Mining monolingual and bilingual corpora

Chiraz Latiri, Kamel Smaïli, Caroline Lavecchia, David Langlois

Mining Monolingual and Bilingual Corpora

Chiraz Latiri[1], Kamel Smaïli [2,3], Caroline Lavecchia[2,3] and David Langlois[2,4]
[1]URPAH Team, Computer Sciences Department
Faculty of Sciences of Tunis
El Manar University, Tunisia
[2]LORIA, Speech Group, Vandoeuvre, France
[3]University Nancy 2
[4] IUFM de Lorraine
chiraz.latiri@gnet.tn
{smaili,lavecchia,langlois}@loria.fr

Abstract—In this paper, we describe two new methods of mining corpora. The first one is based on association rules inspired from classical algorithms of datamining and the second one motivated by the results of triggers on statistical language modeling. Association rules are tested on French newspaper and on a set of scientific documents to expand query. The proposed method outperforms the baseline model. The association rules and triggers are then generalized to mine bilingual corpora. We tested before classical mining algorithms which fail in this task due to the complexity of machine translation corpora which are too huge in comparison to those used classically on textmining.
Association rules have been extended to retrieve inter-lingual association from a bilingual corpus. The triggers are generalized also in order to build a translation table. Both methods have been integrated in a real statistical machine translation. We show the feasibility of two approaches in the context of machine translation mining. By several experiments, we show that inter-lingual triggers achieve better results than the third IBM model.

Index Terms—Formal Concept Analysis, Galois closure operator, Association rule, Generic basis, Triggers, Inter-triggers, Statistical machine translation, BLEU score.

1 INTRODUCTION

Text corpora became since few decades a material raw for several applications such as textmining, speech recognition, machine translation, information retrieval, ontology design, etc. Several techniques have been proposed by different research communities in order to deal with these purposes. Some of them are based on external knowledge provided by an expert while others are based on formal natural language processing. Another way consists in the use of statistical methods and more especially Hidden Markov Models as in speech recognition. In this paper, we are interested in two applications, namely: information retrieval and machine translation.

For information retrieval (IR), the track followed here is based on association rules founded on Galois closure operator [1]. After a remind of the basic concepts necessary to obtain association rules, we will address the issue of expansion query by using the IR system SMART (System for the Mechanical Analysis and Retrieval of Text) [2]. Tests will be conducted on two corpora, namely OFIL and INIST [3]. The second application concerns statistical machine translation. Because the corpus dedicated to translation contains more than 0.5 million words, none of the classical algorithms [4], [5], [6], [7] and [8] succeed to produce closed frequent termsets and their minimal generators. That is why we propose an original way to use association rules in machine translation. Inspired from the work of [9], we adapt the association rules to make them working on parallel corpus, we then propose a new concept called inter-lingual association rule which will be used to produce a translation table. This one is integrated on an operational machine translation. To the best of our knowledge, the association rules have never been used in the context of statistical machine translation.
The second method presented in this paper concerns also an original way to produce a translation table. This method does not need any alignement. It is based on the concept of trigger [10] which we extend it to take into account inter-lingual triggers. Several experiments are conducted in order to find out the best translation table, the one which yields the best result in terms of the BLEU score [11]. We then compare this approach to the third model of IBM [12]. Let us indicate that nowadays machine translation are based on phrases. In this paper we only focus on word-based machine translation. Work on phrase-based machine translation is under progress [13].

Manuscript submitted June 1, 2009

2 Information Retrieval based on association rules

Information retrieval (IR) studies the process of determining the adequacy between a user-defined query and a collection of documents, usually resulting in a subset of relevant documents [2]. With the explosion of the available on line information, IR techniques have to be more precise and more efficient. In this respect, *query expansion* is a technique that aims at reducing the usual query/document mismatch by expanding the query with terms that are highly correlated to those used in the original one [14].

To achieve that, a promising track consists in the application of data mining methods to extract hidden and valuable dependencies between terms. Among these techniques, association rule mining targets to retrieving correlated patterns [15]. A pattern could be any set of terms. An association rule binds two sets of terms: a premise and a conclusion. This means that the conclusion occurs whenever the premise is observed in the set of documents. To each association rule, a confidence value is assigned to measure the likelihood of the association. The use of such dependencies in a query expansion could increase the retrieval effectiveness.

In the sequel, we present association rule mining method based on Galois closure operator [1]. This method has the advantage to reduce redundancy within rules. In other words, it yields a compact representation of rules called *generic bases* [16], [17], [7], [18], [19].

2.1 Association rules and closed termsets mining

In text mining field, an *extraction context* is a triplet $\mathfrak{K} = (\mathcal{D}, \mathcal{T}, \mathcal{R})$ where $\mathcal{D}$ represents a finite set of documents, $\mathcal{T}$ is a finite set of terms and $\mathcal{R}$ a binary relation (i.e., $\mathcal{R} \subseteq \mathcal{D} \times \mathcal{T}$). Each couple $(d, t) \in \mathcal{R}$ means that the document $d \in \mathcal{D}$ contains the term $t \in \mathcal{T}$.

According to a specific application, the set $\mathcal{D}$ can be considered as a set of paragraphs or a set of sentences.

Techniques of association rules start with finding out the frequent sets of terms called *termsets* [1] from the textual context. These termsets must occur more than a fixed user-defined threshold, denoted *minsupp*.

Many representations of frequent termsets were proposed in the literature [15] where terms are characterized by the frequency of their co-occurrence. The ones based on *closed termsets* and *minimal generators* [16], [17] result from the mathematical bases of the Formal Concepts Analysis (FCA) [1]. Indeed, the mining process heavily relies on the *Galois closure operator* [1].

The Galois closure operator splits the set of frequent termsets into *equivalence classes*. Each class contains termsets characterizing the same set of documents. These termsets share the same closure which is obtained by intersecting the associated documents. A *closed termset* represents a maximal group of terms sharing the same documents. While, often several *minimal generators* constitute the *minimal* incomparable elements within each equivalence class. Intuitively, we can say that a closed termset includes the most general terms, while a minimal generator includes one of the most specific terms describing the set of documents.

In the following, we introduce some key results from the Galois lattice-based paradigm in FCA [1] and its applications to association rules mining.

2.1.1 Some reminders: Key FCA Settings

In the remainder, we recall some basic concepts of the theoretical framework presented in [1].

