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 Sciauble, 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 Lopez, 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" bilateral comparable corpus proposed in (Déjean and Gaussier, 2002), we define a "thematic" bilateral 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 bilateral corpora is thus expressed as follows: two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called "thetically" 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 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 hereafter.

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:

$$C_{V,A_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2}$$

with

- $A_{1|2} = \sum_{w_1 \in WC_1 \cap WD_1} \frac{W(w_1, C_1) \cdot \sigma(w_1)}{\tau(w_1, WD_1)}$
- $A_1 = \sum_{w_1 \in WC_1 \cap WD_1} \frac{W(w_1, C_1)}{\tau(w_1, WD_1)}$
- $A_{2|1} = \sum_{w_2 \in WC_2 \cap WD_2} \frac{W(w_2, C_2) \cdot \sigma(w_2)}{\tau(w_2, WD_2)}$
- $A_2 = \sum_{w_2 \in WC_2 \cap WD_2} \frac{W(w_2, C_2)}{\tau(w_2, WD_2)}$

where $W(w_i, C_i)$ 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 $C_i$ 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 $C_1$ and $C_2$ two collections of documents belonging to two distinct linguistic spaces ($\mathcal{L}_1$ and $\mathcal{L}_2$ respectively) in which two native similarity measures $S_{C_1}$ and $S_{C_2}$ are defined. Let $C(\cdot) : S_{C_1} \times S_{C_2} \to \mathcal{R}$ 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
The "thematic" comparability measure used to calculate comparability between pairs of documents of the initial corpus $BC_0$ is calculated during the previous step, with the complexity $O(|BC_0|^2)$. The complexity to evaluate the comparability matrix is quadratic $(O(|BC_0|^2))$ with the size of the initial corpus.

### 3.2. Assisted construction approach

Starting from a raw bilingual corpora $BC_0$ 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 $BC_0$.

The "thematic" comparability measure $C_{YA_2}$ (Eq.1) is used to calculate comparability between pairs of documents in two different languages. The complexity to evaluate the comparability matrix is quadratic $(O(|BC_0|^2))$ with the size of the initial corpus.

**STEP-2:** Filtering of the initial corpus $BC_0$, and production of a bilingual corpus $BC_1$ 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 $BC_1$ produced at the end of this step is substantially smaller than the size of the initial corpus $BC_0$ ($|BC_1| \ll |BC_0|$).

This step involves sorting and filtering documents from the comparability matrix calculated during the previous step, using a minimum comparability threshold, $\beta$, a threshold $\gamma$ that characterizes the minimum node degree of the comparability bipartite graph obtained after pruning the links associated with a comparability below the threshold $\beta$. A third parameter $\sigma = |BC_1|/2$ 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 $\beta$ and we keep only the lines $i$ for which $nl_i > \gamma$.

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 $\beta$ 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 $w$ ($v_i = Max_i \{F_{ij}\}$ and $w_j = Max_j \{F_{ij}\}$) for a clustering $C$ containing $Nc$ clusters are defined as follows:

e) Finally, the corpus $BC_1$ 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 $BC_1$ 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.8) while varying the number of clusters $k$, we perform the calculation of the average intra and inter clusters similarities ($\delta_{\text{intra}}$ and $\delta_{\text{inter}}$ resp., Eq.9) 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. $\delta_{\text{intra}}$ and $\delta_{\text{inter}}$ for a clustering $C$ containing $Nc$ clusters are defined as follows:

$$
\delta_{\text{intra}}(C_i) = \frac{1}{Nc} \sum_{i=1}^{Nc} \sum_{d,d' \in C_i} S_{C_i}(d,d')
$$

$$
\delta_{\text{inter}}(C_i) = \frac{1}{Nc(Nc-1)} \sum_{i=1}^{Nc} \sum_{j \neq i}^{Nc} S_{C_i}(m_i,m_j)
$$

where $S_{C_i}(d,d')$ and $S_{C_i}(m_i,m_j)$ are similarities provided by the native and induced similarities mixing model.

