Grammatical Inference and Textual Information Extraction

Knowledge Extraction from a fragmented textual data base

Alexandre S. Saidi

Ecole Centrale de Lyon
Mathematics and Computer Science Department
Laboratoire LIRIS (CNRS - FRE)
B.P. 163. 69134 Ecully - France
Alexandre.Saidi@ec-lyon.fr

ABSTRACT. *Text Mining tackles the task of searching useful knowledge (patterns) in a natural language document. Given the cost of a (full) morpho-syntactic analysis of a textual database, specially when the linguistic rules are not respected, most text mining techniques process without using the linguistic structure of those documents. In this Information Extraction framework, Grammatical Inference techniques can be used to extract the structure of a text (or of some of its sublanguage). This will allow an informed research of useful information in the textual data bases. In this paper, we present the contribution of the Grammatical Inference in the Text Mining field by reporting an Information Extraction process we applied to a seminar announcement corpus.*

RÉSUMÉ. *L'objectif de l'Extraction de Connaissances Textuelles (ECT) est la recherche de motifs intéressants dans les documents. La plupart des techniques employées dans ce domaine n'utilisent pas la structure linguistique, étant donnée le coût d'une analyse morpho-syntaxique (complète) et l'absence du respect des règles grammaticale (langue naturelle) dans ces textes. Dans ce contexte, l'Inférence Grammaticale peut être utilisée pour extraire la structure d'un texte (ou de ses sous-languages) afin de permettre une recherche informée dans une base de données textuelles. Dans cet article, nous présentons une contribution de l'Inférence Grammaticale dans le domaine d'ECT et exposons les éléments d'un processus d'extraction appliqué à un corpus d'annonces de séminaire.*

KEYWORDS: *Textual Data Mining, Grammatical Inference*

MOTS-CLÉS : *Extraction de Connaissances Textuelles, Inférence Grammaticale*

1. Introduction

The textual data bases constitute the major part of available information. Hence, significant research work concentrate on the Information Extraction (IE) from theses databasis.

In the Information Retreivial field, classification (and clustering) aims to categorise a textual data base into different *corpora*.

Though, the categorisation task is often done without a linguistic (syntactical) analysis of the contents of the latter.

Given a corpus, the information extraction applied to the texts by the techniques of *Text Mining* ([FAY 96], [HEA 97],[FEL 95], [TAN 99], [GRI 97], [AHO 98], [DIX 97], [TAN 94]) consists on the search for nonexplicit information in these texts. Text Mining tries to extract significative informations like the location or the date of a seminar in a conference announcement.

In a basic approach, this task would be difficult if one does not have any *a priori* structural information on the text[1].

Text Mining research field has been focused on since 1991 through MUC programs. However, it is still domain specific and time-consuming to build a new system or to adapt an existing one to a new domain. Although symbolic and statistical methods have been applied in some IE systems (e.g. [CAL 97], [KIM 95], [HUF 96]), not many ones have combined Grammatical Inference with (naive) statistical information. However, it may be noted this approach give similar results closed to the research work done on the *Named-Entity* (see e.g. [BIK 99], [PAL 97]).

In this paper, we report the current development state of an IE system which implements a Machine learning method using Grammatical Inference together with Bayesian values in order to extract information from textual data base.

Given the cost of a syntactical analysis, the search of the morpho-syntactic structure is not of a great interest in the text mining process based on key patterns. However, knowing the structure of the sub-language representing e.g. the *address* in an advertisement of an exposure on the city of *Lyon* which will take place in *Paris* may avoid concluding too quickly (and wrongly) on the place of the exposure upon the simple presence of *Lyon* city name.

On the other hand, in the case of unstructured (free) texts, the rules of linguistic grammars are seldom respected. These texts rather tend to transmit information with few words without using entities such as determinant, verbs and other punctuation.

However, techniques of Grammatical Inference (GI) ([Fu 82], [AHO 94], [MIC 94]) promise to be useful in this field by accompanying the process of *Text Mining* to ex-

1. by *structure*, we mean here any information on the physical, functional, logical and morpho-syntactic structure of the content of the document

ploit the morpho-syntactic structure of patterns (or of sub languages) with a minimum of information on the contents structure.

Techniques of GI attempt to induce the structures of a source data (flow of signs) in the form of production rules of a regular grammar[2]. The induced grammar being an element of a (language inclusion) lattice, the text mining then is concerned by an informed search within a graph of links and possible correlations between the patterns carrying required information and semantics.

In this paper, we consider the case of a textual base (free text) of seminar announcements where several formats are possible. As an example, let us consider the announcement below:

```
Seminar of the Institute of Nuclear physics of Lyon
problem of the mode conversions
Presented by Yves  Hake of Verdière
Fourier Institut of Grenoble
at 14:30 H - Room 27
Paul Dirac Building
```

The aim of the processing of these announcements is to extract various information such as the *Date* or the *Subject* covered in a seminar. Finals measurements like the research fields of a university (or a researcher) can then be extracted.

