

# Advanced Text Mining Methods for Bilingual Lexicon Extraction from Specialized Comparable corpora

No Author Given

No Institute Given

**Abstract.** Recent works rely on comparable corpora to extract efficient bilingual lexicon. Most of approaches in the literature for bilingual lexicon extraction are based on context vectors (CV). These approaches suffer from noisy vectors that affect their accuracy. This paper presents new approaches which relies on some advanced text mining methods to extract association rules between terms (AR) and extend them to *contextual meta-rules* (MR). In this respect, we propose to extract bilingual lexicons by deploying standard context vectors, association rules and contextual meta-rules. These proposed approaches utilize correlations between co-occurrence patterns across language. An experimental validation conducted on a specialized comparable corpora, highlights a significant improvement of bilingual lexicon based on MR compared to the standard approach.

## 1 Introduction

Over several decades, a lot of effort has been put into creation of lexicons with high coverage, high translation quality and for specific domain. At a later time, the researches were oriented towards the exploitation of *comparable corpora*. With comparable corpora are sets of text collections which cover roughly the same subject area in different languages, but which are not translations of each other. Most of works in BLE task from comparable corpora represent the *standard approach* [1, 2, 3, 4, 5] which is based on the context vectors (CV). These vectors store a set of words which are for the neighbourhood of the base word and share the same lexical context. The relation between a base word and its context is called a co-occurrence relation.

Moreover, in text mining (TM) field, one of the main techniques generating knowledge based on co-occurrence relation is association rule (AR) extraction introduced in [6]. These rules provide information on the inter-terms correlations.

For each domain, common language pairs and commercially important subject areas such as medicine, specific dictionaries would be developed. Thus, we encode our intuition into new approaches for extracting bilingual lexicons from specialized comparable corpora. Our proposed approaches relies on enriched representations of the word, especially those derived through Formal Concept Analysis (FCA) paradigm [7] and some advanced TM methods. These methods are deployed to extract AR and extend them to *contextual meta-rule* (MR). These later capture all the words related to AR associated with the base word. This leads to a less sparse representation. Therefore, we focus on how to compute similarities between AR and between MR using their specific metrics, namely support and confidence. Finally, we compare extracted lexicons using these two patterns with regard to that of the standard approach.

The article is organized as follows: in the Section 2, we present the standard approach, their improvements and extended approaches to the task of BLE. Then, we describe, in Section 3, our different models for the BLE based on CV, AR and MR. In Section 4, we describe our combination strategies applicated for extracted lexicons. Section 5 will be dedicated to the linguistic resources and evaluation results of extracted lexicons. The Conclusion section wraps up the article and outlines future works.

## 2 Related Works to the BLE from Comparable Corpora

Most of the works of the state-of-the-art dealing with BLE task from comparable corpora are based on the standard approach [1, 2, 3, 4, 5]. The approaches can be classified in three families, as follows:

1. Standard Approach (SA): The main assumption of the SA states that words with a similar meaning are likely to appear in similar context across languages. Therefore, a word can be represented as a context vector (CV) in source or target language. The dimensions of the source and target vectors are totally different, because each of them is represented by words in a source language and words in a target language. In order to enable the comparison of source and target CV, words in the source CV are translated into the target language using an initial bilingual dictionary. The most popular measure for comparison is the cosine measure, but other authors studied other measures. Thus, we have a list of candidate translations for the base word ranked according to their similarity scores.
2. Improvements of SA: Several contributions have been proposed to improve each step of the SA. [8] combine the information provided by translations context with transliterations<sup>1</sup> and scientific compound words in target language. [9] suggest that CV should be based on the most important contextually relevant words (in-domain terms), and thus propose a method for filtering the noise of the CV. Some researchers looked into adding additional linguistics resources by combining a general dictionary with a specialized dictionary [10] or a multilingual thesaurus [11]. [12] present two techniques for filtering the entries of the initial dictionary (POS-tagging criteria and relative frequency ratio criteria). [4] introduce a word sense disambiguation process that identifies the translations of polysemous words that are more likely to give the best representation of CV in the target language.
3. Extensions of SA: Other approaches have been proposed for BLE that diverge from the SA. In [11], the interlanguage similarity approach avoids the direct translation of the elements of the CV. The principle is to associate to each base word the closest CV to terms in the initial dictionary by using a similarity measure. Other works have used geometric approaches based on two spaces of vectors. The translation step consists of transferring the vectors from the space of the source terms to space of the target terms [13]. We also distinguish the syntactic approach, which is based on the observation that a word and its translation tend to share the same syntactic dependency relations [14]. [15] combine the contextual representation within