Figure 1: Sorting of pairs of documents based on the calculation of the matrix $F_{ij} = nl_i + nc_j$ and vectors $v_i = Max_i \{F_{ij}\}$ and $w_j = Max_j \{F_{ij}\}$.
relationship that tends to be a 1-to-1 manner. The comparability threshold \( \varphi \) of the average degree of the graph, when the comparability 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 \( m_l \) belonging to the same linguistic space \( (l) \), we thus have: 

\[
S_{c1}(d, m_i, m_j) = \alpha S_C(d, m_i) + (1 - \alpha) C(d, m_j)
\]

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:

\[
S_{c2}(d, m_i, m_j) = \alpha S_{C_i}(d, m_i) + (1 - \alpha) S_{C_2}(d, m_i)
\]

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

We use a reject threshold, \( \tau \), on \( S_{c1} \) or on \( S_{c2} \), to decide if the document \( d \) will be finally retained to enrich a cluster or not. In practice, each document in the initial corpus \( BC_0 \) is tested and will enrich the corpus if its \( S_{c1} \) or \( S_{c2} \) value is greater than the threshold \( \tau \). Any additional corpus can naturally be used to further enrich the corpus. The threshold \( \tau \) can be adjusted according to the requirements expressed by the user in matter of average comparability and corpus size. If \( \tau \) is low, we will get more documents in each cluster pair, but these documents will be less comparable in average. However, if \( \tau \) 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, \( BC_F \), 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 \( \alpha \) is used in our mixing model to merge native and induced similarities,
2. parameter \( \beta \) determines the minimum comparability value for filtering the raw corpus \( BC_0 \) before extracting co-clusters,
3. parameter \( \gamma \) corresponds to the minimum degree of nodes (documents) in the comparability bipartite graph once a pruning conditioned by the threshold \( \beta \) (minimal comparability value) has been performed,
4. parameter \( \sigma \) determines the number of documents that are kept in the filtered corpus \( BC_1 \) (in which comparability is maximized),
5. the parameter \( K_0 \) specifies the initial number of clusters extractable from the corpus \( BC_1 \).
6. the parameter $\varphi$ is a comparability threshold used for the extraction of the most similar cluster pairs that constitute the corpus $BC_2$ (becoming $BC_3$ after the manual verification step)

7. the parameter $\tau$ is used as a fitness threshold for adding suitable documents to enrich the corpus $BC_3$ in order to produce the final bilingual corpus $BC_F$ 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 $BC_0$ is composed of 39,729 documents (18,168 English documents and 21,561 French documents). For each document, the non stop words arelemmatized 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.bbc.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.ledevenoir.com/rss/...                 | FR       |
| www.lessentielslu/...                      | 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 ($\alpha=0.5$) 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 $\sigma$ (the number of most comparable documents that we initially keep for each language in the corpus $BC_1$) 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 $\beta$ 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 $\gamma$ (the minimum number of comparability links above $\beta$) to 10 to ensure a minimum degree in the initial bi-partite comparability graph. The parameters $\beta$, $\gamma$ and $\sigma$ are defined in STEP-2.

The parameter $K_0$ (the initial clusters number (STEP-3)) and the parameter $\varphi$ (the reject comparability threshold (STEP-4)) will be determined experimentally. Finally, the parameter $\tau$ (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 $BC_1$

5.1.1. Determination of the initial clusters number $K_0$

We determine here an initial value $K_0$ for the co-clustering of $BC_1$ by analyzing the variations of the average intra and inter similarities $\delta_{intra}$ and $\delta_{inter}$ obtained based on a k-medoids clustering (STEP-3) when $k$ varies.

![Figure 2: Variation of the intra et inter clusters similarities $\delta_{intra}$ and $\delta_{inter}$ on k-medoids clustering when $k$ varies (EN left, FR right).](image-url)
In Figure 2 we see that as \( k \) increases, the curves \( \delta_{\text{intra}} \) and \( \delta_{\text{inter}} \) 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 \( \delta_{\text{intra}} \) is large and the value of inter cluster similarity \( \delta_{\text{inter}} \) 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 \( \varphi \)

Here we determine the comparability threshold \( \varphi \), 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 \( \varphi \) varies (STEP-4).