In the supervised process we consider, the text mining task applied to such a corpus could break up into several phases illustrated by the figure 1. In this process, the important templates slot fillers are already defined by an expert : one knows by advance which kind of information is contained (and sought) in the base[3]. The principal phases of processing illustrated by the figure 1 are briefly described below :

Preprocessing transformation and homogenisation of the characters, sentence extraction, suppression of some common words and punctuation in the text; etc.

Morphological Analysis extraction of lexemes and basic lexical classes; constitution of a dictionary/lexicon of terms and keywords in various slots (e.g. the *institute* for an announcement) starting from positive examples;

Partial syntactic analysis regrouping of the lexemes, constitution of simple and partial syntactic entities according to the structures of the sublanguages;

Grammatical Inference training of the grammar of sublanguages from positive examples together with the description of negative examples;

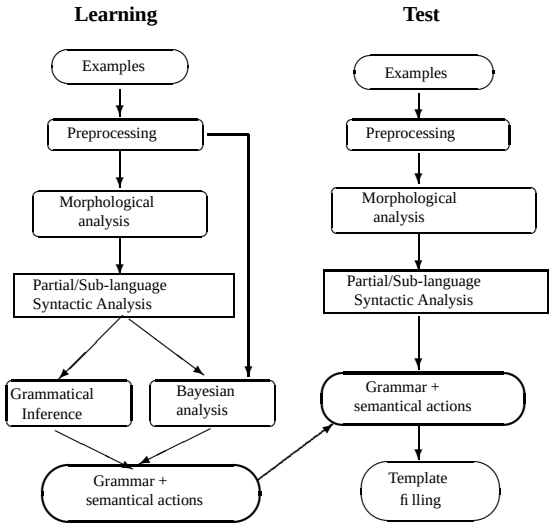
Statistic analysis (Bayesian) extraction of measurements, frequencies, weight and probabilities on the (couple of) patterns in the sample set;

---

2. in the Text Mining field, one is interested in the (so called *surface*) structure of the sublanguage almost governed by regular grammatical rules. Hence, we consider here the *regular* grammatical inference

3. a more interesting part of the IE process is to discover implicit information that the expert is not intended to know ! This is not the case here.

Figure 1. *Learning and test phases of the system*



Adding (semantical) actions to the induced grammar using the results of the Bayesian analysis[4]

Added to this process is a postprocessing and decision making phase (e.g. classification of unprocessed free zones, etc.) which will complete the whole process.

In the test phase , the candidate examples are analysed by the grammar induced in the training phase and information are extracted for the fillers of the template slots.

It is also appropriate to note that examples can be incomplete: for instance, the *Hour* may be missing within an announcement or it can be expressed in a different form (for example, by the "Friday afternoon" expression).

In this process, we are interested in the search of correlations between couples of patterns or a pattern (key) and information.

In the reminder of this paper, we will describe the interest of the Grammatical Inference (section 2) and the Bayesian analysis (section 3) regarded to the text mining task. We describe examples of announcements data base we considered and give some aspects of the realization in progress in the section 8.

4. a semantical action is a term from the *syntax directed* and the *Attributed Grammars* paradigm which denotes (no syntactic) actions based on the attribute values. Distinguished from the pure syntactical analysis, such actions take place in a production rule if the rule applied.

2. Grammatical Induction

In a text mining process, one prefers to avoid the syntactic analysis for several reasons :
- the cost and the complexity of this analysis,
- the very few use of the results of this analysis (the goal is not to correct errors or to translate the text),
- the texts may not follow the correct and complete syntax (of French in our case), etc.

Thus, in a seminar announcement, the subject is similar to a *noun group* but may not follow its rigorous syntax. The inference phase helps in this case to *effectively* retain the rules used in the examples. So the corresponding text mining process will rather be a syntax directed process.

Starting from a sample basis (positive examples and negative cases description, see the section 6.1), the Grammatical Inference (GI) induces production rules of a regular grammar[5] (a deterministic finite state automaton, DFA) of this sample set. In the test phase, the sentences presented to the grammar will be regarded as pertaining (or not) to the language generated by induced grammar.

The Grammatical Inference carries out a classification of the sentences (*accept* or *reject* means belonging or not to a given language) but, in its original form, it does not handle the semantics of these constructions. Hence, Bayesian measures will guide the process by predicting the slot to be submitted to the grammar. The IE process is then achieved with more precision and reliability (see also [FRE 97]).

3. Bayesian measurements

Several techniques of text mining use the Bayesian analysis which (even in its naive form) gives interesting results. In the method known as *naive Bayesian*, the document is presented as a vector of characteristics (e.g. various sections of an announcement). Other presentations such as *bag of words* consider the words present in the text in the form of a collection of words where any internal structure (physical, logical, morpho-syntactic or semantics) is inhibited.

In this approach, when a particular word is present, this presence is noted in the form of an integer value indicating the frequency of the word (0 for its absence) or the weight (see e.g. [SAL 87]) to reflect the importance of the word and its role.

Let us recall the naive Bayesian (conditional probability) formula . Given a hypothesis (e.g. to have such section of the class C in such context inside a message of seminar) and an example of announcement E over C, we have:

$$Pr(C/E) = \frac{Pr(E/C)\,.\,Pr(C)}{Pr(E)}$$

5. we note that the Context-Free grammar induction is an actual and active research fi led facing hard constraints making the general Context Free induction problem undecidable.