Galois closure operator

Two functions are defined in order to map sets of documents to sets of terms and *vice versa*. Thus, for a set $D \subseteq \mathcal{D}$, we define:

$$\Phi(D) = \{t \mid \forall d, d \in D \Rightarrow (d, t) \in \mathcal{R}\} \quad (1)$$

and for $T \subseteq \mathcal{T}$,

$$\Psi(T) = \{d \mid \forall t, t \in T \Rightarrow (d, t) \in \mathcal{R}\} \quad (2)$$

Both functions Φ and Ψ form a *Galois closure operator* between the sets $\mathcal{P}(\mathcal{T})$ and $\mathcal{P}(\mathcal{D})$ [1]. Consequently, the compound operator $\Omega = \Phi \circ \Psi$ is a closure operator.

Formal Concept

A formal concept is a pair $c = (D, T)$, where D is a set of documents, further called *extent*, and T is a termset, further called *intent*. Thus, both D and T are related through the Galois operators, i.e., $\Phi(D) = T$ and $\Psi(T) = D$.

Minimal generator

A termset $g \subseteq \mathcal{T}$ is a *minimal generator* of a closed termset T, if and only if $\Omega(g) = T$ and $\nexists\, g' \subset g$ such that $\Omega(g') = T$ [16].

Galois lattice

When the inclusion operator is performed on the set of formal concepts $\mathcal{C}_{\mathfrak{K}}$, this former constitutes a complete lattice $\mathcal{L}_c = (\mathcal{C}_{\mathfrak{K}}, \leq)$, called *Galois (concept) lattice* [1].

A partial order can be defined on the set of formal concepts as follows:
$\forall\ c_1, c_2 \in \mathcal{C}_{\mathfrak{K}}$, $c_1 \leq c_2$ if and only if $intent(c_2) \subseteq intent(c_1)$, or in an equivalent way $extent(c_1) \subseteq extent(c_2)$.
Given a concept c, we define the set of its immediate successors in the lattice, further called *upper covers* as follows: $Cov^u(c) = \{c_i \in \mathcal{C}_{\mathfrak{K}} | c \preceq c_i\}$, where $\preceq$ is the transitive reduction of $\leq$, i.e., $\forall c_3 \in \mathcal{C}_{\mathfrak{K}}$, $c_1 \leq c_2 \leq c_3$ implies either $c_1 = c_3$ or $c_2 = c_3$.

Lattice operators *join* and *meet* provide respectively the least upper bound (LUB) and the greatest lower

1. By analogy to the *itemset* terminology used in data mining.

bound (GLB) in the concept lattice.

Frequent reduced concept (FRC)
A closed termset $T \subseteq \mathcal{T}$, i.e., $T = \Omega(T)$, is frequent if its support verifies: $Supp(T) = \frac{|\Psi(T)|}{|\mathcal{D}|} \geq minsupp$. Henceforth, at each FRC a support is associated [6].

Iceberg Galois lattice
When the inclusion operator is performed on the set $\mathcal{FRC}$ of frequent reduced concepts , the resulting structure only preserving the LUBs, is an upper semi-lattice called Iceberg Galois lattice [16].

We will call *augmented Iceberg Galois lattice*, the standard Iceberg Galois where each FRC is associated to its minimal generators.

The approach presented in this paper describes only the principle of association rule discovery from the Iceberg Galois lattice [7].

2.1.2 Association rules

The association rule extraction problem has been introduced by Agrawal *et al*. [15]. The derivation of association rules is achieved starting from the set of frequent termsets extracted from the textual context $\mathfrak{K} = (\mathcal{D}, \mathcal{T}, \mathcal{R})$.

Let $T = \{t_1, t_2, \cdots, t_m\}$ be a set of m distinct terms and D a document containing a set of terms of T. Given a subset X of T such that $k = |X|$, then X is referred to as a $k-$termset, and k is its length.

An association rule R is an implication of the form $R : X \Rightarrow Y$, where X and Y are subsets of $\mathcal{T}$, and $X \cap Y = \emptyset$. The termsets X and Y are, respectively, called the *premise* and the *conclusion* of R.

The *support* of a rule $R : X \Rightarrow Y$ is defined as:

$$Supp(R) = Supp(X \cup Y) = |\Phi(X \cup Y)| \quad (3)$$

while its *confidence* is computed as :

$$Conf(R) = \frac{Supp(X \cup Y)}{Supp(X)} = \frac{|\Phi(X \cup Y)|}{|\Phi(X)|} \quad (4)$$

An association rule R is *valid* if its confidence value *Conf(R)* is greater than or equal than a fixed threshold denoted *minconf*. If *Conf(R)*=1 then R is called *an exact association rule (ER)*, otherwise it is called *an approximate association rule (AR)* [20].

2.2 FCA based algorithms for generating association rules

Given a corpus, the problem of mining association rules between terms consists in generating all association rules associated to a user-defined *minsupp* and a *minconf*. This problem can be split in two steps as follows [15]:

1) Extract all frequent termsets that occur in the corpus with a support value $\geq$ *minsupp*.
2) Generate valid association rules between terms from frequent termsets, i.e., rules whose confidence $\geq$ *minconf*.

These steps are performed by a pioneer algorithm in data mining, namely APRIORI [15]. This problem deals with an exciting challenge: how to retrieve only the most pertinent associations from the huge number of possibilities ($2^{|T|} - 2$ for a frequent termset T)?
In fact, during the first step, the set of frequent termsets may grow exponentially with respect to $\mathcal{T}$. Whereas, the second step is an exponential issue depending on the length of the longest frequent termset. These rules can be generated in a straightforward manner, i.e., without any further access to context [15]. Nevertheless, the number of discovered association rules may grow up to several millions [20] while a large number of them could be redundant [7], [18], [19].

Several approaches deal with the redundancy problem. For instance some works relied on the use of other quality measures in addition to the support and confidence, as lift, conviction, dependency, etc. [21], while others introduced user-defined constraints during the mining process or on a post-processing step [22].

More advanced techniques that produce only limited number of rules rely on Galois closure [1]. These techniques focus on extracting irreducible nuclei of all association rules, called *generic basis*, from which the remaining association rules can be derived without information loss [16], [20], [7], [18], [19].

The majority of generic bases convies association rules presenting implications between minimal generators and closed termsets [16], [17], [7], [19]. This ensures obtaining association rules with minimal premise and maximal conclusion part. Such rules convey the maximum of information, and are hence qualified as the most informative ones [16]. An interesting discussion about the main generic bases of association rules proposed in the literature is given in [19].

2.3 Extracting the minimal generic basis of association rules

We propose to adapt the Minimal Generic Basis $\mathcal{MGB}$ of association rules presented in a previous work [7] to the text mining context.