<sup>1</sup> transliterations are words adapted from a source language to a target language, based on their pronunciation

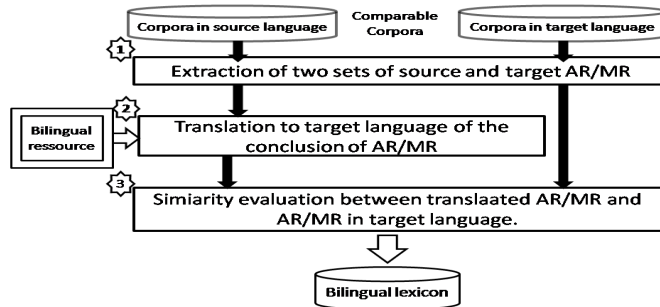
a thematic one. The assumption is that a term and its translation share thematic similarities. Other work combine two representations of the context, namely a contextual representation (by bag of words) and a syntactic representation (by syntactic dependency relations) [16, 17]. [5] introduce an intermediary step that consists of re-estimating the observed word co-occurrence counts either by smoothing or by prediction techniques. Many single terms are compositional (composed of roots and affixes) and this information can be very useful to match translational pairs, especially for infrequent terms where distributional methods often fail. [18] present the compositional approach based on a proposed bilingual morpheme extraction methods.

Our contribution introduced in this paper, can be seen as an extension of the SA aims at using two new patterns except of context vectors. Our approaches differ from other works in two ways: (1) The new patterns are used, according to our knowledge, for the first time for the task of BLE from comparable corpora. These patterns are based on FCA and some advanced TM methods known as association rules extraction. Association rules are also extended to a mined new pattern named contextual meta-rules which capture all the words related to AR associated with the base word. Comparing to CV, MR leads to a less sparse representation. (2) As only one word representing a rule is usually present in a given context, CV fail to integrate all the words related to a given set of association rules. Therefore, it is easier to adapt our approaches to other language pairs and without any condition about the size of the corpora. In addition, we compare the SA using CV to BLE based on AR, BLE based on MR and their combinations.

### 3 Proposed Approaches

Our goal is to extract bilingual lexicons using two patterns AR and MR, providing additional and implicit knowledge. The Figure 1 depicts our proposed approaches: (1) BLE based on AR, and (2) BLE based on MR, it consists in three major steps.

Fig. 1. Overview of our two approaches



### 3.1 Formalisation and Definitions

After introducing some notations, we state the definitions of the concepts used in the reminder of the paper. In this respect, Table 1 provides an overview of the notations used in this and later sections.

**Table 1.** Summary of notations

	Description		Description
$\mathcal{R}_S$	Set of AR in source language	$\mathcal{R}_T$	Set of AR in target language
$\mathcal{MR}_S$	Set of MR in source language	$\mathcal{MR}_T$	Set of MR in target language
$AR_S$	AR in source language ( $AR_S \in \mathcal{R}_S$ )	$AR_T$	AR in target language ( $AR_T \in \mathcal{R}_T$ )
$MR_S$	MR in source language ( $MR_S \in \mathcal{MR}_S$ )	$MR_T$	MR in target language ( $MR_T \in \mathcal{MR}_T$ )
$P_S$	The premise part of $AR_S$ or $MR_S$	$P_T$	The premise part of $AR_T$ or $MR_T$
$C_S$	The conclusion part of $AR_S$ or $MR_S$	$C_T$	The conclusion part of $AR_T$ or $MR_T$
$w_S$	A word in source language ( $w_S \in C_S$ )	$w_S^j$	A candidate translation of $w_S$

In this paper, we use in TM field, the theoretical framework of FCA presented in [7]. First, we formalize an extraction context made up of documents and index terms, called *textual context*.

