.com/...	EN
v1.theglobeandmail.com...	EN
www.thetimes.co.uk/...	EN
rss.nytimes.com/...	EN
feeds.washingtonpost.com/...	EN
feeds.latimes.com/...	EN
www.chinadaily.com.cn/...	EN
feeds.bbci.co.uk/...	EN
www.france24.com/...	EN
rss.cnn.com/rss/...	EN
www.abc.net.au/...	EN
liberation.fr.feedsportal.com/...	FR
www.lavenir.net/rss.aspx...	FR
www.ledevoir.com/rss/...	FR
www.lessentiel.lu/...	FR
rss.feedsportal.com/...	FR
www.romandie.com/rss/flux.xml	FR
rss.lemonde.fr/...	FR
www.courrierinternational.com/...	FR
feeds.lefigaro.fr/...	FR
www.lapresse.ca/...	FR
www.lesoir.be/...	FR

Table 1: List of the collected RSS feeds. All these feeds are from international (world/monde) newswires broadcasted by newspaper or tv networks in English (EN) or French (FR) languages.

The bilingual dictionary that we have used is available at ELRA under the reference ELRA-M0033: it contains 243,580 pairs of lexical entries in French and in English, which decompose into 110,541 lexical entries in English and 109,196 lexical entries in French.

5. Experiments and results

According to our experiments on co-clustering (Ke et al., 2013), we have predefined the following settings:

- a vector model based on a term frequency (*tf*) weighting for representing document content has been preferred to the *tf-idf* model.,
- the thematic comparability measure C_{VA_2} proposed in (Ke et al., 2014) is chosen (Eq.1),
- a median value for the mixture parameter (α=0,5) is selected. Basically the weight of the native and induced similarities are equal in the mixture model, which is generally a good compromise (Ke et al., 2013)

Due to the high complexity ($O(n^3)$) of the calculation of the *induced* similarities, we set the parameter σ (the number of most comparable documents that we initially keep for each language in the corpus $\mathcal{BC}_1$) to 1000. According to our previous experiments, we found that if the comparability value between two documents is greater than 0.1, the two documents may be relatively comparable. Therefore we assign the parameter β value, defining the minimum comparability, to 0.1 to obtain a compromise between the elimination of lowly comparable documents and the keeping of a sufficiently large number of documents. We set the parameter γ (the minimum number of comparability links above β) to 10 to ensure a minimum degree in the initial bi-partite comparability graph. The parameters β, γ and σ are defined in STEP-2.
The parameter K_0 (the initial clusters number (STEP-3)) and the parameter φ (the reject comparability threshold (STEP-4)) will be determined experimentally. Finally, the parameter τ (the enrichment threshold for adding documents) is used to produce the final corpus. It is adjustable by the user according to the requirements and the available processed data (STEP-6).
We then perform the experiment according to the six steps and get the following results.

5.1. Experiments on $\mathcal{BC}_1$

5.1.1. Determination of the initial clusters number K_0

We determine here an initial value K_0 for the co-clustering of $\mathcal{BC}_1$ by analyzing the variations of the average intra and inter similarities δ_{intra} and δ_{inter} obtained based on a k-medoids clustering (STEP-3) when k varies.

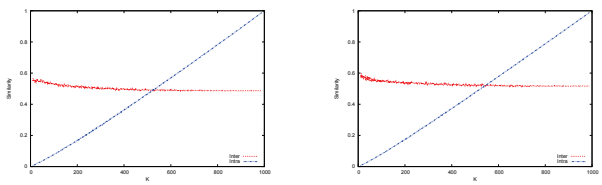

Figure 2: Variation of the intra et inter clusters similarities δ_{intra} and δ_{inter} on k-medoids clustering when k varies (EN left, FR right).

In Figure 2, we see that as k increases, the curves δ_{intra} and δ_{inter} are monotonic decreasing and increasing respectively. They intersect around $K_0 = 550$ for both languages. For a good clustering, it is generally necessary that the value of intra cluster similarity δ_{intra} is large and the value of inter cluster similarity δ_{inter} is small. The intersection point ($K_0 = 550$) of the two curves is empirically a good compromise.

