From the texts to the concepts they contain: a chain of linguistic treatments

Ahmed Amrani* Jérôme Azé† Thomas Heitz†
Yves Kodratoff† Mathieu Roche†

Abstract

The text-mining system we are building deals with the specific problem of identifying the instances of relevant concepts present in the texts. Our system relies therefore on interactions between a field expert and the various linguistic modules we use, often adapted from existing ones, such as Brill's tagger or CMU's Link parser. We have developed learning procedures adapted to various steps of the linguistic treatment, mainly for grammatical tagging, terminology, and concept learning.

Our interaction with the expert differs from classical supervised learning, in that the expert is not simply a resource who is only able to provide examples, and unable to provide the formalized knowledge underlying these examples. We are developing specific programming languages which enable the field expert to intervene directly in some of the linguistic tasks.

Our approach is thus devoted to help one expert in one field to detect the concepts relevant for his/her field, using a large amount of texts. Our approach is made of two steps. The first one is an automatic approach that finds relevant and novel sentences in the texts. The second one is based on the expert's knowledge and finds more specific relevant sentences.

Working on 50 different domains without an expert has been a challenge in itself, and explains our relatively poor results for the first Novelty task.

1 Introduction

In this paper, we present our approach and experiments relative to the Novelty Track of TREC-2004. We only answered the first two tasks of this track. The data available for this track was made of two sets of newspaper articles divided into **Opinion** and **Event** domains. For each of these domains, the set of papers dealt with 25 different topics, each containing at least 25 documents relevant to the topic. We had thus to deal with 50 different topics while our approach is supposed to provide help for one topic only .

The first task was made of two sub-tasks:

- determine the relevant sentences in the documents of each topic,
- find the new sentences among the relevant sentences previously determined.

The second task in which we took part was finding the novel sentences given all the relevant sentences for each of the 50 topics.

In order to fulfill these two tasks, we used the text mining system developed in our team. This system is a complete chain of text mining starting with an initial corpus, building from the

*ESIEA Recherche, 9 rue Vésale - 75005 Paris - France - amrani@esiea.fr
†LRI - Université Paris-Sud - 91405 Orsay Cedex - France - {aze,heitz,yk,roche}@lri.fr

corpus ontologies specific to the corpus, and ending with sets of linguistic instances of the concepts defined by the end-user.

Since this chain of treatments has been used for answering the two tasks we participated in, we will first present this chain and we will detail thereafter our specific approach for the two tasks.

2 Our global chain of text mining

The whole chain of treatments we deal with is shown on figure 1.

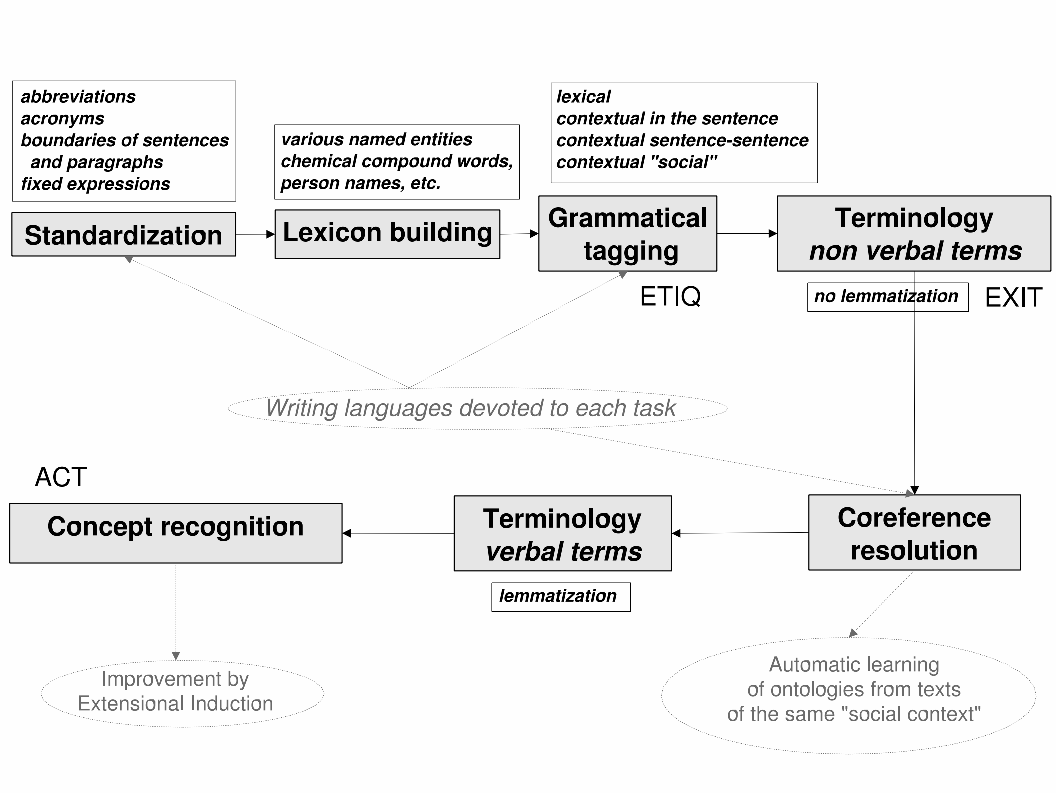

Figure 1: Our global chain of text handling in view of text mining.

2.1 Text retrieval

The first step of any chain of treatments of texts is to build an homogeneous corpus relative to a given domain. In the TREC competition, we did not have to deal with this step since the corpus is provided by the organizers.

The corpus is supposed to be relevant to the studied domain. One particular point for TREC Novelty 2004 is that some irrelevant documents had been inserted in some topics. The topics that contains irrelevant documents can easily be identified since the set of documents then contains more than 25 documents.