**Definition 1 (Textual context).** A textual context is a triplet  $\mathcal{TC} = (\mathcal{D}, \mathcal{V}, \mathcal{I})$  such as:

- $\mathcal{D} = \{d_1, d_2, \dots, d_p\}$  is a finite set of documents of the corpus.
- $\mathcal{V} = \{w_1, w_2, \dots, w_q\}$  is a finite set of  $q$  distinct words of the corpus (i.e., vocabulary of the corpus).
- $\mathcal{I}$  is a binary relation, i.e.,  $\mathcal{I} \subset \mathcal{D} \times \mathcal{V}$ , which connects every document with the words of the corpus which are associated with it.

**Definition 2 (Association rule between terms).** An association rule (AR) binds two termsets<sup>2</sup>, which respectively constitute its premise ( $T_1$ ) and conclusion ( $T_2$ ) parts [6]. Thus, an AR estimates the probability of having the terms of the conclusion ( $T_2$ ) in a document, given that those of the premise ( $T_1$ ) are already there.

The advantage of the insight gained through AR is in the contextual nature of the discovered inter-term correlations. Indeed, more than a simple assessment of pair-wise term occurrences, an AR binds two sets of terms, which respectively constitute its premise and conclusion parts.

Given a rule  $AR: T_1 \rightarrow T_2$ , the *support* and *confidence* of AR are computed as follows:

$$Supp(AR) = Supp(T_1 \cup T_2) \quad (1) \qquad Conf(AR) = \frac{Supp(T_1 \cup T_2)}{Supp(T_1)} \quad (2)$$

**Definition 3 (Contextual meta-rule).** A contextual meta-rule, denoted by MR, is an implication of the form:  $MR: p \Rightarrow w_1, w_2, \dots, w_k$ , such as,  $p \in V$  is the premise and  $\{w_1, \dots, w_k\} \subset V$  is the conclusion.

<sup>2</sup> By analogy to the itemset terminology used in data mining for a set of items.

The support of a  $MR$ , denoted by  $Supp(MR)$ , is the number of documents  $d \in \mathcal{D}$  containing all words  $\in \mathcal{V}$  of  $MR$ . The support is the minimum of support values of  $AR$  selected to construct the  $MR$ . The confidence of a  $MR$ , denoted by  $Conf(MR)$ , is the minimum of confidence values of  $AR$  constituting the  $MR$ . The support and confidence are formally defined as follows:

$$Supp(MR) = \min_{(AR_i \subset MR)} Supp(AR_i) \quad Conf(MR) = \min_{(AR_i \subset MR)} Conf(AR_i)$$

### 3.2 Extraction of Association Rules and Meta-Rules

In order to extract the most representative terms, a linguistic preprocessing is required on the document collections. The textual context document-terms  $\mathcal{TC}$ , is then built by retaining only common nouns, proper nouns and verbs as well the vocabulary  $\mathcal{V}$ . The rationale for this focus is that nouns are the most informative grammatical categories and are most likely to represent the content of documents [19]. A stoplist is used to discard functional terms that are very common.

