Sentence filtering for information extraction in genomics, a classification problem

Claire Nédellec, Mohamed Ould Abdel Vetah, Philippe Bessières

Sentence Filtering for Information Extraction in Genomics, a Classification Problem

Claire Nédellec[1], Mohamed Ould Abdel Vetah,[1,2] and Philippe Bessières[3]

[1]LRI UMR 8623 CNRS
Université Paris-Sud,
91405 Orsay cedex
cn@lri.f

[2]ValiGen SA
Tour Neptune
92086 La-Défense
ould@lri.fr

[3] Mathématique, Informatique et Génome(MIG) INRA,
78026 Versailles cedex
philb@biotec.jouy.inra.fr

Abstract. In some domains, Information Extraction (IE) from texts requires syntactic and semantic parsing. This analysis is computationally expensive and IE is potentially noisy if it applies to the whole set of documents when the relevant information is sparse. A preprocessing phase that selects the fragments which are potentially relevant increases the efficiency of the IE process. This phase has to be fast and based on a shallow description of the texts. We applied various classification methods — IVI, a Naive Bayes learner and C4.5 — to this fragment filtering task in the domain of functional genomics. This paper describes the results of this study. We show that the IVI and Naive Bayes methods with feature selection gives the best results as compared with their results without feature selection and with C4.5 results.

1. Introduction

As an increasing amount of information becomes available in the form of electronic documents, the need for intelligent text processing makes shallow text understanding methods such as Information Extraction (IE) particularly useful. Up to now, IE has been restrictively defined by DARPA's MUC (Message Understanding Conference) program [10] as the task of extracting specific, well-defined types of information from natural language texts in restricted domains with the specific objective of filling pre-defined template slots and databases. We claim that in many domains, IE systems have to rely on deep analysis methods local to the relevant fragments. They should combine the semantic-conceptual analysis of text understanding methods and information extraction by pattern matching; in a first step the relevant textual fragments are filtered based on shallow criteria; in a second step, a representation of the content of the fragments is built by successive interpretation operations based on syntactico-semantic lexicon following a classical approach in text understanding, finally, extraction rules are applied to the resulting interpretations in order to identify the relevant information and store it in a database in the suitable format, usually by filling forms in the MUC case. These three steps differ by the

nature of the knowledge that they exploit and by the complexity of the methods applied. The second step, that is, the syntactico-semantic parsing is the most expensive in terms of resources. The first step, i.e. the filtering of the relevant fragments, allows to limit that analysis to what is needed only, by focussing it on the fragments that potentially contain relevant information. This selection is even more crucial as the information to be extracted is sparser. The sparseness problem had been pointed out in previous research in IE [15] and [16] but no practical solution has been proposed. The main consequence is that the first step must be fast, even if this implies some lack of precision. It must thus be based on a shallow description of the text. The application of learning to the filtering of relevant fragments has received little attention in IE compared to other tasks such as learning for name entity recognition or learning extraction patterns [15, 16]. This lack of interest is due to the type of texts that are generally handled by IE, which are those proposed in the MUC competition. Those texts are usually short and the information to be extracted is generally dense, so that prefiltering is less or not needed at all. The type of information to be extracted such as company names or a seminar starting times often requires only a shallow analysis, the computational cost of which is low enough to avoid prefiltering. This is not the case in other IE tasks such as identifying gene interaction in functional genomics, the application that we describe here.

From a Machine Learning point of view, filtering can be viewed as a classification problem. Textual fragments have to be classified in two classes: potentially relevant for IE or not. The learning examples represent fragments, (sentences in this application) and the example attributes are the significant and the lemmatized words (in a canonical form) of the sentences. We compared experimentally the classification method IVI proposed in [12] for IE in functional genomics, a Naïve Bayes (NB) method [9], and a decision tree-based method, C4.5 [14], on three different datasets in functional genomics described in section 2. As a consequence of the example representation, the datasets are very sparse in the attribute space; the examples are described by few attributes. Thus, in addition to the basic methods, we studied the effect of feature selection as a preprocessing step. The objective of this study is to identify the best classification methods for filtering sentences in functional genomics and to characterize the corpora with respect to these methods. This paper reports our results on comparing classification methods. The methods and the evaluation protocol are detailed in section 3. Section 4 reports and discusses the experimental results. Future work is presented in section 5.