Although we can directly identify these topics, we cannot easily identify irrelevant documents in them. Thus, we decided to ignore them since our system is not devoted to the clustering of relevant documents. We are presently building a unit that will solve this problem in two steps: 1. If the total number of documents is N, then compute two clusters of N-25 and 25 elements using Latent Semantic Analysis, LSA [10]. 2. Define a novelty distance based on the classical

$TF \times IDF$ (as we did for task 2 with some success). 'Shake' the LSA clustering in order to minimize novelty in the new clusters.

2.2 Standardization

Standardization is also often called text cleaning. This is obviously specific to each domain, and only the expert is able to decide what is to be kept, transformed or deleted. For TREC Novelty, we first built a set of abbreviations. This set is useful to unfold the abbreviations found in the texts, thus reducing their ambiguity. For instance, if a text contains the words 'United States', 'United States of America' and the abbreviations 'USA' and 'U.S.', we unfold all of them into one common form, for example 'United_States_of_America'. See table 1 for examples of abbreviations and their developed forms.

Abbrev.	Developed form
Adm	Administrator
Adm.	Administrator
Admin	Administrator
Atty.	Attorney
CEO	Chief_Executive_Officer
CEOs	Chief_Executive_Officers
Capt.	Captain
Cmdr.	Commander
Col.	Colonel
Dr.	Doctor
Dr	Doctor
Drs.	Doctors
Esq.	Esquire
Gen.	General
Gov.	Governor
Hon.	Honorable
Jnr.	Junior
Jr.	Junior
Lieut.	Lieutenant

Table 1: Examples of abbreviations and their developed forms

Another point is connected with the first word of sentences that should be put in lower case, if belonging to common words lexicon, in order not to be confused with proper names. One of our errors, observed after competition time, was to apply this lowercase transformation to all uppercases words. In fact, except first word of sentences, it is only necessary to apply it to all uppercases and capitalized sentences or citations between apostrophes.

Many other transformations have been done in order to reduce the ambiguity. Some of them are: deletion of formatting expressions, development of elisions, formatting of numbers, processing of hyphenations, etc.

The task of putting one sentence by line was already performed by the TREC team in order to enable the comparison of results of different competitor teams. However, it seems that the problem of abbreviations has not been totally taken into account because some sentences boundaries happened at an abbreviation dot. For example:

*<s docid="APW19981018.0674" num="19"> "***Gen.***</s>*

<s docid="APW19981018.0674" num="20"> Pinochet has committed crimes humanity _ many atrocities, torture, murder," Vincete Allegria, an exile who fled Pinochet's regime, told re-

porters outside the clinic.</s>

Another problem takes place with the hyphenations that should be concatenated, for example:

< s docid="NYT19990527.0101" num="20"> The bill would allow workers to receive their share of the Social Security surplus each year, **dir-***</s>*
<s docid="NYT19990527.0101" num="21"> ectly deposited in individual accounts.</s>

Our system could detect these problems and treat them correctly.

2.3 Building the lexicon

In order to build a lexicon, we used external resources such as WordNet 2.0 [11], the Moby Part of Speech II and the Webster dictionary 1913. Our lexicon is about 300,000 words and has been used to perform a first grammatical tagging.

We built a proper-name lexicon with the help of a list of known proper names and the hypothesis that capitalized words around them are also proper names. The establishment of this lexicon was necessary for a further coreference resolution step.

A lexicon of locations was also used to keep them in the summary obtained at the last step of our chain of text handling.

2.4 Tagging of texts

The next step is relative to the grammatical tagging of each word in the texts. We have chosen to use a rule-based tagger, Brill's tagger [2]. Although Brill's tagger has learned its tagging rules from the celebrated Wall Street newspapers corpus, most of the texts need a serious revisiting of the rules and lexicon provided by Brill's basic version. These changes can be done in two ways. The first one is correcting 'by hand' the tags of incorrectly tagged words and train Brill's tagger on the modified texts. This approach is very expensive in human time, and we cannot even ensure that the correct tagging will be subsequently achieved by Brill's tagger.

The second one, the one we used, is based on two external resources, a tagging language that allows a human expert to write, besides some basic tagging rules, also complex tagging rules that can not be learned by Brill's tagger. The language we developed was customized to be more powerful than the rules generated by Brill's tagger. In this tagger, the context examined to build the rules goes either from -3 words around the word to be tagged, or to +3 words around it. The analysis of the badly tagged words reveals that this context is too poor to express the correct tagging rules, so we introduced in our language the possibility to use the whole sentence as context.

Here is an example that could have been learned easily by Brill:

if (0,further,) (+1,, or(VB,VBP,VBD)) then (0,,RB)

meaning: if the word "further" is followed by a verb (VB, VBP or VBD), then tag it as an adverb (RB).

Here is a (still simple) example outside the search space of Brill since it uses a context defined on both sides of the word to be tagged, and the context is variable

if (?1,,DT) (*1,,or(RB,JJ)) (0,,or(VBN,VBG)) (+1,,@NNAll) then (0,,JJ)

meaning: if a word **x** is tagged by VBN or VBG, and it is followed by any kind of noun (@NNAll), and it follows immediately a determinant (DT), with a possible adverb (RB) or adjective (JJ) between the determinant (DT) and **x**, then tag **x** as an JJ (it tags adjectives and premodifiers).

We also introduced specialized ontologies of words that control a special tagging. For instance, the nouns designating a span of time or a place, such as "afternoon" or "region" tend to determine a tag determinant (DT) for a "that" placed just before them. We built several of these specialized categories and used them in our rules.

When quite complex contexts are necessary, our language is devised so that they can be written with relative ease. For instance, many tagging rules take place whenever a possibly large and unknown number of adverbs and adjectives are piled up in front of a noun: this is easily taken into account by formula like

(*1,,RB) (*1,,JJ) (*1,,JJ) ($n,,NN)

that finds the variable place in the sentence, $n, for a noun (NN) following possibly one adverb (RB) and one or two adjectives (JJ).

During the competition, we had no time to learn automatically better rules from the ones we designated ourselves. The competition helped us to realize how much these rules are interesting in practice. This is why we are presently, after (but also because of) the competition, building an active learning system that proposes improvements to the human produced tagging rules. One of the important technical point is that relational learning being very lengthy, we shortcut this difficulty by using features that are the same as, or similar to, the ones used by the human.