When considering a context $\mathfrak{K} = (\mathcal{D}, \mathcal{T}, \mathcal{R})$, the minimal generic basis $\mathcal{MGB}$ is defined as follows [7]:

Given :

- $\mathcal{L}_c$: Iceberg Galois lattice augmented by minimal generators and their supports.
- c_i: frequent reduced concept.
- $Cov^u(c_i)$: upper cover of the frequent reduced concept c_i.
- $\mathcal{G}_{c_i}$: list of minimal generators of the frequent reduced concept c_i.

$$\mathcal{MGB} = \left\{ \begin{array}{l} R : g \rightarrow (c_i - g) \mid g \in \mathcal{G}_{c_i} \wedge c_i \in \mathcal{L}_c \wedge \\ Conf(R) \geq minconf \wedge \nexists\ s \in Cov^u(c_i) \mid \\ \frac{support(s)}{support(g)} \geq minconf \end{array} \right. \quad (5)$$

According to equation 5, non redundant association rules are directly derived, without additional calculations of confidence measure. Approximate rules of the form, $g_1 \Rightarrow (c_2 - g_1)$ are generated where $c_1 \preceq c_2$ and $g_1 \in \mathcal{G}_{c_1}$. However, the derived exact rules have the following form: $g_i \Rightarrow (\Omega(g_i) - g_i)$, given that g_i does not appear as premise of any another valid approximate rule.

In this respect, the problem of mining non redundant association rules is reformulated as follows:

1) **Discover frequent reduced concepts** (GEN-FRC): The CARD algorithm detailed in [6] operates in a level-wise manner to retrieve all frequent closed termsets and their minimal generators.
2) **Discover the upper cover** (GEN-LATTICE): To derive the generic basis $\mathcal{MGB}$, the set of immediate successors of each frequent reduced concept in the Iceberg Galois lattice is needed.
3) **Extract non redundant association rules between terms** (GEN-RULE): Our algorithm GEN-MGB takes the Galois Iceberg lattice as input and returns the approximate and exact association rules.
   The set of approximate rules represents implications from a sub-concept to a super-concept (inter-node implications), assorted with the confidence [23], [24]. These rules are derived while starting from a given node in the augmented Iceberg lattice. On the other hand, exact rules are intra-node implications.

The pseudo-code of GEN-MGB algorithm is given by Algorithm 1. It iterates on all frequent reduced concepts of the augmented iceberg lattice $\mathcal{L}_c$, starting from the border and sweeping downwardly (with respect to $\subseteq$).

Algorithme 1 GEN-MGB algorithm to derive the minimal generic basis $\mathcal{MGB}$ of association rules from a textual context.

Algorithm GEN-MGB
Require: $\mathfrak{K}$: the textual context, the *minsupp* threshold and the *minconf* threshold.
Ensure: The Minimal Generic Basis $\mathcal{MGB}$ through the following steps
1) $\mathcal{FRC}$= GEN-FRC($\mathfrak{K} = (\mathcal{D}, \mathcal{T}, \mathcal{R})$,*minsupp*)
2) $\mathcal{L}_c$=GEN-LATTICE($\mathcal{FRC}$)
3) $\mathcal{MGB}$=GEN-RULE($\mathcal{L}_c$,*minconf*)
Return $\mathcal{MGB}$

3 APPLICATION 1: QUERY EXPANSION IN INFORMATION RETRIEVAL USING ASSOCIATION RULES

The aim of the retrieval activity is to maximize the usefulness of the retrieved documents and to privilege precision over recall. Hence, the idea of query expansion with association rules between terms is to found additional relevant documents and improve their ranking in the list of retrieved ones. Thus, our objective is to enhance precision at low level of recall. We introduce some results of experiments carried out with the two test collections OFIL and INIST of the AMARYLLIS project[2][3]. We will compare the efficiency of the expanded queries with that of original ones.

3.1 Evaluation

In our experiments, we used the information retrieval system SMART[3] [2]. The process of automatic query expansion is the following:

1) *Standard run:* finding the best results of SMART system. These results were evaluated by applying the average precision of the original query (OQ) set at eleven representative recall points.
2) Expanding each query of the collection by all terms that appear in the conclusions of the association rules related to the terms of the original query.
3) *Second run:* A second run is launched with the expanded queries (AREQ) and an evaluation is performed under the same conditions.

For example, in OFIL corpus, the term *conflict* occurs in the premise of 260 valid association rules. Consequently, the query is expanded by all terms that are in the conclusions of these associations rules. For instance, *conflict* has been associated to the following corresponding French words such as: difficulty, solution, Bosnia, security, Serbia, etc.

3.2 Training corpus

We used two different collection :

- Three months of articles from the daily French newspaper *Le Monde* extracted from the OFIL collection [3] which has 11016 heterogeneous articles and 119434 different terms.
- Titles and abstracts of scientific papers catched from the *PASCAL* (four years) and *FRANCIS* (one year) databases extracted from the INIST [3] collection which contains 165431 scientific articles and 174659 different terms.

For a set of queries is associated to each collection and for each query, a set of relevant documents is assigned. We distinguish 26 and 30 queries, respectively, OFIL and INIST collections. In order to extract the most representative terms, a linguistic preprocessing is performed on both corpora. In this application, we focus only on terms related to two grammatical categories: *the common substantives* (SUBC) and the *own substantives* (SUBP). One empty word list is used to discard terms that are very common, e.g. *today*, *yesterday*, etc.

2. AMARYLLIS project is initiated by INIST-CNRS and co-funded by AUPELF-UREF. Its goal is to evaluate French Text retrieval systems.
3. System for the Mechanical Analysis and Retrieval of Text is an IR system

The context document-term $\mathfrak{K}$ is built by selecting only terms corresponding to the selected grammatical categories. The association rules are then generated from the Iceberg Galois lattice using the GEN-MGB algorithm [7]. The minimal threshold of confidence is set to 0% and we varied the minimal and maximal threshold of the support, i.e., *minsupp* and *maxsupp*[4], with respect to the corpus size and to term distributions. While considering the *zipf* distribution of every corpus, minimal and maximal thresholds of support values are set experimentally in order to spread trivial terms as well as marginal ones.

3.3 Results

Table 1 shows the retrieval quality difference between the original queries (OQ) and the expanded ones using association rules (AREQ) derived while considering *minsupp*=5 documents and *maxsupp*=50 documents. These results are expressed in terms of the average precision at the 11 recall points for both corpora OFIL and INIST.

Corpus	OQ	AREQ
OFIL	39.93%	42.51%
INIST	22.10%	22.41%

TABLE 1
Expansion query results by using association rules in term of average precision

Figure 1 shows that for the OFIL corpus, our method yields an improvement of the average precision at 11 recall points of $+6.5\%$. As illustrated in Table 2, we notice that the exact precision at low recall clearly increases at 5, 10, 15 and 30 documents. This means an increase in the number of founded relevant documents by putting them in the head of the list of retrieved documents. **Hence, our query expansion approach based on association rules produces a statistically significant increase in performance in mean of average precision over the baseline system.**

Exact Precision (in %)			
Recall	OG	AREQ	Δ
At 5 documents	42.31	46.92	4.61
At 10 documents	43.08	45.77	2.69
At 15 documents	38.46	43.59	5.13
At 30 documents	32.18	34.87	2.69

TABLE 2
Exact precision at 5, 10, 15 and 30 documents on OFIL corpus

Moreover, we notice that the improvement of the average precision is less significant for higher support values. Extracting association rules, when considering a high support values, leads to some trivial associations between terms that are very frequent in corpora. Therefore, if we expand queries using these terms, we will not improve neither the recall nor the precision.