In order to extract AR, we adapted the CHARM-L algorithm [20] to consider any given  $\mathcal{TC}$ . The algorithm extracts all the frequent termsets as described in [20], with respect to minimal and maximal support thresholds  $minsupp$  and  $maxsupp$ <sup>3</sup>. These thresholds are experimentally set by considering the *Zipf* distribution of each collection. The construction of a meta-rule, *i.e.*, MR consists of grouping the AR having a common label which is presented by the premise. A meta-rule helps to explain finer relations between terms. Indeed, MR provides a global context for the terms that appear together. New support and confidence values ( $Supp(MR)$  and  $Conf(MR)$ ) are defined in section 3.1.

### 3.3 Translation and Disambiguation of association and meta-rules

In SA, dimensions of the source and target vectors are different from each other and source CV have to be translated in target languages using an initial dictionary in order that the dimensions agree [21]. The core of our approach, as the SA, is the initial dictionary, it allows the translation of AR/MR of a candidate word and compare it to all the target AR/MR to identify the correct translation according to a similarity measure.

For each  $AR_S$  or  $MR_S$ , for any source word  $w_{S_i} \in C_S$ , we propose to associate a weight  $f(P_S, w_{S_i})$  to assess the relationship of  $w_{S_i}$  with the premise  $P_S$ . This relation is already given by the confidence value (conditional probability) offered by the AR:  $P_S \rightarrow w_{S_i}$ . We propose to exploit this relation as a local context of each word  $w_{S_i} \in C_S$  with  $P_S$ . This generates the construction of the weight vector for each AR or MR in source and target language. A weight vector  $\vec{V}_R$  for an AR or MR of the form  $P_S \rightarrow w_{S_1}, \dots, w_{S_k}$  is defined as follows:  
 $\vec{V}_R = (f(P_S, w_{S_1}), \dots, f(P_S, w_{S_k}))$ .

<sup>3</sup>  $maxsupp$  means that the termset must occur at most this user-defined threshold.

Let us note that each source word  $w_{S_i} \in C_S$  is translated to the target language using an initial dictionary. We consider all the translations proposed for a source word by attributing for each the same score  $f$  as that of the source word. The word which will have no translation will not be added to the translated AR or MR.

### 3.4 Similarity Evaluation

Each translated AR or MR is compared to the sets  $\mathcal{R}_T$  or  $\mathcal{MR}_T$  to filter the most similar ones by using similarity measures. We adopt two methods for computing similarity between AR and between MR respectively:

1. **Context-based similarity (CBS):** This method relies on vector representations of word meaning. In SA, the word meanings are represented as vectors into a high dimensional space. In our case, conclusion part of AR or MR are represented as a vector with low dimension. We consider the most commonly used measure which is the cosine [1, 5] of the angle formed by two source and target vector :  $Cos(\vec{V}_S, \vec{V}_T) = \frac{\vec{V}_S \cdot \vec{V}_T}{|\vec{V}_S| \cdot |\vec{V}_T|}$ .

The similarity between AR or MR is defined as follows:

$$Sim_{context}(AR_S, AR_T) = Cos(\vec{V}_{AR_S}, \vec{V}_{AR_T}) \quad (5)$$

$$Sim_{context}(MR_S, MR_T) = \frac{Cos(\vec{V}_{MR_S}, \vec{V}_{MR_T})}{Conf(MR_T)} \quad (6)$$

$Cos(\vec{V}_{AR_S}, \vec{V}_{AR_T})$  (respectively  $Cos(\vec{V}_{MR_S}, \vec{V}_{MR_T})$ ) are the cosine between the weight vectors of  $AR_S$  and  $AR_T$  (respectively  $MR_S$  and  $MR_T$ ).

$Conf(MR_T)$  is the confidence value of the target meta-rule  $MR_T$ , which leads to show the importance of a  $MR_T$  based on its confidence value with a  $MR_S$ . We exploit then the *global context* of a MR, namely the support and the confidence, associated for each  $MR_T$ .