The idea is to express the weighted (balanced) probability of the membership of a pattern or a sublanguage within a class C according to the characteristic of the text E but also of those other texts classified and processed as such.

The method is known as *naive* by the assumption of the independence of the considered pattern (word or sublanguage) compared to any other occurrence of the pattern in the text.

To simplify, the probability of having a property (characteristic) will be simply the product of the weighted probabilities of the patterns of the same class computed during the training phase (see section 8 for an example).

Having defined the key patterns to recognise the various (but not all) fillers of an announcement during the training phase, we determine, during the analysis of a pattern p predicted to be the filler of a slot:
- If p is considered (the probability) for the appropriate slot, the remainder of that slot (e.g the $Date$ section) is subjected to the induced grammar that carries out a partial syntactic analysis and extracts the values for the filler;
- Otherwise, if the running pattern is unknown (or rejected), the values of probabilities can help, in the post processing phase (specially when other fillers are recognised), to retain the most probable remaining decision. The confidence coefficient given for each filler completes the information.

Also, the analysis of a filler (whose slot is not deduced with certainty) is committed on the most probable syntactical subtree. In this case, if no sufficient ($\geqslant 0.5$) probability is attached with the current hypothesis, the current unknown word may be ignored (in that step) and the process continues the analysis. The process uses however the backtracking in order to reexamine other possibilities (see the section 8 for the $Subject$ filler).

4. The Announcement Corpus

We considered a textual base of seminar announcements which may have several formats. The aim is to extract various information, for example, the $dates$ or the $subjects$ covered in the seminars. Below, there are several cases of a seminar announcement base we made up via the WEB.

Some of examples below are complete (examples 1, 2 and 3 where significant informations are in the announcement) whereas in examples 4, the $Address$ and the $Place$ miss while in example 5, the $Speaker$ is not given.

Note also that theses examples were originally in French. We give hereafter some of their English translation.

```
1- Seminar of the Institute of Nuclear physics of Lyon
   the problem of the mode conversions
   Presented by Yves Hake of Verdière
   Institute Fourier Grenoble
```

```
   at 14:30 H - Room 27 - 1st floor Paul Dirac Buld.
2- Seminar
   Conference Room, 1st floor,  IRIGM   Buld.
   Thursday April 11 2002, 14h30
   Yves Méheust, ENS Paris
   Flow of Stokes in a rough open fracture
3- Seminar in Toulouse:  Migration towards the free, Utopia or reality?
   By Romance Nicolas.
   05/16/2003 at 13:15
4- seminar:  security and Internet
   Paul-Andre Pays
   Tuesday, Jan 25 1996, 15:15
5- Arithmetic Seminar
   Thursday February 25 at 11h in  Kampé  de Fériet room,  M2  building
```

It may be noted that a Grammatical Inference engine applied to the entire announcement gives the sequential structure of an announcement (the sequence of various slots or sub-languages). This does not bring any relevant information : one obtains a grammar which confirms that the format of an announcement is free (no ordered sections). Instead, we use the grammatical inference in various sub-languages (e.g. the heading or the subject of an announcement) that may contain relevant information. As an example, the heading can contain a topic, a subject or an organiser which can possibly be enriched in the reminder of the announcement; the $Subject$ can add precise details to the $Topic$ of the seminar and vice versa.

5. Slots and Fillers

The following slots are defined for the seminar announcements processing.

$<Topic-Subject>$	the (general) $Topic$ and the $Subject$ of the seminar,
$<Org>$	the organiser, i.e. a university, laboratory, institute, school...,
$<Adr-Place>$	the address and/or the place of the seminar,
$<Speaker>$	the person who will make the talk ,
$<Org-Speaker>$	the organisation of the Speaker (e.g. the lab. of the Speaker),
$<Date>$	the date of the seminar,
$<Hour>$	the beginning hour (or the time range) of the seminar.

Note that an announcement starts with the $seminaire$ (seminar) keyword.

6. The Grammatical Inference Application

It is obvious that a simple textual search cannot be appropriate for extracting knowledge from our seminar announcements. Methods of knowledge extraction based on the Bayesian analysis allow to predict the position of an information in the text to-

gether with its average length (see e.g. [FRE 97]). This technique, based on the learning of the position of a section (e.g. the $<Subject>$) would not be appropriate here because the format of announcements are free and may be different (various slots of an announcement are not ordered). Also, an announcement can be incomplete. Thus, having the induced grammar of e.g. the $<Adr-Place>$ section will make it possible to analyse the content of that sub-language. In a first approach, once that the section $<Adr-Place>$ is located in the text of the announcement (predicted and then confirmed by the keywords), we will follow the production rules in order to analyse this section and then to extract information from it.