5.1.2. Determination of the comparability threshold φ

Here we determine the comparability threshold φ, based on the number of retained clusters, the number of retained documents and the degree of the bipartite graph of aligned clusters as the comparability thresholds φ varies (STEP-4).

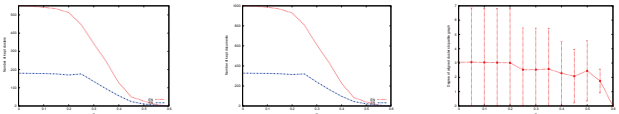

Figure 3: Determination of the comparability threshold φ according to the number of retained clusters (left), the number of retained documents (middle) and the degree of the bipartite graph of the aligned clusters (right).

In Figure 3, we try to determine a value for φ such that the selected clusters contain a sufficient large number of documents and simultaneously such that the comparability relationship between pairs of medoids tends towards a *1-to-1* mapping, i.e. the degree of the bipartite cluster graph tends towards 1. We found and verified that when φ is close to 0.45 all the three decision values (the retained number of clusters number, the retained number of documents number and the degree of the bipartite graph of aligned clusters) are stable. So we set the comparability threshold φ to 0.45. This threshold is used to automatically align bilingual clusters.

5.1.3. Aligned clusters pairs

Four successive executions of the k-medoids have been carried out and a reject comparability threshold $\varphi = 0.45$ (STEP-4) has been used. The clusters pairs have then been manually checked (STEP-5), and 16 pairs of clusters have been finally selected. Each of these retained pairs has one of the 16th highest comparability value. We present in Figure 4 as an example one of the 16 aligned cluster pairs (each cluster is represented by its medoid).

In Figure 5, we manually check the number of new discovered clusters and the number of already extracted (common) clusters based on four successive k-medoids clustering (STEP-5). According to the obtained results, the number of common clusters has a tendency to increase and the number of new discovered clusters has a tendency to decrease with the iteration index. We observe that new clusters are gradually less numerous and after a few iterations (here 3 or 4), the number of extracted clusters becomes stable.

5.1.4. Number of added documents as a function of τ

From corpus $\mathcal{BC}_1$, we study the number of selected documents as the enrichment threshold τ varies based on the two

Comparability value between the two aligned clusters : 0.5960093384724747

<DOC>
<DOCID>63a055c59873577c7a5fec38057e698a</DOCID>
<PUBDATE> null </PUBDATE>
<CURDATE>Sat Mar 16 10:27:03 CET 2013</CURDATE>
<FEEDURI>http://www.chinadaily.com.cn/rss/world_rss.xml</FEEDURI>
<ITEMURI>http://www.chinadaily.com.cn/world/2013-03/15/content_16310532.htm</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>Syria warns against rebels sneaking from Lebanon</TITLE>
<DESC>Syrian foreign ministry stressed Thursday that the Syrian army is practicing the "highest levels of self- restraint" not to hit the rebels' positions on the Lebanese side of the borders with Syria.</DESC>
<TXT>Syria warns against rebels sneaking from Lebanon
Updated: 2013-03-15 09:52 Large Medium Small DAMASCUS - Syrian foreign ministry stressed Thursday that the Syrian army is practicing the "highest levels of self- restraint" not to hit the rebels' positions on the Lebanese side of the borders with Syria. "Over the past 36 hours, the armed terrorists in large numbers have sneaked into Syria at the border villages of Mathoume, Ain al- Sharaa, al-Jousieh and Talkalakh," the ministry said in a statement addressing the Lebanese foreign ministry. The Syrian troops have clashed with the infiltrators and the clashes are still ongoing, the ministry said, adding that many of the assailants have been killed while some others have fled back to Lebanon. The ministry noted that the armed groups' positions on the Lebanese side are seeable by the Syrian troops, which "has been practicing the highest levels of self-restraint." Syria hopes "the Lebanese competent authorities will exert efforts to control its borders with Syria," according to the statement. "Syria expects that the Lebanese side will prevent those (rebels) from using the borders as routes, because they are targeting the security of the Syrian people and infringing upon the sovereignty of Syria," the ministry said. The ministry pointed out that the flow of gunmen and arms have notably stepped up since March 12, charging that those infiltrators are receiving support from Lebanon. The Syrian government has for long complained about the flow of arms and cash from surrounding countries, including Lebanon, whose northern city Tripoli is packed with Islamists who oppose the government of Syrian President Bashar al-Assad. Moreover, some leaked phone recordings made public recently by Lebanese TVs have incriminated some Lebanese officials that appeared coordinating the flow of weapons and money to the rebels in Syria. 8.03K</TXT>
</DOC>