2. **Semantic-based similarity (SBS):** Semantic measures can be reliable because they are based on the judgments of human experts [22]. The semantic similarity between words is assessed based on dictionaries or thesauri. Among the majority of existing semantic similarity measures, WordNet lexical database [23] is used as the underlying resource to calculate semantic similarity. We distinguish six measures using WordNet classified in two categories, namely: (1) Measures based on information content (IC) denoted *RESN* [24], *JCN* [25], and *LIN*. (2) Measures based on path length which simply count the distance between two words in the WordNet taxonomy denoted *PATH* [26], *WUP* [27] and *LCH* [28].

The intuition is to set for each source word  $w_S$ , having several translations, the interlingual similarity  $Sim_{inter}(w_S, C_T)$  of the most similar translation to the words of the target conclusion part  $C_T$ . We compute the semantic similarity  $Sim_{sem}$  between two AR or MR by taking the maximum of  $Sim_{inter}$  of each source word  $w_S \in C_S$ . The final formula for computing  $Sim_{sem}$  between two AR or MR is defined as follows:

$$Sim_{sem}(AR_S, AR_T) = \max_{w_S \in C_S} Sim_{inter}(w_S, C_T) \quad (7)$$

$$Sim_{sem}(MR_S, MR_T) = \frac{[\max_{w_S \in C_S} Sim_{inter}(w_S, C_T)]}{Conf(MR_T)} \quad (8)$$

The advantage of similarity computing between meta-rules is the exploitation of new values of support and confidence. The semantic similarity is weighted by a score based on the confidence score of the  $MR_T$ .

We give an example of AR and MR in source and target language:  $AR_S$ : sein  $\rightarrow$  traitement (1048 ; 0.404946) , . . . , sein  $\rightarrow$  cancer (1747 ; 0.675039).

$AR_T$ : breast  $\rightarrow$  cancer (39276 ; 0.802353), breast  $\rightarrow$  study (12922 ; 0.263978) , breast  $\rightarrow$  treatment (11339 ; 0.23164).

$MR_S$ : sein  $\Rightarrow$  cas , traitement , risque , cancer , tude , taux , (613 ; 0.236862).

$MR_T$ : breast  $\Rightarrow$  cancer , treatment , study , (11339 ; 0.23164).

The bilingual dictionary is applied on each source term of the conclusion part of  $MR_S$ . The translated MR becomes: *sein*  $\Rightarrow$  *case instance* , *treatment treating therapy* , *riskiness risk* , *cancer* , *rates rate* . We follow the two similarity calculation strategies between each translated MR and  $MR_T$ . For each source premise '*sein*', a list of top-ranked translation candidates are selected according to their highest similarities as follows: *sein* : (*breast*, 2.3245), (*aromatase*, 2.1607), (*chemotherapy*, 2.0258), (*study*, 2.0060), (*recurrence*, 1.9804), (*age*, 1.9695) , (*disease*, 1.9591) , etc.

To build the lexicon for each approach, we associate for each entry in the source language from  $AR_S$  (or from  $MR_S$ ) a number of top-ranked candidate translations in the target language representing the premises parts of  $AR_T$  (or  $MR_T$ ) and illustrate as follows:

*sein* breast, aromatase, chemotherapy, study, recurrence, age, disease, etc.

## 4 Experimental Validation

### 4.1 Corpora and linguistic resources

We evaluate our BLE approaches on a specialized comparable corpora. The Breast Cancer corpus is composed of documents collected from the Elsevier website<sup>4</sup>. The documents were taken from the medical domain within the sub-domain of "breast cancer". The same corpus was used in [5]. The documents have been selected and published between 2001 and 2008 where the title or the keywords contain the term *cancer du sein* in French and *breast cancer* in English. The collected documents counts 130 French documents (about 530,000 words) and 1,640 English documents (about 7.4 million words).

The translation is handled using the linguistic resource BabelNet<sup>5</sup> [29] due to its lexicographic and encyclopedic coverage of multilingual terms resulting from a mapping of the Wikipedia pages<sup>6</sup> and WordNet, with other lexical semantic resources.