A brief presentation of the applied GI process is given below (see e.g. [SAI 03] for all details). Given the sample sets I_+ (positive examples) and I_- (negative examples descriptions), one deterministic finite state automaton (DFA) is associated to each example of $I = I_+ \cup I_-$. During the GI process, states of theses automata are merged according to the following predicate :

Predicate Congruence(r_1, r_2) adds constraints to the constraint store θ
Let r_1 and r_2 be the above rules (transitions) with $\alpha,\beta \in \Sigma$
$r_1 : [\alpha] \times s'_1 \rightarrow s_1$ $\quad$ $r_2 : [\beta] \times s'_2 \rightarrow s_2$

(1) if s_1 and s_2 are different final states in $(F_+ \times F_-)$ then set $[s_1] \neq [s_2]$.
(2) if $[\alpha] = [\beta]$ then set $([s'_1] = [s'_2] \Rightarrow [s_1] = [s_2])$ *(DFA condition)*
(3) if $[\alpha] \neq [\beta]$ then set $[s_1] \neq [s_2]$

Here, F_+ (resp. F_-) is the set of final states for the positive examples I_+ (resp. I_-). $[\alpha]$ denotes the equivalence class of $\alpha \in \Sigma$. Identically, $[s_i]$ denotes the equivalence class of the state s_i. The aim of this predicate is to compute the equivalence classes of the states and to create a constraint store θ on the final DFA. Then, given these constraints[6] (that describe a lattice of automata), we pick up a solution which minimises the number of the states, accepting *words* of the *language* of I_+ (a.k.a. L_+) and rejecting those of L_-.

Givens the rules r_1 and r_2 above, the application of the Congruence predicate can produces 3 different configurations (i.e. $[s'_1]=[s'_2] \wedge [s_1]=[s_2]$, $[s'_1]=[s'_2] \wedge [s_1] \neq [s_2]$, $[s'_1] \neq [s'_2] \wedge [s_1] \neq [s_2]$).

Although $[\alpha]=\alpha$ in its simplest form, we introduced the notion of equivalence class for the alphabet using the lexical class function $CL(\alpha)=[\alpha]$ where :

$[\alpha]=[\beta]$ iff $\alpha = \beta$ or $CL(\alpha)=CL(\beta)$, $\alpha,\beta \in \Sigma$.

6. the set θ contains constraints on integers, fbats and boolean expressions. The current system is realised in GNU-Prolog Constraint Logic Programming environment (http://www.inria.fr/)

For example, different city names are considered equivalent. Also, two (possibly different) organisations (university, research laboratory) are equivalent.

Note that if we consider α_1 (resp. β_1) as the *left context* of α_2 (resp. β_2) and α_3 (resp. β_3) as its *right context*, we will cover, in some extend, the case studied in [CAL 97] :

Applying the Congruence predicate to this case will produce 5 different configurations (depending on the equivalence classes of α_i,β_i) with various number of states in which the final induced minimal DFA has 4 states. Constraint store then will decide the final induced DFA considering all transitions and the negative examples.

It maybe noted that the Grammatical Induction upon only positive examples (I_+) tends to over-generalise L_+ (see e.g. [AHO 98]). Hence, the expert may express negative descriptions which are representative of the *words* that are to be rejected. For example, he may state that a *seminar announcement heading containing the Hour value in it* is to be rejected. The I_- set below contains some negative examples for an announcement heading.

6.1. *An Example of the GI Process*

As an example, the results of the grammatical inference on the **heading** of the announcements is the following :

I_+={'SDON', 'S:T', 'S', 'ST','SDT','SàV:T','SN:T', ...}
I_-={'Sa', 'SS', 'S::T', 'S::L', 'S::N','SDD','SOO', 'Sàà', 'Sa:', 'SD:', 'SaVV',...}

where :

S : the "séminaire" keyword (seminar in English),

D : $<Det>$, a determinant (e.g *'du', 'de la', 'des'*) like *'of'* or *'of the'* in English

T : <*Thème*> , an exposed $Topic - Subject$ (e.g. *Algorithm, Complexity*, etc.),

N :$<Nom>$, a Noun, e.g. name of a research laboratory ,

O : $<Org>$, an organisation name (e.g. *institute,laboratory, university, school*...)

V : Ville, name of a City, e.g. *Toulouse*

':' : this character,

'à' : this character (stands for *'at'* or *'in'*, ... in English).

The induced grammar accepts the super language L_+ of I_+ and rejects those of L_-. The final induced automaton accepts the language given below[7]. The rules which

7. Notation : $(X\|Y)$ means (X or Y) and the '.' (dot) denotes the monoid concatenation

reject unsuitable constructions (i.e. words in L_-) are not reported here for the sake of clarity. However, one may observe that a rejection takes place in the induced DFA when a derivation (upon a token) leads to a final failure state (F_-).

The language of the induced finite state automaton :

L+ = "Séminaire" . L1

L1 = ϵ

L1 = (':' || 'à') . L3

L1 = $< Nom >$. L5

L1 = <*Thème*> . L6

L3 = $< Org >$. L6

L3 = ($< Thme > \| < Ville >$) . L1

L5 = ':'. L3

L6 = ϵ

L6 = $< Nom >$. L1

Nota Bene: the induced grammar being in the form of a Definite Clause Grammar (DCG, a sort of logical grammar), predicates expressing the constraints and other actions are then added to its rules (see the example below). For example, while recognising (in their context) :

- a <*Thème*> may contain a part of the $Subject$; then the value corresponding to the $Subject$ will be added to the $< Sujet >$ filler;
- for a $< Ville >$, the corresponding value will be added to $< ADR - Place >$ filler[8].