<DOC>
<DOCID>1e778be2433287d86fb3d7e35eb22d17</DOCID>
<PUBDATE>Fri Mar 15 10:12:16 CET 2013</PUBDATE>
<CURDATE>Sat Mar 16 10:16:32 CET 2013</CURDATE>
<FEEDURI>http://www.courrierinternational.com/rss/all/rss.xml</FEEDURI>
<ITEMURI>http://www.courrierinternational.com/breve/2013/03/15/damas-menace-de-frapper-le-liban</ITEMURI>
<AUTHOR>[]</AUTHOR>
<TITLE>LIBAN • Damas menace de frapper le Liban</TITLE>
<DESC>"Damas menace 'officiellement' le Liban de bombarder son territoire", titre le quotidien libanais.</DESC>
<TXT>Envoyer
Votre message
"Damas menace 'officiellement' le Liban de bombarder son territoire", titre le quotidien libanais. Selon le ministère des Affaires étrangères syrien, des combats opposeraient l'armée régulière syrienne à des "bandes armées qui tentent d'infiltrer la Syrie à partir du Liban", ce qui pourrait pousser Damas à les frapper jusque sur le territoire libanais. Le Conseil de sécurité des Nations unies a réagi en se disant "très inquiet" des "incidents frontaliers répétés" entre les deux pays : tirs par-dessus la frontière, incursions, trafic d'armes.
<<
</TXT>
</DOC>

Figure 4: Alignment of the two clusters (medoids) having the highest comparability.

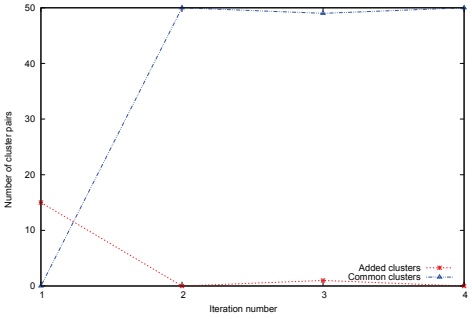

Figure 5: Number of added clusters and number of common clusters when 4 successive iterations of the k-medoids clustering are performed.

proposed enrichment measures S_{v1} (Eq.5) and S_{v2} (Eq.6) (STEP-6).

In Figures 6 and 7, we present the number of added documents according to the enrichment threshold τ (using respectively S_{v1} and S_{v2} measures). Note that the values of τ for the two measures do not represent the same comparability level. For S_{v1}, the value of τ, which significantly reduces the number of added documents is lower than for S_{v2}. Namely, using S_{v1} with $\tau = 0.5$ corresponds almost to the same comparability level than choosing $\tau = 0.7$ with S_{v2}. However, according to our experiments and our intuition, S_{v2} is more suited than S_{v1} to enrich since for a same level of comparability S_{v2} allows for adding more documents than S_{v1}.

5.1.5. Average comparability of each clusters pair without enrichment and with enrichment

In Figures 8 and 9, we show respectively the variation of the average comparability of each cluster pair without enrich-

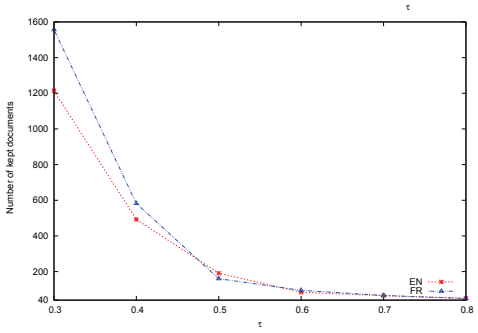

Figure 6: Number of added documents for different enrichment threshold (τ) values when exploiting S_{v1}.