<sup>4</sup> www.elsevier.com

<sup>5</sup> babelnet.org

<sup>6</sup> wikipedia.org

To evaluate the quality of our approaches regards to SA, we built a bilingual reference list. It should be noted that the evaluation of terminology extraction using specialized comparable corpora often relies on lists of a small size: 100 in [3], 125 and 79 in [4]. For Breast Cancer, we selected randomly 170 French words, corresponding to source premises of MR, with their translations in English from the online dictionary *Word Reference*<sup>7</sup>.

## 4.2 Evaluation

The rank of the correct translation can be considered as an important characteristic when evaluating extracted lexicons. This characteristic is taken into account only by measuring the mean average precision (MAP) defined in [5]. Let  $n$  is the number of terms of the reference list,  $N$  is the length of candidates translation and  $r_i$  is the rank of the correct candidate translation  $i$ . If the correct translation does not appear in the top  $N$  candidates,  $\frac{1}{r_i}$  is set to 0. The MAP is defined as follows:

$$MAP = \frac{1}{n} \sum_{i=1}^N \frac{1}{r_i} \quad (9)$$

**Fig. 2.** Precision of  $SA_{CV}$ ,  $L_{MR}$  and  $L_{AR}$  for Breast Cancer (FR-EN)

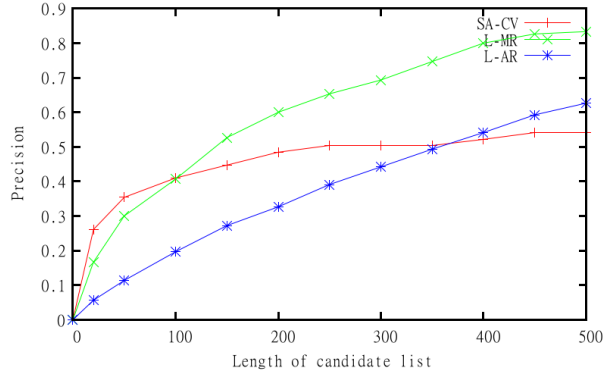


Figure 2 shows the different precision values obtained for the standard approach ( $SA_{CV}$ ), BLE based on Meta-rules ( $L_{MR}$ ) and BLE based on association rules ( $L_{AR}$ ). We vary the length of candidate list  $N$  from 20 to 500. The low scores obtained for the top 20 candidates are explained that unlike previous work, we do not limit ourselves to very frequent words for evaluation. This is ensured by the step of association rules extraction by setting minimum thresholds of support and confidence.  $L_{MR}$  significantly outperforms  $L_{AR}$  and  $SA_{CV}$  approaches. Indeed, the overall precision of the  $L_{MR}$  and  $L_{AR}$  approaches increases for large lengths of candidates list as  $N =$

<sup>7</sup> wordreference.com



200, 300, 400, and 500. We can see in Figure 2 that the performance increases by increasing  $N$ . This can be explained by the fact that the probability of obtaining correct translations increases with the candidate list growth. While after  $N = 150$ , performance remains almost constant. In the experiments below, we set  $N = 200$ . We note that  $Top_N$  means that the correct translations of a given word is present in the first  $N$  candidates.

### 4.3 Results and Discussion

**Table 2.** MAP at  $Top_{200}$  with different semantic-based similarity measures

	JCN	LIN	RESN	PATH	WUP	LCH
$L_{MR}$	0.3503	0.3988	0.3555	0.3895	<b>0.4039</b>	0.3535
$L_{AR}$	0.2081	0.3836	0.1975	0.2531	<b>0.3885</b>	0.2222

The experiments for similarity evaluation adopt two methods for computing similarities as presented in section 3.4. The semantic-based similarity method is performed with respect to the six semantic similarity measures described above. Table 2 displays the obtained results measured in terms of MAP at  $Top_{200}$  for scenarios  $L_{MR}$  and  $L_{AR}$  by varying semantic similarity measures. From these results, we notice that the overall MAP is improved with WUP measure regards to other measures. This increase is important and shows that the similarity between a base word and its candidate translations relies on more information. This information is valuable with semantic measures based on a lexical database.