4. *maxsupp* means that the termset must occur less than this user-defined threshold.

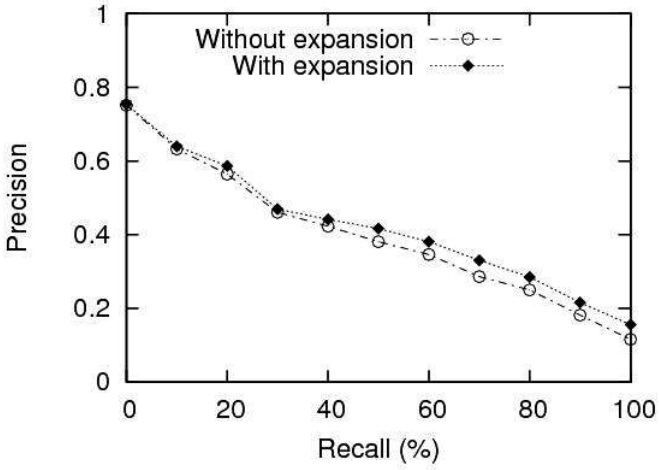

Fig. 1. Recall/precision diagram of OFIL corpus for a *minsupp*=5 and a *maxsupp*=50 (Improvement $+6,50\%$)

According to Table 1, we notice that the improvement of the average precision is smaller for INIST than for OFIL corpus. It can be justified by:

- INIST is a scientific collection where terms have a very weak distributions and marginally co-occurs .
- An important part of the vocabulary is not used, since is not correctly analyzed, due to the tagger which does not identify specific and scientific terms of INIST.

4 APPLICATION 2: MACHINE TRANSLATION

The second application of association rules is to use them in the context of machine translation. In the following we will explain how to achieve a translation table from association rules and to compare them to an original method in machine translation proposed in [9].

In order to understand how to use them in this area, we propose to give an overview about machine translation. In the following, we will focus only on Statistical Machine Translation (SMT).

Indeed, statistical techniques have been widely used and have been really successful in automatic speech recognition, machine translation and in natural language processing over the last two decades. This success is due to the fact that this approach is language independent and requires no prior knowledge. This technique requires a large amount of suitable data to carry out the estimation of significant parameters. In SMT, one needs bilingual aligned corpora to estimate all the necessary models. It is then very exciting to investigate text mining techniques and especially those used in this paper to retrieve from parallel corpora inter-lingual associations required for translation.

4.1 Principle of Statistical Machine Translation

In this framework, the translation process is essentially the search for the most probable sentence in the target language e given a sentence in the source language f. Let $f = f_1, .., f_i$ be the source sentence (i.e., to be translated) and $e = e_1, .., e_j$ be the sentence generated by the system:

$$\hat{e} = \arg\max_e P(e|f) \tag{6}$$

By using Bayes rule and by omitting the denominator (which does not depend on e), we obtain:

$$\hat{e} = \arg\max_e P(e)P(f|e) \tag{7}$$

In Equation 7, $P(e)$ is estimated by a *language model*. Its role is to propose a sentence supposed to be correct in the target language. Notice that $P(f|e)$ is computed from a *translation model* and is supposed to reflect the truthfulness of the translation. The decoder then generates the best hypothesis by making a compromise between, at least, these probability distributions.

4.1.1 Evaluation of Translation

The best way to evaluate a translation system is to ask a human to score each output of the system. Unfortunately, a machine translation system is developed over several incremental optimization steps. Therefore, it is time consuming and expensive to ask a human to evaluate at each step. That is why several measures have been proposed to automatically evaluate such systems. Among these measures, we can cite: WER, NIST, BLEU, etc. BLEU (BiLingual Evaluation Understudy) is one of the most used measure, it has been proposed by [11]. The way that BLEU and other automatic evaluation metrics work is to compare the output of a machine translation system to reference human translations. BLEU is a n-gram precision based over several references. Its objective is to increase score of the solutions which looks like, in terms of n-grams, the references.

$$BLEU = BP \times \exp^{\sum_{n=1}^{N} \alpha \log P_n} \tag{8}$$

with P_n is the probability of a sequence of n words, α gives a weight to the translated n-gram and BP is a brevity penalty which decreases the score of the translations which are shorter than the reference [11].

4.1.2 Bilingual corpus

Statistical machine translation needs a material raw in order to estimate different parameters. In the following experiments are carried out on the proceedings of the European Parliament [25]. We used a French-English parallel corpus of 596831 sentence pairs. The French side has a total of 17 million words (77567 unique tokens). The English side has a total of 16 millions words (60331 unique tokens).

Table 3 gives details about the used parallel corpus EUROPARL.

		French	English
Train	Sentences	$596K$	
	Words	$17.3M$	$15.8M$
	Singletons	$26.6K$	$22.2K$
	Vocabulary	$77.5K$	$60.3K$
Development	Sentences	1444	
	Words	$15.0K$	$14.0K$
Test	Sentences	500	
	Words	$5.2K$	$4.9K$

TABLE 3
Quantitative description of the EUROPARL corpus

4.2 Another way to achieve translation: Inter-lingual triggers

IBM proposed in [12] five models allowing to compute the translation probabilities. These methods become unavoidable and are used by so many people in machine translation community. In the sequel, we will introduce a new idea to achieve a translation table which has been proposed in an earlier work [9]. It is based on mining a parallel corpus in order to find out the couple of words or phrases which are translation one of others.

4.2.1 A brief remind of Triggers

The concept of triggers has been largely used in statistical language modeling. Roughly, a statistical language model yields a probability to each potential sequence of words belonging to a vocabulary. Triggers are a special kind of language model which is inspired from the Cache model [26]. The Cache model enhances the probability of a word w_i when it occurs in its left context which makes the sequence to which it belongs more likely. A trigger model goes further and enhances the probability of a list of words which are correlated to w_i [27]. To achieve that, all the correlated words are retrieved as illustrated in the example of Figure 2.