The other resource is a software, ETIQ [1], that enables the expert to see the context of each word and then determine if the tag associated to a particular word is correct or not.

For the Novelty Track, our expert has produced a set of some 450 rules in order to improve the quality of the tagging. Some of these rules are specific to the corpus studied, but most of them are complex rules that are valid for any English text.

The tags we used reproduce almost exactly, at least during tagging proper names, the ones used by Brill. At further stages we introduced more semantic tags as will be explained later. The only significant difference we introduced, more or less in order to take into account the journalistic style of the corpus, is the difference between proper names of individuals, we called NPP, the countries we called NPL and the proper names of organizations we called NP. This difference is extremely important for newspapers since coreferences to individuals are extremely different to the ones to NPL and NP. In particular, journalists seem to be very skillful at using the gender of a person in order to build anaphora impossible to solve without knowing this gender.

2.5 Terminology

Once a correct tagging is obtained, we can make a further step by finding terms often used in the texts. The tagging is an important step since, as we shall see, we used grammatical patterns to extract the collocations (i.e., sets of words forms or lexemes happening in succession with an 'unexpected' high probability) and we need to have a correct tagging if we want to obtain significant collocations.

We used simple syntactical patterns to extract the collocations. These patterns are Adjective-Noun, Noun-Noun, Noun-Preposition-Noun, etc. Significant collocations are collocations that represent an occurrence of a concept in a text. Collocations can be ordered using statistical criteria [4, 14, 6, 12]. We used a software built in our team, EXIT, [13] (see figure 2).

2.5.1 Statistical measurements

EXIT (EXtract Iteratively complex Terms) uses different statistical measures to order the collocations

We present here three basic measurements made on the TREC corpus. Mutual information [3] is based on the dependence of the two words x and y composing the collocations. Let $P(x, y)$

represent the probability of observing x and y together in this order. $P(x)$ (resp. $P(y)$) represents the probability of observing x (resp. y). The mutual information $MI(x, y)$ is defined by (1):

$$MI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

The second measure we used is the loglikelihood [5]. Before defining it, we must give several definitions. Consider a contingency table associated with each couple containing the words (x_i, x_j).

	x_j	$x_{j'}$ with $j' \neq j$
x_i	a	b
$x_{i'}$ with $i' \neq i$	c	d

The values a, b, c and d define occurrences of a couple where $a + b + c + d = N$ is the number of occurrences of all couples.

The loglikelihood $L(x_i, x_j)$ is defined by (2) :

$$\begin{aligned} L(x_i, x_j) \quad = \quad & a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) \\ & -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \\ & +N\log(N) \end{aligned} \tag{2}$$

These two measurements provide the likelihood score of the couples (x, y). They differ mainly because mutual information takes only into account the independence of x and y while loglikelihood takes also into account the number of collocations without x nor y.

We noticed that for all proportions of terms found, loglikelihood always gives the higher proportion of correct terms [12]. This confirms Daille's assertion that loglikelihood is the best measure for the detection of terms [4].

Another widely used ranking function, referred to as Occ_L, is defined by ranking terms according to their number of occurrences, and breaking the ties by the likelihood.

2.5.2 Iterative Method

In our approach, the words, composing binary terms, are linked by a hyphen in modified texts. For example, when finding the term "family planning," words "family" and "planning" are linked by a hyphen to form "family-planning." One originality of our approach is the iterative procedure used to extract the terms, that is, to extract the terminology during the n-th iteration, we use the terms found at the $(n-1)$th one in order to build new terms, more specific than the ones obtained at the n-th iteration. For instance, in the corpus provided by TREC, during the first iteration, we extracted the binary term "family-planning". Then, during the second iteration, we extracted the term "family-planning-organization".

2.5.3 Parameters added

In order to improve the precision, we modified the scores of the terms by parameters [13]. For example, we added a parameter to favor the collocations found in several different texts rather than in a single one. Another parameter we used is a simple pruning, fixing a minimum number of occurrences of a collocation, whatever its loglikelihood, in order to accept it as a term.

For the TREC Novelty Track, we extracted terms linked to Opinions or Event. These terms have been associated to concepts using ACT [9, 12] (see section 2.7).