2. The application domain: functional genomics

2.1 A genomics point of view on IE

The application problem to which applying IE is here about modeling the gene interactions from text, in the domain of functional genomics. This problem has been previously described in [1, 12, 11, 18] among others. The existence of numerous scientific and technical domains sharing strong common aspects with functional

genomics, from a document point of view, will allow adapting the methods developed here to other application domains. This is typically the case for related domains in biology, but more generally, the methods will be transposable and exploitable in any application of knowledge extraction from scientific and technical documents.

Modeling interactions between genes is of significant interest for biologists, because it is a prerequisite step towards the understanding of the cell functioning. To date, most of the biological knowledge about these interactions is not described into databanks, but only in the form of scientific summaries and articles. Therefore, their exploitation is a major milestone towards building models of interactions between genes. Actually, genome research projects have generated new experimental approaches like DNA chips at the level of the whole organisms. A research team is now able to quickly produce thousands of measurements. This very new context for biologists is calling for automatic extraction of knowledge from text, to be able to interpret and making sense of elementary measurements from the laboratory by linking them to scientific literature. The bibliographic databases can be searched via Internet using keyword queries that retrieve a superset of the relevant paper abstracts. For example, the query "*Bacillus subtilis transcription*" related to the gene interaction topic retrieves 2209 abstracts.

Extract of a MedLine abstract on Bacillus subtilis.

```
UI  - 99175219 [..]
AB  - [..] It is a critical regulator of cot genes encoding
proteins that form the spore coat late in development. Most cot
genes, and the gerE gene, are transcribed by sigmaK RNA polymerase.
Previously, it was shown that the GerE protein inhibits
transcription in vitro of the sigK gene encoding sigmaK. Here, we
show that GerE binds near the sigK transcriptional start site, [..]
```

Then the biologist has to identify the relevant fragments, (in bold-face in the example) in the abstracts and to extract the useful knowledge with respect to the goal of identifying gene interaction. Then, he has to represent it in a structured way so that it can be recorded in a database for further querying and processing. The more general goal is to identify all the interactions and molecular regulations and to build a functional network.

Example of a form filled with the information extracted from the sentence in the example.

Interaction	**Type**: negative		
	Agent: GerE protein		
	Target:	**Expression**	**Source**: sigK gene
			Product: sigmaK protein

This domain is representative of the scope of our study on automatizing filtering of relevant fragment for IE: the information to be extracted is local, mainly located in single sentences or part of sentences. It is very sparse in the document set. For instance, only 2.5 % (470) of the 20000 sentences contain relevant information on gene interaction in the 2209 *Bacillus subtilis* abstracts mentioned above. We contend that the information extraction has to rely on a deep analysis. Indeed previous approaches based on shallow descriptions of the texts (e. g. IE techniques such as transducers defined manually and based on significant verb and gene names [1, 11, 18])

or on statistic measures of keywords co-occurrences [12, 17] (e.g. information retrieval-based techniques) yield limited results with either a bad recall or a low precision. The following example illustrates some of the problems encountered:

"**GerE** stimulates **cotD** transcription and inhibits **cotA** transcription in vitro by sigma K RNA polymerase, as expected from in vivo studies, and, unexpectedly, profoundly inhibits in vitro transcription of the gene (**sigK**) that encode **sigma K**.".

The IE methods based on keywords or gene names (bold-face) and interaction verbs (framed) are not able to identify the inhibition interaction between GerE and sigK gene transcription (28 words far) or, if they will, also erroneously identify interactions between cotD and sigK and between cotA and sigK. Extracting relevant knowledge in the selected documents thus requires more complex IE methods such as syntaxico-semantic methods based on lexical and semantic resources specific to the domain[1]. The characteristics of this application thus perfectly fit the requirements for applying classification methods for filtering relevant text as an IE preprocessing step.