Other possible adjustments are achieved during the postprocessing phase.

6.2. *An example : the* $Date$ *automaton and actions*

As an example, some of the induced DCG rules (together with the actions) for the expression of the $< Date >$ filler are given below. The absence of any part of a $Date$ is denoted by an empty rule (ϵ-rule) not reported here[9].

$< \mathbf{Date} >$:: $[\text{"}date\text{"}][\text{" : "}][\text{"}le\text{"}][< Day - name >][< Mid - day >][\text{"}le\text{"}]$
$< Day > [< Sep >] < Month > [< Sep >] < Year >$.

$< \mathbf{Mid - day} >$:: $< Word >$ {\$1$\in$ {"$matin$"} ; $add(part\_of\_Heure, \text{"}8h - 12h\text{"}, 100)$
OR \$1 $\in$ {"$aprs - midi$"} ; $add(part\_of\_Heure, \text{"}14h - 18h\text{"}, 100)$}.

$< \mathbf{Month} >$:: $< Number >$ {\$1 $\in$ {1..12} ; $add(part\_of\_Date, \$1, 100)$}

8. in the case and the presence of $< Adr - Place >$ context.
9. $[xxx]$ means optional xxx; \$k : the value of the kth literal (as in $yacc$ compiler compiler).

$$\| < Word > \{\$1 \in \{"jan"..."dec"\}\ ; add(part\_of\_Date, \$1, 100)\}\ .$$

$<\mathbf{Day-name}>:: < Word > \{\$1 \in \{"lun"..."sam"\}\ ; add(part\_of\_Date, \$1, 100)\}.$

$< \mathbf{Day} >:: < Number > \{\$1 \in \{1..31\}; add(part\_of\_Date, \$1, 100)\}.$

$< \mathbf{Year} >:: < Number > \{\$1 \geq 1990\}; add(part\_of\_Date, \$1,100)\}.$

$< \mathbf{Sep} >:: \quad '/' \| ' :' \| ' -' \| ... \qquad --a\ separator$

Nota Bene: the value 100 (parameter of the predicate add) indicates the confidence coefficient of the value assigned to the filler. Here, the case of $< Date >$ is relatively simple and follows a relatively known format. We may however notice that the presence of "matin/après-midi" (AM/PM in English) of the $< Date >$ will complete the $< Hour >$ slot filler.

7. Bayesian Analyse and Measurements

Considering a sample set of 100 examples, the percentage values are given below ($'Org - Sp'$ abbreviates $'Organiser - Speaker'$, $Pres$ stands for $Present$, Sub for $Subject$ and Spk for $Speaker$) :

Table 1. *Frequency table of various sections in the seminar announcement data base*

	Sub	Org	Date	Hour	Place	Adr	Spk	Org-Sp	End	Pres
Annonce	14	9	41	4	14	14	9	0	0	100
Sub	0	0	4	0	9	0	23	0	9	45
Org	0	0	4	0	4	0	0	0	0	9
Date	0	0	0	77	9	0	4	0	4	95
Hour	9	0	4	0	41	0	18	0	14	86
Place	4	0	23	0	0	36	4	0	23	91
Adr	4	0	9	0	4	0	0	0	32	50
Spk	4	0	9	0	4	0	0	36	4	59
Org-Sp	14	0	0	4	4	0	0	0	14	36

We add to this table two other values : 77% of the announcements contain a $Topic$ in their heading, and 18% of the headings contain an indication on the organiser (Org). The *present (pres)* column indicates that e.g. the $Subject$ is present only in 45% of the announcements. The cells containing 0% are of a particular interest because they give indications on the cases that do not occur. For example, $< Org - Sp >$ never follows the heading of an announcement.

As an example, we apply the conditional probability to the section $Subject$ of the example (1) of the section 4 where the slot of the second line is not determined. This example shows how the post-processing will help deciding the slots filler.

Given the table 1 above, the probability so that the unknown section (2nd line of this example) in this announcement be a $Subject$ (surrounded by the *Heading* and

the *Speaker*) is 12%. However, this announcement does not comprise a $Topic - Subject$ in its heading and, the $< Speaker >$ is the successor of a $Subject$ in $\frac{23}{45}$ cases. Therefor, the filler is predicted at 23% (weighted 51%) to be the $Subject$.

More precisely, in the post processing phase, we will gradually eliminate the recognisable sections $Place$ and Adr (14%) to retain only the $Subject$. Hence, the best of the probabilities for determining Y is chosen from those of X $\rightarrow$ **Y** (eg. announcement followed by the $Subject$: 14%) and **Y** $\rightarrow Z$ (the $Speaker$ which follows the $Subject$ in $\frac{23}{45}$ cases and its weighted value is 51%).

Note that the strongest probability of the section which follows the heading is the $Date$ section. However, one can recognise a $Date$ by the keywords in the induced grammar.