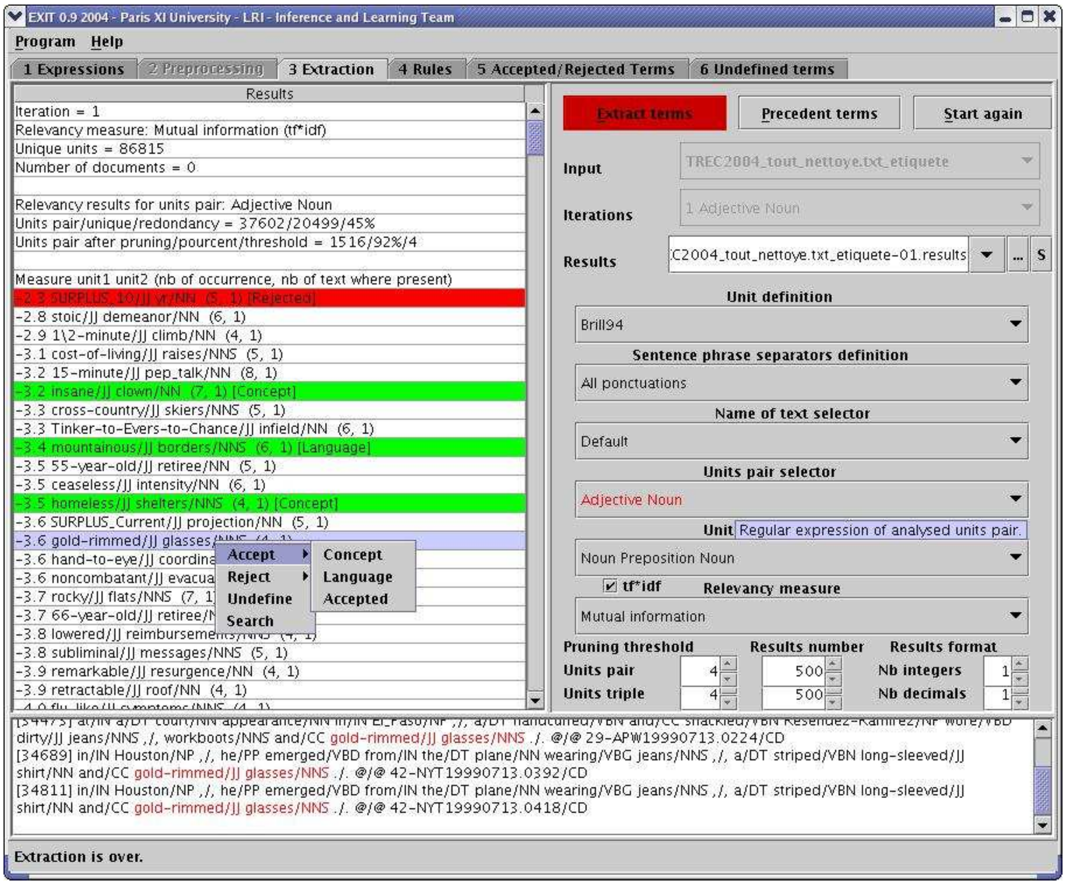

Figure 2: Screenshot of EXIT. This module allows the expert to iteratively extract terms as shown in the list of the left panel.

We also extracted collocations from the description of topics. Then, using the software FASTR [8], we looked for variations of terms in the texts associated to each topic. For example, in the description of the topic N88, the collocation “Vieques Island” have been extracted. This collocation appears only 7 times in the texts associated to this topic. But it appears 20 times in another form, which is a variation of “Vieques Island”: “Island of Vieques”.

Terms and their variants occurring both in the topic’s description and in the associated texts have been used to compute a score for each sentence of the texts, as explained later in section 3.1.1.

2.6 Coreference resolution

During the competition, we dealt only with anaphora referring to human beings.

One of the main difficulty we found in the TREC texts is the coreference problem. In other words, how can we automatically determine the person referred to by the words “he” or “she” in a given sentence ?

The resolution of the coreferences on the names is performed topic by topic using predefined lexical patterns, lists of possible first names and lists of possible social roles (see Table 2 for some examples).

Lists	**Examples**
first names	chadwick, geoffrey, rogelio, tiffaney, kelle, lisandra, . . .
social roles	Administrator, Doctor_of_the_Church, Madam, bodyguard, . . .

Table 2: Examples of the different lists used.

For each person, we look for the functions which can be associated with this person (using the social roles list). We also look for possible pseudonyms and the gender for each person.

This knowledge is used to replace, as much as possible, all the references of the type: he, she, his, her and I by the person referred to.

For example, in the topic N51 entitled “General Pinochet Arrested”, many possible coreferences have been detected in the sentences presented in table 3. These coreferences are divided in two types: “he/his” or a social role (Foreign_Minister).

Table 4 shows how the different coreferences have been solved.

sentence number and doc id	**text**
14-NYT19981017.0086	British authorities refused to say where Pinochet is being held, nor did they set a date for when **he** would be questioned.
15-NYT19981017.0086	ever since **he** led a violent coup to overthrow Salvador-Allende-Gossens, the elected socialist president in 1973, Pinochet has been a political icon throughout Latin-America , representing the excesses of a long period of military rule.
16-NYT19981017.0086	an estimated 3000 Chileans were shot in the-streets or ”disappeared” during **his** rule, and a senior member of **his** regime was imprisoned under United-States pressure for the murder of former **Foreign-Minister** Orlando-Letelier in Washington in 1976.

Table 3: Sentences extract from the article with docid: NYT19981017.0086.

These ontologies have been gathered in order to achieve coreference resolution (see Tab 5).

The ontologies attributing a gender, a family function, and a social role to each person of the text have been built “by hand” by us just before the end of the competition time. They are very crude and enabled us to solve a limited number of anaphora. Nevertheless, as a consequence of the competition, we realized how our tagging language could be easily extended in order to be able to build these ontologies automatically, by extracting information from the text. This effort is taking place presently and we are currently building ontologies for family relationships and for 10 forms of social roles (e.g., army, justice, etc.). For instance, this enables us to recognize that Pinochet is a "father", a "general" or a "dictator" the last one being qualified as "former" or "retired", etc.

2.7 Concept recognition in texts

Once the terminology is completed, we use it to recognize the occurrences of a concept in a text, by clustering the terms into classes, each term being a linguistic instance of a concept. We are currently building a system of concept recognition in texts called ACT [9, 12].

ACT allows the user to build ontologies from the texts using the terms, the words, their tags and the context of each word or term.

In the present state of ACT, we use two types of information in order to spot the presence of a concept in the texts. As in [7], we use a superficial syntactic parser in order to obtain syntactic relationships among the words. In fact, most of them can be looked upon as terms. For instance, consider the set of syntactic relationships (as exemplified in in table 6) we actually used in order to recognize concepts.

In each case, you can notice that the grammatical relationship is not as important as the co-occurrence of the words in order to define a concept. This is why we are presently developing an extension of our terminology programs to apply them to detect noun-verb and verb-noun collocations that are quite enough to understand the topic of the sentence, once the determination of active/passive form of the verb has been done. This determination is quite simple as long the past and present participles have been correctly recognized in their verbal function (as opposed to their premodifier role). We developed a set of rules, written in our language, in order to perform this operation. Then, the distinction between material and auxiliary roles of the verbs 'have' and 'be' are easy to recognize, hence the passive forms.

2.7.1 Concept definition

The step of concept definition is entirely in the hands of the domain expert. It is important to stress that this entire process is strictly domain specific. The expert has to define what the interesting concepts are. One of the main difficulty of the Novelty TREC Track was that we had to deal with 50 different topics for which we did not have any available expert.