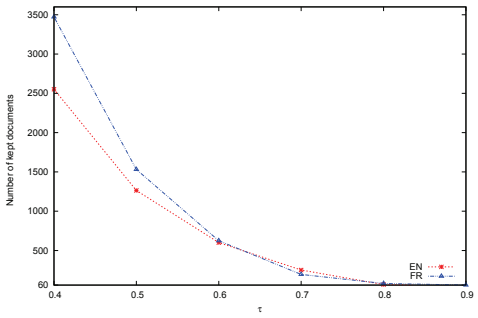

Figure 7: Number of added documents for different enrichment threshold (τ) values when exploiting S_{v2}.

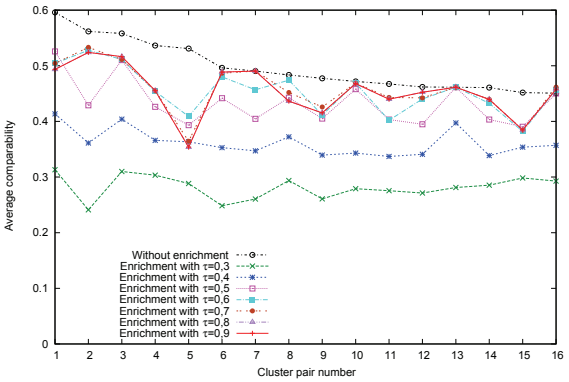

Figure 8: Average comparability of each pairs of cluster pair without enrichment and with enrichment controlled by the adding threshold τ on S_{v1}.

ment and with enrichment controlled by the adding threshold τ on S_{v1} and S_{v2} (STEP-6). In Figure 8, we find that in general the average comparability of each clusters pair increases when τ increases, however, when $\tau \geq 0,7$, the average comparability value is almost stable. This shows that

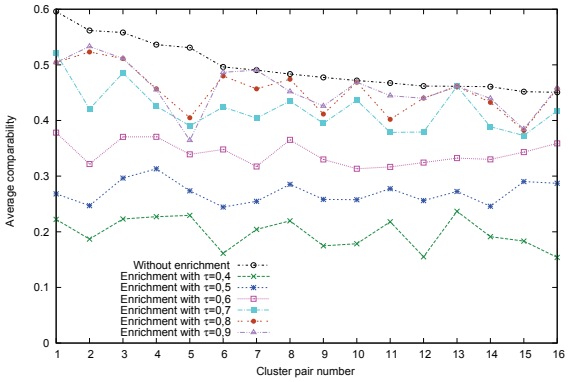

Figure 9: Average comparability of each pairs of clusters without enrichment and with enrichment controlled by the adding threshold τ on S_{v2}.

on S_{v1}, a relevant τ can be initially set around 0.5. In Figure 9, the average comparability of each clusters pair generally increases also when τ increases, but when $\tau \geq 0,8$, the average comparability value becomes very stable. This shows that for S_{v2}, a relevant τ can be initially set around 0.7. Comparing the two figures, we see again that the fusion of *native* and *induced* similarities has a positive impact in the enrichment because when $\tau \geq 0,8$ (even $\tau = 0,9$), more added documents are added. Empirically again, S_{v2} seems to be more interesting than S_{v1}.

6. Conclusion

We have proposed a semi-supervised approach for the construction of comparable corpora having a controlled thematic cohesion. This approach aims at producing aligned thematic clusters more or less comparable by using i) the "thematic" comparability measure C_{VA_2} as defined in (Ke et al., 2014), ii) a k-medoids co-clustering as define in (Ke et al., 2013) with a median value for the parameter α used to merge *native* and *induced* similarities (α=0.5). This approach is based on 6 steps and requires fixing seven important parameters such as the initial number of clusters K_0 for a k-medoids clustering, the reject comparability threshold φ used to prune the the aligned cluster pairs, the enrichment threshold τ used to increase the size of the aligned cluster pairs, etc. We tested our approach on a real case study based on the collecting of published news from RSS news wires during a six month period. We studied some of the effects of these parameters such as the number of extracted clusters, the size of the clusters when varying the enrichment threshold τ, etc. Our approach integrates an enrichment step that allows for providing a comparable corpus of larger size while ensuring a high "thematic" comparability in average between clusters. The big advantage of this approach is that it provides thematic and comparable aligned clusters that serve as basic constituent for the construction of "thematic" comparable corpora.
However, as we have integrated a manual verification step to ensure the quality of the cluster alignment, this approach

is not entirely automatic. Nevertheless, this manual verification only considers the alignment of cluster medoids, and in practice the cost impacting the user is maintained as low as possible and remains acceptable inpractice.