2.2 Textual corpora and learning sets

The robustness of the classification methods has been evaluated with respect to different writing styles, different biological species, and then different gene interaction models. The classification methods chosen have been applied, evaluated and compared on three different datasets. These sets have been built from paper abstracts about three species: the first set, denoted *Dro*, is about a fly, *Drosophila melanogaster*[2], the second, denoted *Bs*, is about a bacterium, *Bacillus subtilis*[3] and the third, denoted *HM*, is about the mouse and the human[4]. They come from two bibliographic databases with different writing styles. The Dro dataset is from FlyBase, the database devoted to *Drosophila* genes. Its abstracts are concise, 2 or 3 sentences long, the sentences short and the syntax quite simple. The two others are from MedLine, the generalist biology bibliographic database. The abstracts of MedLine are longer, around 10 sentences, in more complex syntactic forms than those of FlyBase. The abstracts have been selected by the queries "*Bacillus subtilis transcription*" for Bs dataset and *Telomere*, *Apoptose*, *DNA replication*, *DNA repair*, *cell cycle control*, *two-hybrid* and *interaction* for HM. The examples sets have been selected in the abstracts under the locality assumption that the sentence level is the suitable granularity degree in this IE application, as it is often the case in Machine Learning for IE applications, [15] and [16]. It is assumed that the potentially relevant sentences in the Bs and HM sets contain at least two gene or protein names denoting the agents of the interaction as in previous work. In the Dro set as it has been provided to us, the sentences contain exactly two gene or protein names. This difference should not affect the filtering phase but the extraction phase only. The identification of gene names identification for the

[1] This is the goal of the Caderige project of which this research is part.

[2] The Dro example set has been provided as such by B. Jacq and V. Pillet from LGPD-IBDM.

[3] This set has been built by P. Bessières (MIG, INRA) in the Caderige project.

[4] It has been provided as such by the LGPD-IBDM and the ValiGen company.

Dro and HM set has been done manually by LGPD-IBDM biologists. This manual selection results in 530 abstracts Dro set, and 105 abstracts and 962 sentences for HM set that have been provided to us as such. This manual processing affects the classification results as it will be shown in section 4. The sentence selection for the Bs set has been automatically done with the help of a list of gene and protein names of *Bacillus subtilis* and their derivations provided by MIG and manually completed by new derivations observed in the corpus. The problem of the automatic identification of gene names in genomics document has been recently studied and recognized as a prerequisite for any further automatic document processing because of the lack of exhaustive dictionary and because of the varying notation [2, 5, 6, 13].

Table 1. Features of the example sets.

	Dro	Bs	HM
Document data base	FlyBase	MedLine	
# bibliographic references	> 100 000	around 16 Millions	
# sentences per abstract	2, 3	approximatively 10	
species	*Drosophila*	*Bacillus subtilis*	mouse - human
# biblio. references to the species	20 300	15 213	4 067 879
# abstracts selected (queries)	20 300	2209	32448
# abstracts selected after manual step	530	Not relevant	105
# sentences in the abstracts	5 244	around 20 000	962
# sentences filtered (at least 2 gene names) = # examples	1197	932	407
# attributes	1701	2340	1789
# positive examples (PosEx)	653	470	240
# negative examples (NegEx)	544	462	167

Training example of Bs dataset built from the sentence, which illustrates section 2.1.

```
Example : addition stimulate transcription inhibit transcription
vitro RNA polymerase expected vivo study unexpectedly profoundly
inhibit vitro transcription gene encode
Class : Positive
```

The attributes that describe the learning examples represent the significant and lemmatized words of the sentences. They are boolean in the case of C4.5 and they represent the number of occurences in the sentence in the other cases, i.e, IVI and NB. The examples have been classified into the positive and the negative categories, i.e. describing *at least one* interaction (positive) or none at all (negative). The HM and Bs sentences have been lemmatized using Xerox shallow parser. Stopwords such as determinant have been removed as non-discriminant with the help of the list provided by Patrice Bonhomme (LORIA). It initially contains 620 words and it has been revised with respect to the application. After stopwords removal, the three example sets remain very sparse in the feature. Half of the attributes describe a single example. The capacity to deal with data sparseness was thus one of the criteria for choosing the classification methods.