As our results will show, our tools are not really dedicated to this task when used by a non-expert. But it is quite important to notice that even if we were not expert of the different fields, we have successfully used ACT to define concepts that helped us in the relevant and novelty task.

2.7.2 Inductive step

The last step in this process of concept characterization is the inductive phase during which an existing categorization is automatically completed. Once the expert has provided the set of interesting concepts, and as many as possible linguistic instances of these concepts (called **seeds** of the linguistic traces of the concepts), the induction algorithm is able to complete these seeds so as to obtain a set of linguistic instances, exhaustive for the corpus under study. This algorithm is based on the determination of a few thousand groups, called 'potential concepts', containing words having the largest possible number of syntactic relations in common. These groups are grown by adding new syntactic relations to them, when this does not result in a too great reduction of the number of shared relations. The notion of "reduction of the number" is fixed by parameters chosen

sentence number and doc id	text
14-NYT19981017.0086	British authorities refused to say where **augusto-pinochet** is being held, nor did they set a date for when **augusto-pinochet** would be questioned.
15-NYT19981017.0086	ever since **augusto-pinochet** led a violent coup to overthrow **salvador-gossens**, the elected socialist president in 1973, **augusto-pinochet** has been a political icon throughout Latin-America , representing the excesses of a long period of military rule.
16-NYT19981017.0086	an estimated 3000 Chileans were shot in the-streets or ”disappeared” during **augusto-pinochet** rule, and a senior member of **augusto-pinochet** regime was imprisoned under United-States pressure for the murder of former Foreign-Minister(*orlando-letelier*) **orlando-letelier** in Washington in 1976.

Table 4: Coreference resolution for sentences 14, 15 and 16 from the article with docid: NYT19981017.0086.

<person>

<id>augusto-pinochet</id>

<lastName>pinochet</lastName>

<firstName>augusto</firstName>

<gender>1</gender>

<pseudo>pinochet</pseudo>

<socialRole>General</socialRole>

<socialRole>Mister</socialRole>

<socialRole>President</socialRole>

<socialRole>Senator</socialRole>

<socialRole>dictator</socialRole>

<socialRole>leader</socialRole>

<socialRole>president</socialRole>

</person>

<person>

<id>orlando-letelier</id>

<lastName>letelier</lastName>

<firstName>orlando</firstName>

<gender>1</gender>

<socialRole>Foreign-Minister</socialRole>

</person>

Table 5: Part of the ontologies built from topic N51.

Instance of the concept : legal-charges-crimes
(charge:Verb,for:Preposition,death:Object)
(suspect:Verb,in:Preposition,kill:Object)
(white-man:Object,accuse:Verb)
(accuse:Verb,of:Preposition,kill:Object)
Instance of the concept : aircraft-accident
(crash:Subject,kill:Verb)
(jet:Subject,crash:Verb)
(plane:Subject,crash:Verb)
(kill:Verb,in:Preposition,crash:Object)

Table 6: Examples of instances of two concepts.

by the user. We experimentally fixed the parameters so that the algorithm proposes about 50,000 such potential concepts. These are compared to the ones obtained by hand using a complex distance measure that will not be described here. When the distance between a potential concept and an expert-defined concept is small enough, then the relations belonging to the potential concept are added to those of the expert-defined concept.

3 Task 1 of Novelty Track

We used two different approaches to answer this task: an automatic approach and a semi-automatic approach. The automatic approach used all the steps of our chain of treatments except concepts. The semi-automatic used all the informations including concepts. In both cases, however, we never used any query expansion technique on the description of the relevance. This weakness might explain our results at detecting relevance.

3.1 Relevant Sentences Retrieval

3.1.1 Automatic approach

For this approach, we have proposed two runs using different types of information. The first run (Run 1) only used information identified in the topic definition: persons, noun, verb and numbers. The second run (Run 2) used these same information and all persons identified in the sentences of the texts.

For the person names, we differentiate two different coreferents. We will identify them as C_{3p} and C_{1s}. C_{3p} is used to identify persons referred with one of these coreferents: "he, his, she, her". C_{1s} is used to identify persons referred with the coreferent "I". The table 7 presents some examples for each of these coreferences.

All the named entities have been included in C_{3p}.

The terminology we built and the references and coreferences to individuals are used to compute a value of each sentence relevance. The computation is done as follows: we gather

- the individuals present in the topic and/or the texts processed as coreferences;
- the nominal terms and their various forms;
- the locations names present in the subject;
- the numbers (including the dates) present in the topic;
- the verbs and nouns present in the topic.

coreference	coreferent	example
C_{3p}	he	Nina brook also criticized the house, and **Republican Speaker David_Wilkins** for the failure to get a compromise. "a very real obstacle remains, a majority in the house who say they will not vote for a compromise, and the speaker of the house who says **he** will not vote for a compromise," Brook said.
C_{1s}	I	"**I** have not seen a significant change or shift in the house's position on that," **Wilkins** said.

Table 7: Coreferences examples.

Run 1	
Information	Weights
C_{3p} from topic	1
C_{3p} from texts	0
C_{1s} from topic	1
C_{1s} from texts	0
nominal terms	1
location	1
numbers	0.4
verbs	0.1
nouns	1

Run 2	
Information	Weights
C_{3p} from topic	1
C_{3p} from texts	0.1
C_{1p} from topic	1
C_{1p} from texts	0.1
nominal terms	1
location	1
numbers	0.4
verbs	0.1
nouns	1

Table 8: Weights used for Runs 1 and 2 of task 1.

Each sentence is replaced by a summary containing only these informations.

If a sentence contains none of the above information, its reduced form is empty - the sentence obtains a score of zero relevance. If not, the sum of the weights associated to each of this information provides a score of relevance for the whole sentence. The last step consists in selecting among the whole set of sentences the most relevant ones. The rule used considers relevant sentences if their scores are strictly higher than the average scores.

Table 8 shows the weights used for Run 1 and 2.