![Figure 3: Determination of the comparability threshold \( \varphi \) 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 \( \varphi \) 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 \( \varphi \) 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 \( \varphi \) 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 been one of the highest comparability value. We present in Figure 4 as example one of the 16 aligned cluster pairs (each cluster is represented by its medoid).

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

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 decrease 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 \( \tau \)

From corpus BC1, we study the number of selected documents as the enrichment threshold \( \tau \) varies based on the two 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 \( \tau \) (using respectively \( S_{v1} \) and \( S_{v2} \) measures). Note that the values of \( \tau \) for the two measures do not represent the same comparability level. For \( S_{v1} \), the value of \( \tau \), 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-
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 $C_{V,A_2}$ as defined in (Ke et al., 2014), ii) a $k$-medoids co-clustering as defined in (Ke et al., 2013) with a median value for the parameter $\alpha$ used to merge native and induced similarities ($\alpha=0.5$). 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 $\varphi$ used to prune the aligned cluster pairs, the enrichment threshold $\tau$ 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 $\tau$, 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 veri-
cation only considers the alignment of cluster medoids, and
in practice the cost impacting the user is maintained as low
as possible and remains acceptable in practice.
In addition, the parameters that need to be set up or op-
timized may not be easily tunable because they may vary
(hopefully not critically, but this has to be verified) ac-
2

gording 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 num-
ter of times to extract sufficient clusters, which also com-
plicates the approach.
Regarding the perspectives, we can define a finer time pe-
riod 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 re-
trieval). Furthermore, as the similarity-comparability mix-
ing model we use for co-clustering purpose is quite depen-
dent on the coverage of a bilingual dictionary, it is quite ap-
pealing to use the constructed "thematic" comparable cor-
pus that is produced to extract bilingual lexicon to enrich
the initial bilingual dictionary. Once the bilingual diction-
ary has been enriched, we can re-use it to refine the "thet-
matic" comparable corpus, and further iterate until no im-
provement 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 in-
formation retrieval based on document alignment tech-
niques. In Research and Advanced Technology for Dig-
tal Libraries, volume 1513 of Lecture Notes in Computer
Science, pages 183–197. Springer.
Déjean, H. and Gaussier, E. (2002). Une nouvelle ap-
proche a l’extraction de lexiques bilingues à partir de cor-
pus comparables. Lexicometrika, Numéro spécial, cor-
pus alignés:1–22.
Ke, G., Marteau, P.-F., and Ménier, G. (2013). Improv-
ing 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 evalu-
ations on synthetic french-english comparable corpora.
In 9th Conference on International Language Resources
and Evaluation (LREC 2014). European Language Re-
sources Association (ELRA).
Li, B. and Gaussier, E. (2010). Improving corpus compa-
rability for bilingual lexicon extraction from comparable
corpora. In COLING, pages 644–652.
Li, B. (2012). Measuring and Improving Comparable Cor-
pus Quality. Ph.D. thesis, Université de Grenoble.
McNamee, P., Mayfield, J., and Nicholas, C. (2009).
Translation corpus source and size in bilingual re-
trieval. In Proceedings of Human Language Technolo-
gies: 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 mul-
tilingual 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 cor-
pus 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 Tech-
ology, AIRS’11, pages 458–469. Springer.
Resnik, P. (1999). Mining the web for bilingual text. In
Proceedings of the 37th annual meeting of the Associ-
ation 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 Interna-
tional Conference on New Methods in Language Pro-
cessing, 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 sys-
tem. 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 com-
parable 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 re-
trieval, 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 cor-
pora. In Proceedings of the 12th Conference of the Euro-
pæan Chapter of the Assoc. for Comp. Ling., EACL ’09,
pages 843–851, Stroudsburg, PA, USA.