3. Classification methods

3.1 Method descriptions

The classification method *IVI* had been applied to Dro dataset [12]. It is based on the example weight measure defined by (2), which is itself based on the attribute weight measure defined by (1) where $occ(Att_i,ex_j)$ represents the value, (i.e., the number of occurrences) of the attribute i for the example j. The class of the example is determined with respect to a threshold experimentally set to 0. Examples with weights above (resp. below) the threshold are classified as positive (resp. negative).

$$Weight(Att_i) = \frac{\sum_{ex_j^+ \in PosEx} occ(Att_i, ex_j^+) - \sum_{ex_j^- \in NegEx} occ(Att_i, ex_j^-)}{\sum_{ex_j \in Ex} occ(Att_i, ex_j)} \quad (1)$$

$$IVI(ex) = \sum_{i=1}^{|Att(Ex)|} Weight(Att_i) \quad (2)$$

The Naïve Bayes method (NB) as defined by [9], seemed to be suitable for the problem at hand because of the data sparseness in the attribute space. As IVI, NB estimates the probabilities for each attribute to describe positive examples and negative examples with respect to the number of their occurrences in the training set. The probability that a given example belongs to a given class is estimated by (4), the product of the probability estimations of the example attributes, given the class. The example is assigned to the class for which this probability is the highest.

$$Pr(Att_j | Class_i) = \frac{\sum_{ex_k \in Class_i} occ(Att_j, ex_k)}{\sum_{l=1}^{|Class|} \sum_{ex_k \in Class_l} occ(Att_j, ex_k) + |Class|} \quad (3)$$

$$Pr(ex | Class_i) = \prod_{j=1}^{|Att(ex)|} Pr(Att_j | Class_i) \quad (4)$$

The Laplace law (3) yields better results here as compared with the basic estimate because its smoothing feature deals well with the data sparseness. The independence assumption of the attributes is obviously not verified here also previous work has shown surprisingly good performances of NB despite of this constrain [4]. The third class of methods applied is C4.5 and C4.5Rules. Compared to NB and IVI, the decision tree computed by C4.5 is more informative and explicit about the combination of attributes that denote interactions, and thus potentially on the phrases that could be useful for further information extraction.

3.2 Feature selection

The data sparseness is potentially a drawback for C4.5 Feature selection appears here as a good way to filter the most relevant attributes for improving classification [19] but also for selecting the suitable corpus for other IE preprocessing tasks such as

semantic class learning (section 5). This latter goal has motivated the choice a filtering method for feature selection instead of a wrapper method selection [7], where the classification algorithms would be repeatedly applied and evaluated on attribute subsets in order to identify the best subset and the best classifier at the same time [8]. The measure of attribute relevance used here is based on (5). It measures the capacity of each attribute to characterize a class, independently of the other attributes and of the classification method. The attributes are all ranked according to this measure and the best of them are selected for describing the training sets (section 4).

$$\text{DiscrimP}(\text{Att}) = \frac{\sum_{i=1}^{|\text{Class}|} \text{Max}\{\Pr(\text{Att},\text{Cl}_i), 1-\Pr(\text{Att},\text{Cl}_i)\}}{|\text{Class}|} \quad (5)$$

3.3 Evaluation metrics

The methods have been evaluated and compared with the usual criteria, that is, recall (7), precision (8), and the F-measure (9), computed for the three datasets.

$$\text{Recall}(\text{Class}_i) = \frac{|\text{Ex} \in \text{Class}_i \text{ and assigned to Class}_i|}{|\text{Ex} \in \text{Class}_i|} \quad (6)$$

$$\text{Precision}(\text{Class}_i) = \frac{|\text{Ex} \in \text{Class}_i \text{ and assigned to Class}_i|}{|\text{Ex classified in Class}_i|} \quad (7)$$

$$F = \frac{(\beta^2+1) * \text{Precision} * \text{Recall}}{(\beta^2 * \text{Precision}) + \text{Recall}} \quad (8)$$

More attention is given to the results obtained for the positive class because the examples classified as positive only will be transferred to the IE component. The recall rate for this class should therefore be high even if this implies some lack of precision. The β factor of the F-measure has been experimentally set to 1.65 in order to favor the recall. IVI and BN have been evaluated by leave-one-out on each dataset. For performance reasons, C4.5 and C4.5Rules have been only trained on 90 % of the learning sets and tested on the remaining 10 %. The results presented here are computed as the average of the test results for ten independent partitions.