3.1.2 Automatic approach making use of domain knowledge

In this approach we used the concepts built by the expert with ACT. These concepts are combined with the informations gathered by the automatic step. In other words, we add the name of the concept as a (semantical) tag to its linguistic instances in the text. We did three runs (Run 3, 4 and 5) that used this approach.

We named P_a the sentences determined as relevant using the automatic approach, and P_c the sentences satisfying at least one rule in the set of concept's based rules that have been determined by the expert. These rules allow us to determine if a sentence if relevant or not.

Table 9 shows the different weights used to combine P_a and P_c sentences.

3.1.3 Results

For this first task, we have obtained poor results. But it is interesting to note that our approach achieves the best precision associated to a low recall.

	Relative Weights of	
	P_a	P_c
Run 3	0.1	0.9
Run 4	0.5	0.5
Run 5	0.9	0.1

Table 9: Weights used to combine automatic and concept's based approach.

topic	**Ranks**				
	Run 1	Run 2	Run 3	Run 4	Run 5
N51 (Event)	12	**3**	**1**	**2**	7
N54 (Event)	7	**4**	60	58	57
N55 (Event)	**5**	**4**	**1**	**2**	14
N61 (Opinion)	7	6	**1**	**3**	16
N70 (Opinion)	**3**	**1**	60	11	**4**
N76 (Opinion)	26	**1**	40	41	37
N78 (Opinion)	34	**3**	60	20	28
N82 (Event)	16	**1**	59	60	9
N85 (Event)	**4**	**3**	10	**2**	**1**
N94 (Opinion)	55	55	**1**	**2**	57
N95 (Event)	26	7	**1**	**2**	10
N96 (Opinion)	37	**4**	58	17	29

Table 10: Best 12 results for the relevant sentences retrieval of task 1 on a total of 50 topics.

We used the results given by TREC to evaluate our different approaches. Table 10 shows our best results (at least one run in the 5 first runs) for the relevant sentences retrieval task.

For this task, our better results are not linked to a particular domain (Event/Opinion). As a global conclusion, we can see that Run 2 gives better results than Run 1. We can then conclude that it seems useful to take into account the persons identified in the sentences of the texts even if they do not appear in the definition of what is relevant for the topic. Unfortunately, we have not used this run as a basic run for the semi-automatic approach. Results obtained for Runs 3, 4 and 5 have to be compared with those obtain with Run 1.

We can see that the use of concepts gives better results when the concepts were well defined (see topics N51, N55, N61, N94 and N95). Inversely, when the concepts were not useful, the semi-automatic approach gives worse results than the automatic one (see topics N54, N76, N78 and N96).

As we have already mentioned, we do not have an expert for the topics under study and it seems interesting to notify that using our tool (ACT) to quickly analyze the texts, we found some useful concepts for the task of relevance.

Table 11 shows the average Fscore obtain for each run on this task.

Run	**average Fscore**
Run 1	0.306
Run 2	0.356
Run 3	0.255
Run 4	0.299
Run 5	0.302

Table 11: Average Fscore associated to each run for the relevance sub-task of task 1.

topic	Ranks				
	Run 1	Run 2	Run 3	Run 4	Run 5
N51 (Event)	**4**	12	**1**	**2**	7
N55 (Event)	6	**2**	**1**	**3**	53
N57 (Event)	**2**	**3**	**1**	**3**	**3**
N60 (Opinion)	59	50	**1**	26	60
N61 (Opinion)	9	**5**	**1**	**2**	**2**
N62 (Opinion)	17	**4**	54	45	**1**
N63 (Opinion)	54	**4**	10	12	60
N70 (Opinion)	**1**	**2**	29	6	**3**
N76 (Opinion)	59	20	**2**	**1**	59
N77 (Opinion)	**1**	**3**	57	29	**2**
N84 (Opinion)	**1**	6	58	7	**3**
N91 (Opinion)	54	**2**	46	33	8

Table 12: Best 12 results for the novelty detection of task 1 on a total of 50 topics.

3.2 Task 1: Novelty Detection

For this task, we simply used our summary of each sentence to determine if a sentence was novel or not. We consider that a sentence is novel when it contains at least one information that has not been previously seen. As for the previous task, we have two automatic runs: Runs 1 and 2, and three semi-automatic runs: Runs 3, 4 and 5. In the two first runs, sentences cannot contain concepts and in the other ones, sentences contain this information.

3.3 Results

Table 12 shows our best results for this task. Quite surprisingly, some of these results were obtained for topics in which we were not able to achieve good results for the relevance task. For instance, for topics N57, N62, N63, N77, N84 and N91, our approach did not appear in the five best results for the relevance task but are in the five first for the novelty one.

More generally, we can see that the use of concepts improve the results for many runs (see topics N51, N55, N57, N60, N61, N62 and N76).

Finally, we can see that our approach performs better on Opinion topics than on Event ones.

Table 13 shows the average Fscore obtained for each run on this task.

Run	average Fscore
Run 1	0.066
Run 2	0.108
Run 3	0.098
Run 4	0.098
Run 5	0.072

Table 13: Average Fscore associated to each run for the novelty sub-task of task 1.

4 Task 2 of Novelty Track

For this second task, we used the given relevant sentences for each topic and we changed our novelty detection system.

Our results are better than those of the first task. We can see on the figure 3 the performance of each system participating to this task. Our results are prefixed with the name "lriaze2".