In addition, the parameters that need to be set up or optimized may not be easily tunable because they may vary (hopefully not critically, but this has to be verified) according to different corpus context (themes, sources, genre, etc.). Finally, as the k-medoids co-clustering depends on the initial conditions, we need to execute it a certain number of times to extract sufficient clusters, which also complicates the approach.

Regarding the perspectives, we can define a finer time period to filter documents and alleviate the computation cost of comparability and similarity matrices. It is also relevant to integrate features such as TNC (title and content), LIU (language independent unit) and MTD (monolingual terms distribution) as suggested in (Vu et al., 2009) to complete the "thematic" specification for the corpora construction.

We can also extend this approach to the construction of "thematic" comparable corpora for other pairs of languages and test the "thematic" comparable corpus that is produced against a specific application (such as bilingual lexicon or terminology extraction or cross-language information retrieval). Furthermore, as the similarity-comparability mixing model we use for co-clustering purpose is quite dependent on the coverage of a bilingual dictionary, it is quite appealing to use the constructed "thematic" comparable corpus that is produced to extract bilingual lexicon to enrich the initial bilingual dictionary. Once the bilingual dictionary has been enriched, we can re-use it to refine the "thematic" comparable corpus, and further iterate until no improvement can be further expected.

7. Acknowledgements

This work has been partially funded by the French National Research Agency (ANR-METRICC project).

8. References

Braschler, M. and Scäuble, P. (1998). Multilingual information retrieval based on document alignment techniques. In *Research and Advanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 183–197. Springer.

Déjean, H. and Gaussier, E. (2002). Une nouvelle approche a l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Numéro spécial, corpus alignés:1–22.

Ke, G., Marteau, P.-F., and Ménier, G. (2013). Improving the clustering or categorization of bi-lingual data by means of comparability mapping. Technical report.

Ke, G., Marteau, P.-F., and Ménier, G. (2014). Variations on quantitative comparability measures and their evaluations on synthetic french-english comparable corpora. In *9th Conference on International Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA).

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING*, pages 644–652.

Li, B. (2012). *Measuring and Improving Comparable Corpus Quality*. Ph.D. thesis, Université de Grenoble.

McNamee, P., Mayfield, J., and Nicholas, C. (2009). Translation corpus source and size in bilingual retrieval. In *Proceedings of Human Language Technologies: NAACL-2009, Short Papers*, NAACL-Short '09, pages 25–28, Stroudsburg, PA, USA. ACL.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *ACL*.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pages 265–272.

Otero, P. G. and López, I. G. (2009). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 21–25.

Prochasson, E. E. (2009). *Alignement multilingue en corpus comparables spécialisés*. Ph.D. thesis, Université de Nantes.

Rahimi, Z. and Shakery, A. (2011). Topic based creation of a persian-english comparable corpus. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology*, AIRS'11, pages 458–469. Springer.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 527–534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Schmid, H. (2009). TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th annual international ACM SIGIR conference*, SIGIR '96, pages 58–65, New York, NY, USA. ACM.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Trans. Inf. Syst.*, 25(1), February.

Talvensaari, T. (2008). Effects of aligned corpus quality and size in corpus-based clir. In *Proceedings of the IR research, 30th Eur. conf. on Advances in information retrieval*, ECIR'08, pages 114–125. Springer.

Tao, T. (2005). Mining comparable bilingual text corpora for cross-language information integration. In *In KDD*, pages 691–696.

Vu, T., Aw, A. T., and Zhang, M. (2009). Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the Assoc. for Comp. Ling.*, EACL '09, pages 843–851, Stroudsburg, PA, USA.