4. Evaluation

4.1 Comparison of the IVI, C4.5 and BN methods

The first experiments allow the comparison of C4.5, C4.5Rules, NB and IVI on the three datasets (Table 2). As recall and precision computed for two classes yields to the same rates, they appear in a same line. NB has been applied here with the Laplace law. In the three cases, NB and IVI results are better than C4.5 and C4.5Rules results. This can be explained by the sparseness and the heterogeneity of the data. The global precision rate is 5 to 8 % higher and the precision rate for the positive class is 4 to

12 % higher. However, the good behavior of the IVI-BN family is not verified by the recall rate for the positive on the Dro dataset: C4.5 recall rate is better than NB and IVI on this set (13 %) but worse on Bs' and HM's ones (-12 to -13 %). The origin of Dro dataset could explain these results: it comes from FlyBase where the sentences are much shorter than those of MedLine, from which Bs and HM are extracted. Thus Dro examples are described by *less attributes* although the ratio of the number of attributes to the examples is similar to Bs one. This could explain the overgenerality of C4.5 results on Dro set illustrated by the high recall and bad precision rates. The analysis of NB and IVI results shows that NB behaves slightly better at a global level.

Table 2. Comparison of C4.5, C4.5Rules, IVI and BN on the three datasets.

Corpus	Dro				Bs				HM			
Method	C4.5	C4.5 R	BN	IVI	C4.5	C4.5 R	BN	IVI	C4.5	C4.5 R	BN	IVI
Recall Positive	**88,9** ±2.4	86,8 ±2.6	75,3 ±2.9	69,1 ±3.5	63,9 ±4.3	71,4 ±4.1	**85,7** ±3.2	82,6 ±3.4	88,3 ±4.1	84,5 ±4.1	**97,1** ±2.1	90 ±3.8
Precision Positive	68,1 ±3.6	70,5 ±3.5	82 ±3.2	**83,1** ±2.8	63,4 ±4.3	62,8 ±4.4	66,6 ±4.3	**67,4** ±4.2	63,7 ±6.1	64,2 ±6.1	68,5 ±5.9	**70,3** ±5.8
Recall-precision for all	72 ±2.5	73.6 ±2.5	**77,5** ±2.4	75,4 ±2.4	62,4 ±3.1	62,9 ±3.1	**71,1** ±2.9	71 ±2.9	63,7 ±4.1	63,4 ±4.7	**72** ±4.4	71,5 ±4.4

However, their behaviors on the positive examples are very different: NB achieves a higher recall than IVI (3 to 7 %) while IVI achieves a better precision than NB (1 to 2 %) but the difference is smaller. The higher recall and precision rates for positive on HM compared to Bs is explained by the way the HM set has been built. The selection of the sentences in the abstracts has been done manually by the biologists among a huge number of candidate sentences (Table 1) and the bias of the choice could explain the homogeneity of this dataset compared to Bs which has been selected automatically. This hypothesis has been confirmed by further experiments on the reusability of the classifiers learned from one corpus and tested on others. As a better recall is preferred in our application, the conclusion on these experiments is that NB should be preferred for data from MedLine (Bs and HM) while for FlyBase (Dro), it would depend on how much the IE component would be able to deal with sentences filtered with a low precision. C4.5 should be chosen if the best recall is preferable while BN should be chosen for its best recall-precision tradeoff.

4.2 Feature selection

As described in section 3, the attributes for each dataset have been ranked according to their relevance. For instance, the best attributes for the Dro set are, `downstream`, `interact`, `modulate`, `autoregulate`, and `eliminate`. The effect of feature selection on the learning results of IVI, NB and C4.5Rules methods has been evaluated by selecting the best n attributes, n varying from hundred to the total number of attributes, by increments of hundred.