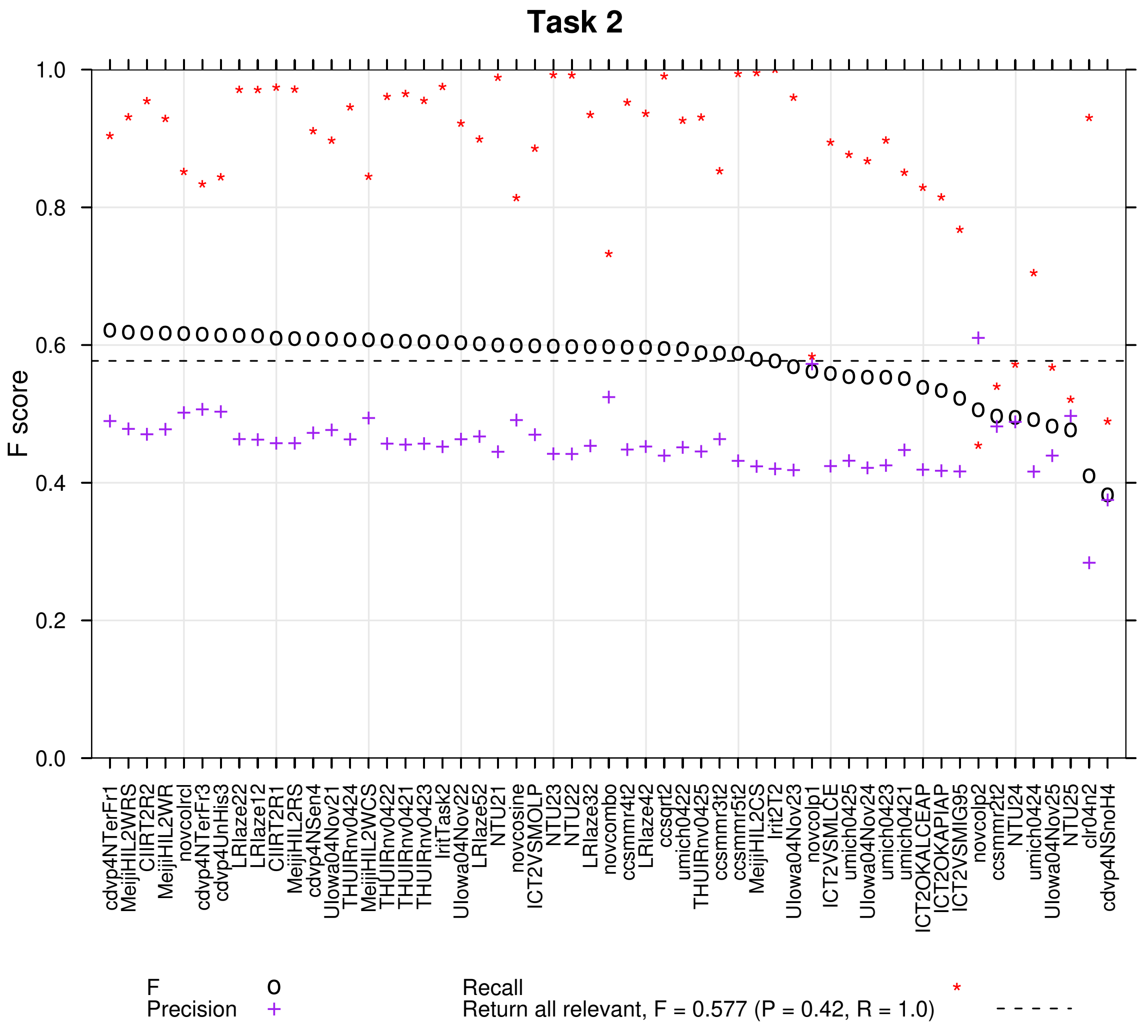

Figure 3: Fscore results for task 2. Our results are identified with LRIaze prefix

First of all, we applied some filters on each sentence. The following informations were removed from the sentences:

- punctuation characters,
- all words except non modal verbs, nouns and adjectives,
- and words: **be have get do can must should would could need shall will say says said relevant irrelevant**.

For each filtered sentence, we have computed the $TF \times IDF$ score and we have normalized it by the number of words of the sentence. This information ($TF \times IDF$) was the basic information used to determine if a sentence is novel or not.

To determine the novelty, we combined two different strategies: the use of a static threshold and the use of a dynamic threshold. Thus, a sentence is novel if

	no concept	using concepts
using coreference resolution	Run 1	Run 3
without coreference resolution	Run 2	Run 4 & 5

Table 14: Combination used for the different runs of task 2.

topic	Ranks				
	Run 1	Run 2	Run 3	Run 4	Run 5
N56 (Event)	**2**	**2**	39	39	41
N65 (Opinion)	14	14	4	9	25
N70 (Opinion)	**3**	**3**	**5**	**5**	**2**
N73 (Event)	**2**	**1**	11	10	7
N75 (Opinion)	**2**	**2**	**5**	**5**	8
N86 (Opinion)	12	12	**2**	**2**	**2**
N88 (Event)	**3**	**2**	25	21	23
N89 (Opinion)	**2**	**2**	38	38	30
N91 (Opinion)	**4**	**4**	36	37	31
N93 (Opinion)	**3**	**3**	35	35	11
N96 (Opinion)	**3**	**3**	6	6	22
N99 (Opinion)	**3**	**3**	6	**5**	23
N100 (Opinion)	18	18	8	20	**2**

Table 15: Best 13 results for the novelty detection of task 2 on a total of 50 topics.

- its score is higher than the static threshold.
- its score is higher than the one of the previous novel sentence: dynamic threshold

Two different approaches have been used to summarize the sentences.

- using or not the coreference resolution step
- using or not the concepts.

Table 14 shows the different combinations we used. Run 5 is the same as Run 4, but we changed the static threshold in order to be more permissive.

4.1 Results

Table 15 shows our best results for this task.

In this task, our approach obtains the first place only once. But, surprisingly the automatic runs provide the best results. For this task, our results seem to be stable.

As for the novelty detection task 1, the results are better for the texts relative to Opinion. Finally, it seems that coreference resolution doesn't affect the global results.

Table 16 shows the average Fscore obtained for each run on this task.

5 Conclusions

The text-mining system we are building deals with the specific problem of identifying the instances of relevant concepts found in the texts. This has several consequences.