**Table 3.** MAP improvement achieved with all scenarios for Breast Cancer. The symbols  $\dagger$  indicates statistically significant improvement over the best run in bold,  $p - value < 0.05$

	$SACV$ (Baseline)	$L_{MR}$	$L_{AR}$	$LCV+MR$	$LCV+AR$	$LCV+MR+AR$
CBS	0.383	<b>0.4576</b> $\dagger$	0.4061	<b>0.4768</b> $\dagger$	0.4309	<b>0.5101</b> $\dagger$
SBS	-	0.4039	0.3985	0.4527	0.4054	0.4866

Table 3 shows obtained results in terms of MAP. We interpret that our approach based on meta-rules  $L_{MR}$  were improve significantly for the three corpora compared to  $SACV$ . This can be explained by the noisy nature of the context vectors with respect to the filtered meta-rules by setting thresholds of support and confidence for their construction. The  $L_{MR}$  is better in terms of MAP compared to  $L_{AR}$ , and this can be explained by the fact that the rank of the correct translations found for the lexicon based on MR is more important than the correct translations found for the lexicon based on AR. Moreover, the extracted lexicon have a significant improvement with MR deploying Context-based Similarity (CBS) than those deploying Semantic-based Similarity

(SBS) methods. We demonstrate that approaches which combines CV with MR and AR ( $L_{CV+MR}$  and  $L_{CV+AR}$ ) were improved significantly more than  $L_{MR}$  and  $L_{AR}$ . Besides, approaches which integrates MR, AR and CV ( $L_{CV+MR+AR}$ ) gives the highly significant difference ( $p < 0.01$ ) for Breast Cancer (with a  $p$  equal to 0.0059. This can be explained by the fact that the vocabulary used in the breast cancer field is specific and less ambiguous. The obtained results for  $L_{MR}$  in term of MAP (45.19%) are better than comparable results reported in [5] (42.3% which is the best MAP assessed for the unbalanced version of the corpus) and those reported in [30] (42.4% for the weighted combination).

## 5 Conclusion and perspectives

We introduced new approaches for bilingual lexicon extraction considered as an extension of the SA aims at using two new patterns except of context vectors. These patterns are represented by association rules and meta-rules which are extracted based on FCA and some advanced TM methods. We evaluated our approaches on a specialized comparable corpora from the medical domain. More precisely, our different experiments show that using a specialized comparable corpus always improves significantly the quality of extracted lexicons in term of MAP. Moreover, similarity evaluation between meta-rules deploying Context-based Similarity gives the best MAP than those deploying Semantic-based Similarity measures. The results showed that the lexicon based on MR provide a solution for noisy problem, bearing in mind that the contexts are generally limited to few words around the base word, encountered with context vectors. Furthermore, the quality of extracted lexicons combining context vectors, association rules and meta-rules is highly significant. As future work, we first plan to improve the quality of the extracted lexicons. Secondly, we propose a CLIR model based on the extracted lexicons.

## References

1. Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: Machine Translation and the Information Soup, third Conference of the Association for Machine Translation. Lecture Notes in Computer Science, Springer (October 1998) 1–17
2. Rapp, R.: Automatic identification of word translations from unrelated english and german corpora. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. (1999) 519–526
3. Morin, E., Daille, B., Takeuchi, K., Kageura, K.: Bilingual terminology mining-using brain, not brawn comparable corpora. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07), Prague, Czech Republic. (2007) 664–671
4. Bouamor, D., Semmar, N., Zweigenbaum, P.: Towards a generic approach for bilingual lexicon extraction from comparable corpora. In: Proceedings of the 15th Machine Translation Summit, Nice (September 26 2013) 143–150
5. Morin, E., Hazem, A.: Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. Natural Language Engineering **22**(4) (2016) 575–601