4.2.1 Effect of feature selection on NB results

For the three sets, the recall noticeably increases and the precision noticeably decreases with the number of relevant attributes selected, which is what is expected, (Fig. 2, Fig. 3 and Fig. 4). The F-measure increases in the first quarter, more or less stabilizes on a plateau on a half, slightly increasing since recall is predominant over precision in our setting of F-measure (section 3), and then decreases in the last quarter or fifth, after a small pick in the case of Dro and Bs sets. According to the F-measure, the best attribute selections in terms of the recall - precision compromise are thus at the end of the plateau around 3/4 - 4/5 of the total number of attributes. For the Dro set, it is around 1400 attributes and for Bs set it is around 1900 attributes. One can notice that the recall for positive examples for the Dro and Bs sets is 10 to 15 % higher than the global recall and that is the opposite for the precision, which is exactly what is desirable in our application.

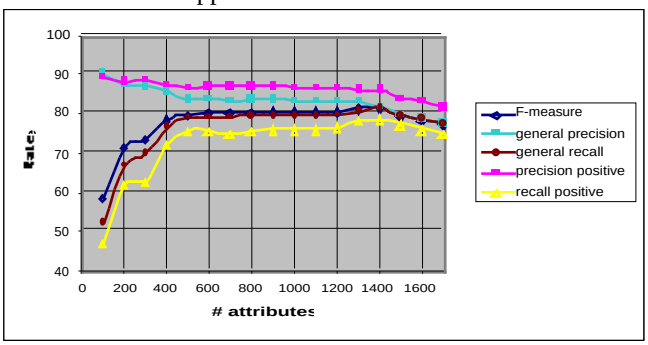

Fig. 2. NB classification results after feature selection on Dro set.

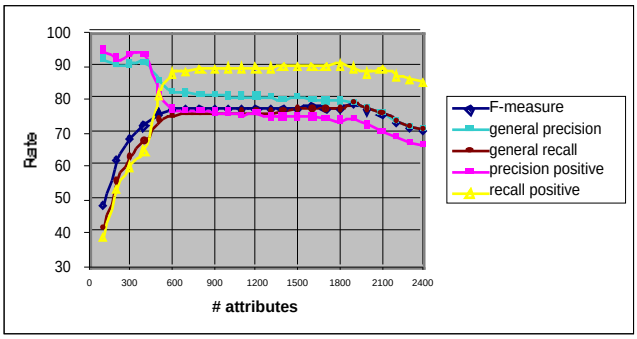

Fig. 3. NB classification results after feature selection on Bs set.

For the HM set, this phenomenon is even more noticeable: the recall of the positive is very high, close to 100 %, and 20 % higher than the global recall (Fig. 4). Compared to the other sets the plateau is more horizontal between 400 et 1900 attributes after a slight increase between 400 and 800, and there is no pick before the decrease, then the global recall-precision rate is stable between 800 and 1400 and all points are equivalent in this interval. This could be explained by the homogeneity of the HM dataset that affected the initial classification results in the same way (4.1).

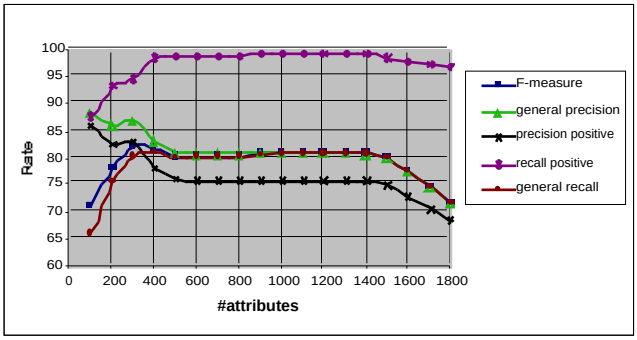

Fig. 4. NB classification results after feature selection on HM set.

Table 3 presents a summary of the results obtained with NB without and after feature selection for the best attribute. NB results are improved by feature selection. The gain is very high for HM, around 10 %, less for Bs (6-7 %), and 4-5 % for Dro.

Table 3. Comparison of NB results with the best feature selection level.