We develop a chain of linguistic treatments such that the n-th module improves the semantic tagging of the (n-1)-th. This chain has to be friendly towards at least two kinds of experts:

a linguistic expert, especially for the modules dealing mostly with linguistic problems (such as correcting wrong grammatical tagging), and a field expert for the modules dealing mostly with the meaning of group of words. Our definition of friendliness includes also developing learning procedures adapted to various steps of the linguistic treatment, mainly for grammatical tagging, terminology, and concept learning.

In our view, concept learning requires a special learning procedure that we called Extensional Induction. Our interaction with the expert differs from classical supervised learning, in that the expert is not simply a resource who is only able to provide examples, and unable to provide the formalized knowledge underlying these examples. This is why we are developing specific programming languages which enable the field expert to intervene directly in some of the linguistic tasks.

Our approach is thus not particularly well adapted to the TREC competition, but our results show that the whole system is functional and that it provides usable information.

In this TREC competition we worked at two levels of our complete chain. In one level, we stopped the linguistic treatment after completion of the terminology (i.e., detecting the collocations relevant to the text). Relevance was then defined as the appearance of the same terms in the task definition (exactly as given by the TREC competition team) and in the texts. Our relatively poor results show that we should have been using relevance definitions extended by human-provided comments. Novelty was defined by a $TF \times IDF$ measurement which seems to work quite correctly, but that could be improved by using the expert-defined concepts as we shall now see.

The second level stopped the linguistic treatment after the definition of the concepts. Relevance was then defined by the presence of a relevant concept and novelty as presence of a new concept. For each of the 5 runs, this approach proved to be less efficient than the simpler first one. We noticed however that the use of concepts enabled us to obtain excellent results on specific topics (and extremely bad ones as well) in different runs.

We explain these very irregular results by our own lack of ability to define properly the relevant concepts for all the 50 topics since we got our best results on topics that either we understood well (e.g., Pinochet, topic N51) or that were found interesting (e.g., Lt-Col Collins, topic N85).

We knew before the competition that our approach was devoted to one topic dealt with by a specialist of this topic, and interested in solving the task. We thus knew that our approach was not particularly well-suited to the 50 TREC Novelty tasks. We nevertheless observed that when we devote enough attention to one topic, then both precision and recall are largely increased. Inversely, we may say that for topics we found uninteresting, the decrease in performance is not as bad as could be expected.

From our experience of participating in TREC 2004, we observed that our approach should be improved in several ways. The first observation is that the very classical supervised learning techniques (such as the one used in Brill's system, now improved by using Hidden Markov Chains, or Support Vector Machines) are very efficient when an already correctly tagged corpus is available. This is, say, never the case in real life, especially if we consider a semantical tagging (such as is necessary for technical fields). We thus have to develop learning techniques that follow the human

Run	**average Fscore**
Run 1	0.614
Run 2	0.614
Run 3	0.598
Run 4	0.597
Run 5	0.602

Table 16: Average Fscore associated to each run for Novelty task 2.

performing some kind of annotation 'by hand', and both learn from each improvement done by the human, but also help him/her in performing further improvements. This is a 'very active' learning which should be more developed.

The second observation is that coreference resolution is very topic dependent, and needs ontologies almost specific to the text under study. It follows that these ontologies have to be built 'on the fly' from the texts at hand. Once we have gathered a set of texts for which a given ontology is relevant, and we have built automatically this ontology, the problem is to determine the new texts it will be applicable upon. Our present working hypothesis is that Latent Semantic Analysis will be powerful enough to determine if a given new text belongs or not to the same 'social context' as a given existing group. We shall test this hypothesis in a very near future.

The third observation is relevant to an application to more technical texts, such as in the track Genomics. In this case, the notion of 'social context' is almost trivial since it corresponds to a given topic of the research in molecular biology. Inversely, the building of the ontologies is made much more difficult by the existence of a very large variety of linguistic forms used to describe the chemical entities involved. Our language for building ontologies is not yet able to handle this variability, but we already found some of the functionalities that have to be added in order to cope with it.

One last consequence of our participation to TREC Novelty is therefore that it forced us to build a language to express linguistic contexts for the relatively simple language of the newspapers. The limitations of this language shows us clearly the way to improvements that will enable us to extract relevant information for the case of the complex technical language of molecular biology.

References

[1] A. Amrani, Y. Kodratoff, and O. Matte-Tailliez. A semi-automatic system for tagging specialized corpora. *Proceedings of the Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, 3056:670–681, 2004.

[2] E. Brill. Some advances in transformation-based part of speech tagging. In *AAAI*, volume 1, pages 722–727, 1994.

[3] K-W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29, 1990.

[4] B. Daille, E. Gaussier, and J.M. Lang. An evaluation of statistical scores for word association. In *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications*, pages 177–188, 1998.

[5] T-E. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[6] S. Evert and H. Kermes. Experiments on Candidate Data for Collocation Extraction. In *Proceedings of the 10th Conference of The European Chapter of the Association for Computational Linguistics*, pages 83–86, 2003.

[7] D. Faure and T. Poibeau. First experiments of using semantic knowledge learned by asium for information extraction task using intex. *Ontology Learning, ECAI-2000 Workshop*, pages 7–12, 2000.

[8] C. Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings, 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), p. 341–348.*, 1999.

[9] Y. Kodratoff. Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. *Machine Learning and its Applications*, pages 1–21, 2001.

[10] T. Landauer and S. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[11] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography 3 (4)*, pages 235–244, 1990. revised august 1993.

[12] M. Roche, J. Azé, O. Matte-Tailliez, and Y. Kodratoff. Mining texts by association rules discovery in a technical corpus. In *Proceedings of IIPWM'04 (Intelligent Information Processing and Web Mining), Springer Verlag series "Advances in Soft Computing"*, pages 89–98, 2004.

[13] M. Roche, T. Heitz, O. Matte-Tailliez, and Y. Kodratoff. EXIT: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In *Proceedings of JADT'04 (International Conference on Statistical Analysis of Textual Data)*, volume 2, pages 946–956, 2004.

[14] F. Xu, D. Kurz, J. Piskorski, and S. Schmeier. A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping. In *LREC 2002, the third international conference on language resources and evaluation*, 2002.