6. Agrawal, R., Skirant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20<sup>th</sup> International Conference on Very Large Databases, VLDB 1994, Santiago, Chile (September 1994) 478–499
7. Ganter, B., Wille, R.: Formal Concept Analysis. Springer-Verlag (1999)
8. Prochasson, E., Morin, E.: Points d’ancrage pour l’extraction lexicale bilingue à partir de petits corpus comparables spécialisés. In: Traitement Automatique des Langues (TAL). Volume 50. (2009) 283–304
9. Ismail, A., Manandhar, S.: Bilingual lexicon extraction from comparable corpora using in-domain terms. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING10), Beijing, China (2010) pages 481489
10. Morin, E., Prochasson, E.: Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In for Computational Linguistics, A., ed.: Proceedings of the 4th Workshop on Building and Using Comparable Corpora, Portland, Oregon (Juin 2011) 27–34
11. Déjean, H., Gaussier, E.: Une nouvelle approche à l’ide lexicques bilingues à partir de corpus comparables. In: VÉRONIS, J., directeur de la publication : Lexicométrica, Alignement lexical dans les corpus multilingues. (2002) 1–22
12. Hazem, A., Morin, E.: Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey, European Language Resources Association (ELRA) (may 2012)
13. Gaussier, E., Renders, J.M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL’04). (2004) 526–533
14. Yu, K., Tsujii, J.: Bilingual dictionary extraction from wikipedia. In: Proceedings of NAACL HLT 2009. (2009) 121–124
15. Rubino, R., Linarès, G.: Une approche multi-vue pour l’extraction terminologique bilingue. In: CORIA’11. (2011) 97–111
16. Andrade, D., Matsuzaki, T., Tsujii, J. In: Effective Use of Dependency Structure for Bilingual Lexicon Creation. Springer Berlin Heidelberg (February 20–26 2011) 80–92
17. Hazem, A., Morin, E.: Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. In: Computational Linguistics and Intelligent Text Processing - 15th International Conference (CICLing 2014). Volume 8404 of Lecture Notes in Computer Science., Springer (April 6–12 2014) 310–323
18. Hazem, A., Daille, B.: Bilingual lexicon extraction at the morpheme level using distributional analysis. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (may 2016)
19. Barker, K., Cornacchia, N.: Using noun phrase heads to extract document keyphrases. In: 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, London, UK, Springer-Verlag (2000) 40–52
20. Zaki, M.J., Hsiao, C.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. on Knowl. and Data Eng. **17**(4) (april 2005) 462–478
21. Kim, J., Kwon, H., Seo, H.: Evaluating a pivot-based approach for bilingual lexicon extraction. Comp. Int. and Neurosc. **2015** (2015) 434153:1–434153:13
22. Lintean, M.C., Rus, V.: Measuring semantic similarity in short texts through greedy pairin and word semantics. In: Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference. (May 2012)
23. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11) (1995) 39–41

24. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. (1995) 448–453
25. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR (1997)
26. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics. (1989) 17–30
27. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics. ACL '94, Association for Computational Linguistics (1994) 133–138
28. Leacock, C., Chodorow, M.: chapter Combining local context and WordNet sense similarity for word sense identification. In: WordNet: an electronic lexical database. MIT Press (1998)
29. Navigli, R., Ponzetto, S.: Babelnet: building a very large multilingual semantic network. In Hajic, J., Carberry, S., Clark, S., eds.: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10), The Association for Computer Linguistics (2010) 216–225
30. Chebel, M., Latiri, C., Gaussier, E.: Bilingual lexicon extraction from comparable corpora based on closed concepts mining. In: Advances in Knowledge Discovery and Data Mining - 21st Pacific-Asia Conference (PAKDD 2017). (23-26 May 2017) 586–598