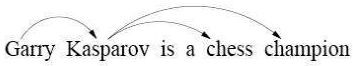

Fig. 2. Example of correlated words

Triggers are determined by computing mutual information between two random variables X, Y, each of them takes its values on the list of words belonging to the vocabulary V of the language model. Then for two words x,y, the correlation is given by:

$$I(x,y) = P(x,y)log_2\frac{P(x,y)}{P(x)P(y)} \tag{9}$$

For each vocabulary entry the n best correlated words in terms of mutual information are kept. We call a trigger a set made up of a trigger and its triggered words.
In language modeling triggers are used as a new language model which is interpolated with a classical n-gram [28].

4.2.2 Inter-lingual triggers

Inter-lingual triggers have been also used in [29] to enrich resource deficient languages from those which are considered as potentially important.
An inter-lingual trigger is henceforth a set made up of a word (or a sequence of words) f in a source language, and its best correlated words in a target language $e_1, e_2, \ldots, e_n$. This will be written as: $Trig(f) \longrightarrow e_1, e_2, \ldots, e_n$. The method we propose, could produce intra-language triggers (classical one) and inter-language triggers. That means Source-Source, Target-Target, Source-Target and Target-Source triggers are calculated. In order to find out these triggers, all the pairs of parallel sentences have been concatenated inside the same corpus as shown in Figure 3. The triggers in which we are interested are depicted.

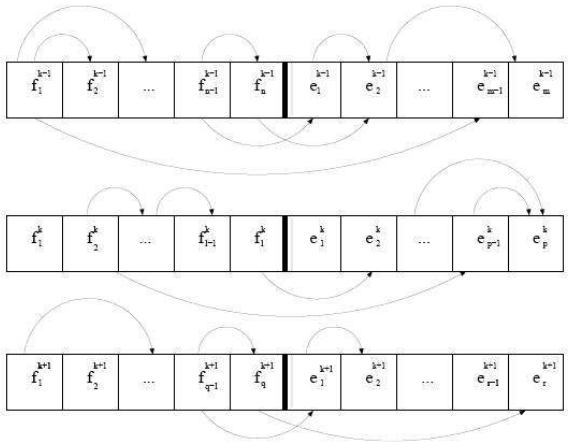

Fig. 3. Inter-Lingual Triggers

In this work, we will only focus on pairs of triggers such that the trigger word is in a source language and the triggered one is in a target language as in the exemple depicted in Figure 4.

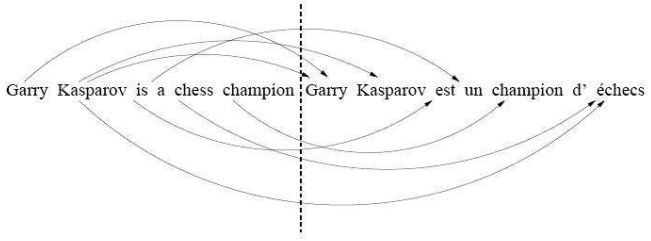

Fig. 4. Example of inter-lingual triggers

Inter-lingual triggers are determined on a parallel corpus according to the following formula:

$$MI(f,e) = P(f,e) log \frac{P(f,e)}{P(f)P(e)} \quad (10)$$

where f (respectively e) is a French (respectively English) word. $MI(f,e)$ denotes the mutual information assigned to e and f and $P(e)$, $P(f)$ and $P(f,e)$ are defined as follows:

$$P(X) = \frac{N(X)}{N} \qquad P(f,e) = \frac{N(f,e)}{N} \quad (11)$$

where $N(X)$ is the number of sentences where X occurs, $N(e,f)$ is the number of sentence pairs where e and f co-occur and N is the number of sentence pairs in the training corpus.
For each French word f, we keep as inter-lingual triggers, the k best English words $e_1 \ldots e_k$ in terms of MI.

The above formula looks like the one used in the literature but is not exactly the same. In fact, our objective is to lead to machine translation dictionary without using any external knowledge. That is why the mutual information is calculated inside a window which has the length of a concatenated pair of sentences (for which one is the translation of the other). Clearly, we would like to retrieve the words in a target language $E = e_1, e_2, \ldots, e_n$ which are correlated to a word f in a source language. Among the set E, we hope to find a subset T which is made up only by the translations of f. The translation table will be composed of a set of French trigger f_i, the associated English triggered words $E_i = e_1^i, e_2^i, \ldots, e_n^i$ and for each e_j^i is assigned a probability which is induced from the inter-lingual mutual information such as:

$$\sum_{j=1}^{|E_i|} P(e_j^i | f_i) = 1 \quad (12)$$

4.3 Inter-lingual association rules

The idea to use associations between terms in machine translation is owe to the work presented above [9]. This approach is based on inter-lingual triggers to provide automatically a bilingual dictionary in multiple languages. This method drive us to adapt the association rules to make them working on parallel corpus. For that, we introduce the concept of *inter-lingual association rule* (ILAR). Let R be an implication of the form $X \Rightarrow Y$ where X and Y are two termsets and $X \cap Y = \emptyset$ and *Language(X)* $\neq$ *Language(Y)*. Because we consider a parallel corpus as a single document in which a sentence and its translation are gathered in the same one, we keep the definition of support and confidence given respectively in equations 3 and 4.

We extracted closed frequent termsets and their minimal generators. We tried a battery of algorithms dedicated to this task such as PRINCE [8], CLOSE, A-CLOSE [30], CHARM [4], TITANIC [5] and CARD [7]. We noticed that none of these algorithms succeeds while considering a so highly sized textual context which contains 596381 sentences and 137898 different terms. This kind of corpus is not considered as huge in applications as speech recognition or machine translation, whereas in textmining it is not often to use a such highly sized

corpus. Our experiments showed the limits of algorithms based on FCA to extract formal reduced concepts and generic bases of association rules since we consider raw textual corpora having huge sizes. However, these algorithms are considerably more successful if we consider structured data such as synthetic datasets.

To overcome this limit, we adapted the GC-GROWTH algorithm [31] for the extraction of the frequent closed termsets, i.e. frequent reduced concepts set $\mathcal{FRC}$, and their generators by varying the minimal support as depicted in Table 4. We obtain a very important number of frequent reduced concepts and the corresponding generators.

In order to generate associations between terms from reduced frequent concepts set $\mathcal{FRC}$, we adapt our algorithm GEN-MGB to get out only associations having a single term in the premise and a single term in the conclusion, applying the decomposition axiom [7]. The idea of the conditional decomposition is that from each rule $r : f \Longrightarrow e_1\ e_2\ \ldots\ e_n \in \mathcal{MGB}$, the different association rules $r : f \Longrightarrow e_i$ are derived as valid rules. When the frequent termset (fe_i) is not in the $\mathcal{FRC}$ then $\forall\ i \in [1..n]$ we assign to it as support value the one corresponding to the smallest concept in $\mathcal{FRC}$ including (fe_i).