The system is parametered by the depth of the Morpho-Syntactic analysis. Thus, if needed, the (partial) linguistic class from this filler can be extracted giving a (partial) Noun Group (even without the initial determinant).

8. The Realisation

This section describes briefly the state of the experimentation on the basis of seminar announcements. We used several tools, in particular, the morphological analyser **Unitex** ([UNI 03]) and the dictionaries of ABU ([ABU 03]).

Following the preprocessing phase, the system learns various rules from the sublanguages (sections of an announcement corresponding to the template fillers) on the training examples set. Then we produce a set of production rules for the partial morpho-syntactic analysis of this corpus.

The morphological analysis of a portion of an announcement can give multiple results as mentioned in many examples[10]. In order to produce the induced grammar, the expert will use some heuristics in order to eliminate the useless combinations and to reduce the linguistic analysis phase to its bare minimum and to limit more supervision. The rejection of the combinations requires a partial syntactic analysis. However, in the training phase, one can retain the rules able to eliminate these combinations (e.g. keyword presence).

Given that the seminar announcements do not follow the linguistic rules, we do not decide completely about a section without having used all the knowledge we will extract. For example, the presence (alone) of the word *Lyon* cannot conclude on a $< Adr - Place >$. Also, a determinant ($< Det >$) misses from the *"Écoulement de Stokes dans une fracture rugueuse ouverte"* (in English : "Stokes Flow in an open rough fracture") in the example (2) of the section 4. We will consider announcements as they are provided and have only some elementary information (e.g. the $< Subject >$ part has a syntax close to a *Noun group*, the $< Date >$ and the $< Hour >$ use numbers, etc.)

10. as from the English example *I spring in the spring on the spring like a spring*

Some details of the analysis phase of the example (1) of the section 4 follows.

```
1- Séminaire de l'Institut de Physique Nucléaire de Lyon
   Le problème des conversions de modes
   Présenté par  Yves Colin de Verdière
   Institut Fourier Grenoble
   à 14:30 H - Salle 27 - Rez de chaussée - Bât. Paul Dirac
```

Nota Bene : $part\_of\_X$ means part of the filler of the slot X. If the decision cannot be made, the character string in hand is considered as $free$ standing by the post processing.

Hereafter, the symbols K stands for a *Keyword* and Gn for a *Noun Group* (in French), 'w' stand for word and 'ws' for sequence of words.

```
INIT :  for I in {Annonce,Date,Sujet,Heure,Lieu,Adr,Orateur,Org,Org-Orateur}
        do  i.present=false; i.val=Date.val=empty;
            part_of_i.val=empty;  part_of_i.cf=0;
        done
```

Regrouping and useless words elimination on the morphological outputs gives:

```
A- k("Séminaire", N:ms) w("de l'",Det+Dind:fs:ms)   k("Institut", N:ms)
ws("de Physique Nucléaire de Lyon")
==> {Annonce.present=true; part_of_org.val=$3.$4; part_of_org.cf=100}
```

Here, the value 100 is an indication of the confidence on the fact that the string can belong to Org.
Nota Bene : the word *institute* is a keyword for Org. Here, $'Topic - Subject'$ is absent in the heading.

```
B- w("Le problème des conversions de modes", GN);
==> {push_back(Free, \$1);}
```

N.B. there is no keyword present; hence, the string is considered as $free$.

```
C-k("pésenté",V:Kms) k("par",Prep)   ws("Yves Colin de Verdière")
==> {Speaker=true; Speaker.val=\$3; }

D- k("Institut",n:ms) w("Fourier", N+Prenom:ms) w("Grenoble", N)
==> {part_of_Org.val=\$*; part_of_Org.cf=9;
     part_of_Org_Sp.val=\$*; part_of_Org_Sp.cf=61; }
```

The value 61 is an indication of the (weighted) confidence on the fact that the string can belong to Org.

```
E- k("à",Prep) ws("14", Nb) key(":") ws("30", Nb) k("H") k("Salle",N:fs)
ws("27", Nb) k("Rez de chaussée",N)   k("Bât.",N) ws("Paul Dirac", N)
==> {Hour.present=true; Hour.val=$2.$3.$4.$5;
     Place.present=true; Place.val=\$6-;}   -- the reminder of the line
```

In the post processing phase, the statistical measurements enable us to conclude that the line 2 of this announcement is not one of {*Date, Heure, Lieu, Adr, Org, Org-Orateur, Orateur*}. Hence, as seen before, it can be part of the $Subject$ (51%).

8.1. *Postprocessing and filling the slots*

Having slots of the template partially filled, we reconsider the calculus to decide if any filler can be provided, completed or left vacuum. At this stage, the templates will contain the following values (coefficients added beside if not 100):

```
part_of_Org         = "Institut de Physique Nucléaire de Lyon"
part-of-Org         = "Institut Fourier Grenoble"             (9)
part_of_subject     = "Le problème des conversions de modes"  (51)
Speaker             = "Yves Colin de Verdière"
part_of_Org_Sp      = "Institut Fourier Grenoble"             (61)
Hour                = "14:30 H"
Place               = "Salle 27 - Rez de chaussée - Bât. Paul Dirac"
Subject             = "Le problème des conversions de modes"  (51)
Address             = "Institut de Physique Nucléaire de Lyon"
Date                = ""
Org                 = "Institut Fourier Grenoble"             (9)
```

9. Performances Evaluation

We considered about 100 examples to measure the performance. We applied then a ten-fold cross validation (for the training and test phases example generation) and observed that the results were not significantly changed for more examples.

Metric definitions : $Precision$ and $Recall$ measures are the computed percentages of the well known $soundness$ et $completeness$ properties (in the LP community). Using a given a corpus of announcements, evaluation metrics are based on the filler presence and prediction. In this case, $Relevant \cap Retrieved$ will denote the number of present slot values which are accurately detected and assigned to the fillers (i.e. correct and computed). The *Relevant* value is then the value of effectively relevant (just) present slot values and the *Retrieved* value is whatever slot that has been claimed (predicted) to be correctly assigned. Hence, we have :

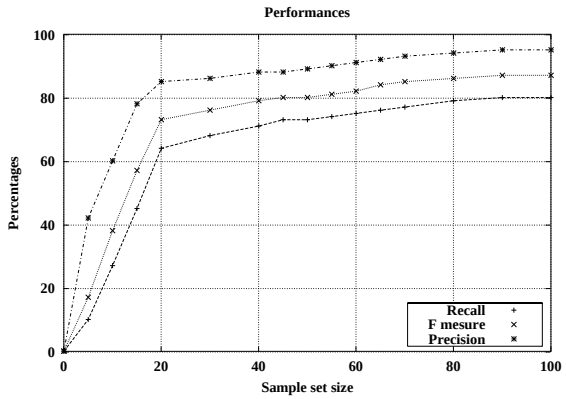
$$Precision = \frac{Number\ of\ the\ present\ and\ Correctly\ assigned\ slots}{Number\ of\ slot\ claimed\ to\ be\ identified}$$

$$Recall = \frac{Number\ of\ present\ and\ Correctly\ assigned\ slots}{Number\ of\ correct\ present\ slots}$$

Also, a harmonic measure called F-measure (see e.g. [LEH 91]) is used to give the mean of the above values : $F - mesure = \frac{Precision \times Recall}{\frac{1}{2}(Precision + Recall)}$

The diagram of the figure 2 shows the performance percentages we obtained. As one may observe for the seminar announcements corpus, it is not surprising to have high performance values (95% and 80%) given the intended slots and the relative low risk

Figure 2. *Performance evaluation*



of error. The system is quite domain specific and may even be enhanced. Hence, without appropriate modification, it can not be applied as such to other kind of corpus.

10. The Related Work

Several textual IE system have been proposed since the focus on researches started by MUC program of DARPA (e.g. [DAR 92], [LEH 91]).

The use of patters dictionnary is common to many systems. Some uses clustering to create patterns by generalising those identified by an expert(see e.g. [SOD 95]). The dictionnary we use in the present work contains basically the keyword (and their lexical class) that are then used during the analysis.

Syntactic informations can be used as in Autoslog ([RIL 93], [RIL 96]) which uses a set of general syntactic patterns validated by an expert. Among theses systems, some uses advanced syntactic analysis to identify the relationship between the syntactic elements and the linguistic entities (e.g. in [HUF 96]). This analysis is costly (when the sematic information is not used) and may limit the system specially if linguistic rules are not respected like in our seminar announcement examples.

In many IE systems, human interaction is highly required through different phases of training. Machine Learning techniques like decision trees are used ([McC 95]) to extract coreferences using the annotated coreference examples.

Among these systems, the current work is closed to PAPIER system ([CAL 97]). RAPIER is an ILP system that takes pairs of documents and filled templates and induces rules that directly extract fillers for the slots in the template. This system uses constraints on words and part-of-speech tags surrounding the fillers' left and right contexts. In some extend, our system can be seen from this point of view since,

as mentioned in the GI section (2), our grammatical Inference engine implements this technique implicitely. Also these results should be compared with those of the *Named-Entity* research work (see e.g. [BIK 99], [PAL 97]) and aims to learn *names* by identifying all named locations, persons, organisations dates and so on.

11. Conclusions

We presented an IE system that fills slots of a template associated to seminar announcements using Grammatical Inference and Bayesian measurements.

Once the template are slots filled, current techniques of Data Mining (see e.g. [AHO 98]) can then be applied to the data base made up since the resulting values of the slots describe simply a relational database scheme. One current use of the system is to extract information like the research field of universities, laboratories or researchers in order to guide PHD students in their researches.

This is a work in progress but the performance results are encouraging to continue the project. We plan to first enhance and then extend the system to other corpora like job announcements and marine weather announcements in order to establish statistics on marine catastrophes and previsions. The system will be integrated to a classical Datamining engine in order to establish important information on marine events.

12. References

[ABU 03] ABU. Divers Dictionnaires *Association de Bibliophiles Universels*, http://abu.cnam.fr/

[AHO 94] H. Ahohen, H. Mannila. *Forming Grammars for structured documents*. RR. U. Helsinki. 1994.