Nevertheless, for low support values and even with the use of optimizations brought by the GEN-MGB algorithm, the huge number of closed termsets as well as that of minimal generators, constitute an actual hamper towards an efficient extraction of inter-lingual association rules, based on Iceberg Galois lattice sweeping. In fact, the lower the support value is, the higher density of the extraction context is. For that reason, we are not able to lower the *minsupp* threshold beyond 20 sentences. This threshold is considered as very weak with respect to what is usually used by the data mining community. Some runs are summarized in Table 4, where MinThresh is the *minsupp* threshold, N(FRC) the number of FRC and N(Gen) the number of generators. The number of FRC and generators are expressed in millions.

MinThresh	N(FRC)	N(Gen)
30 sentences ($25 \times 10^{-4}\%$)	3.50	2.70
25 sentences ($16 \times 10^{-4}\%$)	4.70	3.64
20 sentences ($11 \times 10^{-4}\%$)	5.20	6.70

TABLE 4
Size of frequent reduced concepts and their generators for three fixed *minsupp* values

4.4 How to obtain a translation table from inter-lingual association rules?

We constructed a unique dictionary including English and French words. The vocabulary is built up from the union of the most frequent French words and English ones according to the fixed *minsupp* threshold. For all the "tool" words (small words in English and French as: or, it, in, thus,..., de, la, le, donc, etc), we generate in a first time a parallel corpus containing only these words. Inter-lingual associations rules between them are then derived using GEN-MGB algorithm. Their translations are then added into the final dictionary.

The basic idea is to automatically provide a bilingual dictionary from the discovered inter-lingual associations. The potential translations of a French term f which appears in a premise of an association rule are obtained by selecting all the English terms $e_1, e_2, ..., e_n$ which are present in conclusions of the same inter-lingual association.

Namely, an entry in a French-English dictionary D is defined as:

$$f \Longrightarrow e_1, e_2, \ldots, e_n \in D \Leftrightarrow \forall\ j \in [1..n],\ r : (f \Longrightarrow e_j, Conf_j) \in \mathcal{MGB}\ \wedge\ Conf_j \geq minconf \quad (13)$$

To achieve a translation table using inter-lingual association rules, we assign to each potential term translation which occurs in the conclusion part a probability computed from the confidence value such as:

$$\forall\ f, e_i \in PT(f), P(e_i|f) = \frac{Conf(f \Longrightarrow e_i)}{\sum_{e \in PT(f)} Conf(f \Longrightarrow e)} \quad (14)$$

where $PT(f)$ stands for the potential translations of the French term f.

In most cases, experiments showed that the exact rules, i.e., with full confidence equal to 1, achieve correct translations of French entries. We observe that, our algorithm GEN-MGB generates also approximate associations with a strong confidence. Their conclusions represent potential translations of the terms within their premises. However, we notice that inter-lingual association rules explore non significant translation which introduce noise in the bilingual dictionary. These associations can appear with high or low confidences since the filtering is based only on statistical metrics, namely *minsupp* and *minconf*.

Examples produced by inter-lingual association rules (ILAR) and inter-lingual triggers (ILT) translations are illustrated in Table 5.

5 RESULTS OF INTER-LINGUAL ASSOCIATION RULES AND INTER-LINGUAL TRIGGERS ON MACHINE TRANSLATION

To deeply evaluate inter-lingual associations and inter-lingual triggers, we will integrate them in machine translation system. They will be used to build a translation table instead of those usually used by the community (IBM models). For the language model, all the experiments below are conducted by using a 3-gram model. Decoding is achieved by PHARAOH [32] and the models are evaluated by using the BLEU score. As mentioned in the beginning of this paper, the tests below do not use a phrase-based machine translation.

French	ILT	$P_{ILT}(e_i\|f_i)$	ILAR	$P_{ILAR}(e_i\|f_i)$
Coopération	cooperation	0.52	cooperation	0.76
	development	0.06	development	0.12
	countries	0.04	countries	0.11
Pêche	fisheries	0.34	fisheries	0.45
	fishing	0.26	policy	0.12
	fish	0.06	fishing	0.35
Difficulté	difficulty	0.43	problem	0.17
	difficulties	0.12	difficulties	0.22
	difficult	0.09	difficulty	0.60
Compétences	powers	0.31	powers	0.55
	competences	0.13	competences	0.16
	competence	0.10	skills	0.09
Alimentaire	food	0.50	safety	0.28
	safety	0.16	industry	0.02
	chain	0.07	food	0.68
Tempêtes	storms	0.50	storms	1.0
	floods	0.07	-	-
	storm	0.06	-	-

TABLE 5
Examples of English units translated by respectively inter-lingual triggers (ILT) and inter-lingual association rules (ILAR)

Minsupp	*Minconf*	Size of ILAR-Dic	BLEU
20	0.05	65520	20.26
	0.01	355181	**22.07**
30	0.05	49581	20.06
	0.01	243932	21.86

TABLE 7
Decoding test results with Inter-Lingual Association Rules

5.1 Experiments on inter-lingual associations

While using inter-lingual association rules on machine translation, we considered some settings bounded to the size and to the nature of parallel corpora, namely:

- To keep the maximum of the vocabulary words in our mining process, we set the *minsupp* threshold to 20 sentences. For such a threshold, an overwhelming number of associations are generated.
- To maximize the number of entries in the bilingual dictionary, we consider very weak values of the *minconf* threshold. This is a major hamper because we accept pairs of terms which are not strongly correlated and do not necessarily express correct translations.

Table 6 gives the best values of the parameters necessary to run PHARAOH. They have been optimized on the development corpus described in Table 3. These parameters will be explained in subsection 5.2.4.

ttable-limit	ttable-threshold	wd	wl	wl	w
30	0.1	0.2	0.8	1	-1

TABLE 6
Optimization of PHARAOH parameters on the development corpus for ILAR

To the best of our knowledge, this work is the first where association rules have been adapted and generalized in order to retrieve inter-lingual associations. In addition, we show the feasibility of this approach. A BLEU score of 22.07 has been achieved.

5.2 Experiments on inter-lingual triggers

One of the most crucial component of machine translation is its translation table. That is why in the following, we present few experiments in order to find out the best one. In order to achieve that, we compare inter-lingual triggers to the third model of IBM [12].

5.2.1 Baseline Triggers: Trig-n

In this case, we consider all the triggered words of an entry as potential translations. We call these triggers *Trig-n* with n the number of potential translations accepted for each entry. The evolution of the BLEU score is given by the chart of *Trig-n* of Figure 5. We notice that by varying n from 0 to 200, the BLEU score increases by 2 points between trigg-10 and Trigg-20. This shows that basically, the best translations are in the first twenty triggered words. Beyond 20 potential translations the contribution is not significant.

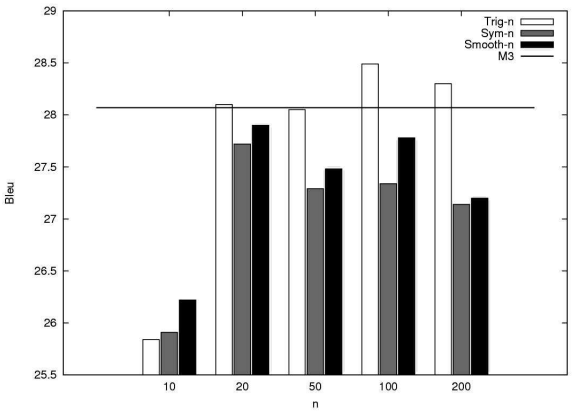

Fig. 5. Several inter-lingual trigger experiments based on different translation tables

5.2.2 Symmetric translation table: Sym-n

The second way to build a dictionary consists in considering as possible translations the couples (f_j, e_i) which respect the following constraint:

$$e_i \in Trig\text{-}n(f_j) \;\; and \;\; f_j \in Trig\text{-}n(e_i) \qquad (15)$$

This means that translations of a word e are obtained by selecting all the target triggered words $f_1, f_2, \ldots, f_n$ which trigger the source word e as illustrated in Figure 6. This allows us to refine the triggers of f_j by taking into account only those which are relevant. In other words, if e_i is correlated to f_j and f_j is triggered by e_i, then it

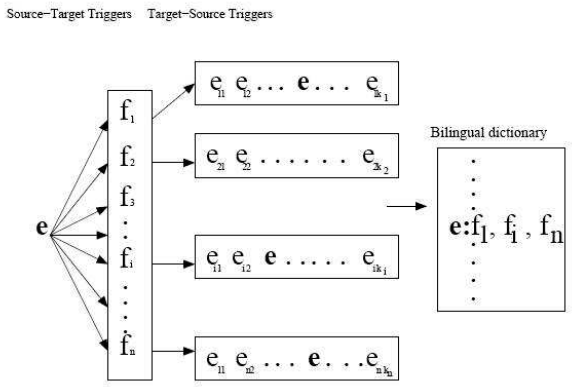

Fig. 6. A symmetric dictionary

is likely to guess that one is the translation of the other. Figure 5 show that the results of *Sym-20* overcome those obtained by *Sym-10*. We also notice a slight improvement in comparison to *Trig-10*. However, since 20 potential translations, this model is not better than *Trig-n*. Indeed, this constraint is too strong and discards several possible translations for an entry.

5.2.3 Smoothed translation tables: Smooth-n

In order to avoid zero-probability to the couples of translation which do not respect the symmetry constraint, we propose in this experiment to smooth the probability. In statistical language models, smoothing techniques are systematically used [33]. To achieve that, we use a simple smoothing method for which each translation probability is decreased. The gained mass probability is distributed over other translation possibilities.

$$\forall e_j \in Trig\text{-}n(f_j),\ P_s(e_i|f_j) = \begin{cases} P_{sym}(e_i|f_j) - \varepsilon \\ \qquad if\ e_i \in Sym\text{-}n(f_j) \\ \alpha\ otherwise \end{cases} \tag{16}$$

5.2.4 Comparison with IBM Model

In order to evaluate the relevance of inter-lingual triggers, we compare it to the third IBM model trained with GIZA++ [34]. The result is illustrated in Figure 5 (see line M3) and shows that Trig-100 achieves a slight improvement. In fact Trig-100 gives a score of 28.49 whereas M3 reaches a score of 28.07. Let notice that improving BLEU score is very difficult, and even a slight improvement is considered positively. In order to go further, we tried to set the parameters of PHARAOH differently. Indeed, for optimization reasons, PHARAOH restricts its research for each word to the 20 best translations. We then optimized two parameters *ttable-limit* and *ttable-threshold* for both models (M3 and Trig-100). The corresponding results are given in Table 8.

Trig-100 is optimal when we set the number of potential translations to 22 with a probability greater than 0.04 whereas IBM model (M3) is optimal when it uses 53 words for each translation without any restriction on the probability of each entry. We can conclude that with less words, *Trig-100* outperforms M3 by $2,4\%$ which means that inter-lingual triggers is a good principle to find out the best translations of an entry. The optimisation carried out with PHARAOH shows that the best translations of an entry are in the top 22 list whereas for M3, the best translations are within the 53 first translations.

Model	ttable-limit	ttable-threshold	BLEU
Trig-100	22	0.04	28.95
M3	53	0.00	28.27

TABLE 8
Optimization of parameters *ttable-limit* and *ttable-threshold* on development corpus

After the optimization of the previous parameters on the development corpus, we optimized the weights of the different components used by PHARAOH decoder, namely :

- tm: The weight of the translation table.
- lm: The weight of the language model.
- d: The distorsion model.
- w: The penalty word.

This optimization improves respectively Trig-100 by 2.07 and M3 by 0.28. These results are summarized in Table 9.

Model	tm	lm	d	w	BLEU
Trig-100	0.9	0.8	0.4	-3	**31.02**
M3	0.6	0.7	0.4	-1	**29.23**

TABLE 9
Optimization of parameters tm, lm, d, w

After all these optimizations we evaluate Trig-100 on a test corpus of 500 sentences with the best parameters obtained on the development corpus. The achieved results are illustrated in Table 10.

Model	BLEU
Trig-100	**30.97**
M3	**29.57**

TABLE 10
System evaluation on the test corpus

6 DISCUSSION

We have described two approaches to address the issue of machine translation. The first one based on inter-lingual association rules inspired from classical methods of datamining. Association rules are first tested on query expansion. Experiments are conducted on two different corpora. The first one extracted from the French

newspaper Le Monde and the second one extracted from scientific documents. On both corpora, the use of association rules outperforms the results in terms of recall in comparison to the basic queries.

To take advantage from association rules, we boosted them and make them managing two different languages. Henceforth, the left side of a rule uses a word in a source language and the right side uses words from a target language. This allows us to build a translation table and to integrate it in an operational machine translation. The results obtained showed the feasibility of this approach and achieved a BLEU score of more than 22 on EUROPARL. The work of extending association rules to handle phrases is in progress and it is based on closed frequent sequences.

The second method presented in this paper is based on a new concept called inter-lingual triggers. This concept is a generalization of the one used in statistical language modeling. English and French corpora have been concatenated at the sentence level. Inter-lingula triggers have been then retrieved. The best inter-lingual triggers have been then selected to constitute the translation table. After several experiments we lead to the best one which achieves a BLEU score of 30.97 whereas the model M3 of IBM leads to a score of 29.57. In addition, we showed that for inter-lingual triggers the best translations are in the 22-top list whereas for the one proposed by IBM is on the 53-top list.