[AHO 98] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. *Applying data mining techniques for descriptive phrase extraction in digital documents*. In Proc. Advances in Digital Libraries (ADL98), Santa Barbara, CA, 1998.

[BIK 99] , D. Bikel, R. Schwartz, R. Weischedel *An Algorithm that Learns What's in a Name*, 1999

[CAL 97] M.E. Califf, R.J. Mooney, *Relational Learning of Pattern-Match Rules for I.E.*, Proc. of AAAI Symposium on Applying Machine Learning to Discourse Processing, 1997.

[DAR 92] ed., *Proc. of 4th and 5th DARPA Message Understanding Evaluation and Conference*. Morgan Kaufman. 1992, 1993.

[DIX 97] M. Dixon, *An Overview of Document Mining Technology*, 1997, http://citeseer.nj.nec.com/dixon97overview.html.

[FAY 96] U. fayyad & all *From DataMining to Knowledge discovery : An overview*. in Advances in Knowledge Dicovery and DataMining, MIT Press, Cambridge, Mass 1996.

[FEL 95] R. feldman, I. Dagan *Knowledge Discovery in textuel databases (KDT)*. Proc. of the 1st Int. Conf. on Knowledge Discovery and DataMining (KDD-95), Montreal, Ca, AAAI Press, 1995.

[FRE 97] D. Freitag. *Using Grammatical Inferenec to Improve Precision in*. ICML'97. 1997.

[Fu 82] H.S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice Hall, N.Y. 1982.

[GRI 97] R. Grishman, *Information Extraction: Techniques and Challenges*, http://citeseer.nj.nec.com/grishman97information.html.

[HEA 97] M.A. Hearst, *Text Data Mining : Issues, Techniques and Relationship to Information Axxess*, Presentation Notes for UW/MS Workshop on data mining, July 1997.

[HON 01] Theodore W. Hong, Keith L. Clark, *Using Grammatical Inference to Automate Information Extraction from the Web*, Lecture Notes in Computer Science - 2168, 2001.

[HUF 96] S. B. Huffman, *Learning information extraction from examples*. in Wermter, Riloff, Scheler ed. Berlin, 1996.

[KIM 95] J. T. Kim, D.I. Moldovan, *Acquisition of linguistic patterns for KB information extraction*. IEEE 7(5), 1995

[LEH 91] W. Lehnert, B. Sundheim, *A performance evaluation of text-analysis technologies*, AI Magazine 12(3), 1991.

[LEW 94] D.D. Lewis, W.A/ Gale , *A sequential algorithm for training text classifier*, Proc. of the 7th Int. Conf. on Research and Development in Information Retrievial, 1994.

[McC 95] J. McCartht, W. Lehnert, *Using decision trees for coreference resolution.*, in Proc. of 4th Int. Conf. on IA, 1995.

[MIC 94] L. Miclet. *Grammatical Inference*, Syntactic and Structural Pattern Recognition. H. Bunk and SanFeliu eds. World Scientifi c.

[MOO 95] E. Califf, R.J. Mooney, *Induction of first order decision lists : Results on learning the past tense of English verbs*, Journal of AI Research., 1995

[PAL 97] D.D. Palmer, D.S. Day, *A Statistical Profile of the Named Entity Task*, in proc. of th 5th. Conf. on Applied Natural Language Processing, Washington D.C., 1997 ACL.

[RIL 93] E. Riloff, *Atomatically constructing a dictionnart for information extraction tasks*. in Proc. of 11th National Conf. on AI, 1993.

[RIL 96] E. Riloff, *Atomatically generating extraction patterns from untagged text*.. in Proc. of 13th National Conf. on AI, 1996.

[SAI 03] A. Saidi. *A Constraint satisfaction framework for Documents Recognition*, Workshop on Multimedia Discovery and Mining , ECML-PKDD 2003.

[SAL 87] G. Salton, C. Buckley , *Term Weighting approaches in automatic Text Retrievial*, Technical report 87-881, Departement of CS, Cornell University, 1987.

[SOD 95] S. Soderland, D. Fisher, J. Aseltine and W. Lehnert, *Crystal : inducing a conceptual dictionnary*, in Proc. of the 14th Int. Conf. on AI, 1995

[SOD 96] S. Soderland, D. Fisher, J. Aseltine and W. Lehnert, Srystal : *Issues in inductive learning of domain specific text extraction rules*, in LNCS in AI, 1996.

[TAN 94] Y. Yan Tang, C. De Yan, C. Y. Suen *Document processing for Automatic Knowledge Acquisition*. IEEE transactions on Knowledge and Data Engineering. 6(1). 1994.

[TAN 99] A.H. Tan. *Text Mining : The state of the art and the Challenges*. Proc. PAKDD'99 workshop on Knowledge Discovery from Advanced Databases. Beijing, 1999.

[UNI 03] Le sytème Unitex développé par S. Paumier *http://www-igm.univ-mlv.fr*. Voir aussi *http://www-igm.univ-mlv.fr/ unitex/*.