In conclusion, two methods of corpora mining have been proposed to discover pair of translations. They showed that it is possible to handle statistical machine translation differently. We work right now on developing phrase-based machine by using these two new concepts. We proposed several algorithms in order to extend both association rules and inter-lingual triggers to make them supporting phrases on the left and right sides. The work is under progress.

REFERENCES

[1] B. Ganter and R. Wille, *Formal Concept Analysis*. Springer-Verlag, Heidelberg, 1999.

[2] G. Salton, "The SMART retrieval system: experiments in automatic document processing," *Prentice-Hall series in automatic computation, New Jersey*, vol. 1377, March 1971.

[3] J. P. Chevallet, H. Haddad, and M. Géry, "Actes de la campagne de tests 'AMARYLLIS II : Expérimentations et résultats," in *Atelier final de la campagne Amaryllis 2, Paris*, Avril 2000.

[4] M. Zaki and C. Hasio, "CHARM: An efficient algorithm for closed association rule mining," in *Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington, VA, USA*, April 2002.

[5] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal, "Computing Iceberg concept lattices with TITANIC," *IEEE Transactions on Knowledge and Data Engineering*, vol. 2, no. 42, pp. 189–222, July 2002.

[6] C. Latiri, S. BenYahia, and G. Mineau, "Conceptual non-redundant association rules discovery: Application to query expansion," in *Proceedings of the First International Conference on Formal Concept Analysis: The State of the Art (ICFCA03), Darmstadt, Germany*, February-March 2003.

[7] C. Latiri, W. Bellagha, and S. BenYahia, "VIE-MGB: A Visual Interactive Exploration of Minimal Generic Basis of Association Rules," in *proceedings of the The third International Conference on Concept Lattices and their Applications (CLA'05), Olomouc, Czech Republic*, September 2005, pp. 179–196.

[8] T. Hamrouni, S. Ben Yahia, and Y. Slimani, "PRINCE: An algorithm for generating rule bases without closure computations," in *Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'05), Copenhagen, Denmark, LNCS*, vol. 3589. Springer-Verlag, August 2005, pp. 346–355.

[9] C. Lavecchia, K. Smaïli, D. Langlois, and J. P. Haton, "Using inter-lingual triggers for machine translation," in *Proceedings of the Tenth Interspeech, Antwerp, Belgium*, August 2007.

[10] R. Rosenfeld, "Adaptive statistical language modeling: a maximum entropy approach," Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1994.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, 2001, pp. 311–318.

[12] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[13] C. Lavecchia, K. Smaïli, and D. Langlois, "Discovering phrases in machine translation by simulated annealing," in *Proceedings of the Eleventh Interspeech Conference, Brisbane, Australia*, September 2008.

[14] M. Adriani and C. J. V. Rijsbergen, "Term similarity-based query expansion for cross-language information retrieval," in *Proceedings of the International Conference ECDL'99, LNCS*, vol. 1696, 1999, pp. 311–322.

[15] R. Agrawal and R. Skirant, "Fast algorithms for mining association rules," in *proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile*, June 1994, pp. 478–499.

[16] Y. Bastide, N. Pasquier, R. Taouil, L. Lakhal, and G. Stumme, "Mining minimal non-redundant association rules using frequent closed itemsets," in *proceedings of the international Conference DOOD'2000, LNCS, Springer-verlag*, July 2000, pp. 972–986.

[17] N. Pasquier, Y. Bastide, R. Taouil, G. Stumme, and L. Lakhal, "Generating a condensed representation for association rules," *Journal of Intelligent Information Systems*, vol. volume 24(1), pp. 25–60, 2005.

[18] M. Z. Ashrafi, D. Taniar, and K. Smith, "Redundant association rules reduction techniques," *International Journal Business Intelligence and Data Mining*, vol. 1, no. 2, pp. 29–63, 2007.

[19] S. BenYahia, G. Gasmi, and E. M. Nguifo, "A new generic basis of factual and implicative association rules," *In Intelligent Data Analysis (IDA)*, To appear 2009.

[20] M. J. Zaki, "Mining non-redundant association rules," *Journal of Data Mining and Knowledge Discovery (DMKD)*, vol. 9, pp. 223–248, 2004.

[21] F. Guillet and H. J. Hamilton, *Quality Measures in Data mining*. Studies in Computational Intelligence, Springer-Verlag, 2007.

[22] R. Srikant, Q. Vu, and R. Agrawal, "Mining association rules with item constraints," in *proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California*, August 1997.

[23] V. Duquenne and J. L. Guigues, "Famille minimale d'implications informatives résultant d'un tableau de données binaires," *Mathématiques et Sciences Humaines*, vol. 95, no. 24, pp. 5–18, 1986.

[24] M. Luxenburger, "Implications partielles dans un contexte," *Mathématiques, Informatique et Sciences Humaines*, vol. volume 29, no. 113, pp. 35–55, 1991.

[25] P. Koehn, "EUROPARL: A multilingual corpus for evaluation of machine translation," in *MT Summit, Thailand*, 2005.

[26] R. Kuhn and R. DeMori, "A cache-based natural language model for speech recognition," *IEEE Trans. PAMI*, vol. 12, no. 6, pp. 570–582, 1990.

[27] C. Tillmann and H. Ney, *Selection criteria for word trigger pairs in language modeling*. LNAI, Springer Verlag, 1996, vol. 1147, pp. 98–106.

[28] ——, "Word trigger and the EM algorithm," in *Proceedings of the Conference on Computational Natural Language Learning*, Madrid, Spain, 1997, pp. 117–124.

[29] W. Kim and S. Khudanpur, "Lexical triggers and latent semantic analysis for cross-lingual language model adaptation," *ACM*

Transactions on Asian Language Information Processing (TALIP), vol. 3, no. 2, pp. 94–112, 2004.

[30] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proceedings of the 7th International Conference on Database Theory (ICDT'99), LNCS, Springer-Verlag*, vol. 1540, January 1999, pp. 398–416.

[31] L. Haiquan, L. Jinyan, L. Wong, M. Feng, and Y. P. Tan, "Relative risk and odds ration: A data mining perspective," in *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Baltimore, Maryland, USA*, June 2005, pp. 368–377.

[32] P. Koehn, "PHARAOH: a beam search decoder for phrase-based statistical machine translation models," in *Proceedings of Meeting of the American Association for Machine Translation (AMTA)*, 2004, pp. 115–124.

[33] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech and Language*, vol. 8, pp. 1–38, 1994.

[34] F. J. Och and H. Ney, "Improved statistical alignment models," in *Association of Computational Linguistics*, Hongkong, China, October 2000, pp. 440–447.