Dataset	Dro		Bs		HM	
# attributes	all att. 1701	1400	all att. 2340	1800	All att. 1789	900-1300
Rec. Positive	75,3 ± 2.9	**79**±3.1	85,7±3.2	**90,8**±2.6	97,1±2.1	**99,6**±0.8
Prec. Positive	82 ±3.2	**86,4**±2.6	66,6±4.3	**74,1**±4.00	68,5±5.9	**76,1**±5.4
Prec.-Rec. for all classes	77,5 ±2.4	Rec. **81,8** ±2.2 Prec. **82,1** ± 2.2	71,1±2.9	Rec. **77,5**±2.7 Prec. **79,9**±2.6	72±4.4	Rec. **81,1** ±3.8 Prec. **81,3** ±3.8

4.2.2 Effect of feature selection on C4.5 and IVI results

Similar experiments have been done with C4.5. There are summarized in Table 4.

Table 4. Comparison of C4.5 results with the best feature selection level.

Dataset	Dro		Bs		HM	
# attributes	all at. 1701	1400	all at. 2340	1600	All at. 1789	1300
Recall Pos.	**86,8** ±2.6	84,5 ±2.8	**71,4** ±4.1	70,1 ±4.2	84,5 ±4.6	**84,6** ±4.6
Precision Pos.	70,5 ±3.5	**75**±3.33	62,8 ±4.4	**71,4** ±4.13	64,2 ±6.1	**78,8** ±4.6
Prec-Recall for all	73.7 ±2.5	**75,3** ±2.4	62,9± 3.1	**71,1**±3	63,4 ±4.7	**74,9** ±5.2

The conclusions are similar to NB ones: feature selection improves the global classification results for all sets, the global improvement is important for Bs and HM (9 %), and less for Dro (1,6 %) for the same reasons related to the origin of the corpora as previously pointed out.

The similar experiments done with IVI are summarized in Table 5. The improvement is higher for IVI than for the two other methods. Its range is between approximately +6 % for Dro, +10 % for Bs to +16 % for HM.

Table 5. Comparison of IVI results with the best feature selection level.

Dataset	Dro		Bs		HM	
# attributes	all at. 1701	1300	all at. 2340	1900	all at. 1789	1400
Recall Pos.	69 ±3.5	**77,9** ±3.2	82,6 ±3.42	**91,5**±2.5	90 ±3.8	**98,3** ±1.6
Prec. Pos.	83,6 ±2.9	**88,4** ±2.5	67,4 ±4.23	**78,3**±3.7	70,3 ±5.8	**83,4** ±4.7
Prec.-Rec. for all	75,4±2.4	Rec. **81,9**±2.2 Prec. **84,1**±2.1	71±2.9	Rec. **82,8**±2.4 Prec. **83,2**±2.4	71,5±4.4	Rec. **87,5**±1.6 Prec. **87,5**±4.7

4.2.3 Conclusion on the effect of feature selection on classification

The comparison between the experimental results with C4.5, NB and IVI for the best feature selection shows that IVI globally behaves better than the two others do. With respect to the recall rate for positive, NB behaves slightly better or similarly to IVI (1 to 2 %) while IVI precision rates are better than NB ones (2 to 7 %). Therefore, in the case where the good positive recall is preferred NB with feature selection should be chosen for all datasets except for those like Dro that are less sparse and more homogeneous and where C4.5 without feature selection is better. In the case where a best recall-precision compromise is preferred, IVI with feature selection should be applied.

5. Future work

This research focuses on the classification of sentences represented by their significant and lemmatized words. The methods studied yield global recall and precision rates higher than 80 % and high recall rates for the positive class with feature selection by prefiltering. Other criteria should be tested for selecting the attributes, such as information gain and mutual information. Better results should also be obtained with classification with more information gain global measures that would take into account the dependency between the words which form significant noun phrases. For instance the results of the ongoing work at LIPN on the acquisition of terminology for gene interaction should reduce both the number of attributes and their dependency. We also plan to study the reduction of the number of attributes by replacing in the examples, the words by the concept (the semantic class) they belong to as learnt from a biological corpus. Moreover, classification should be improved by reducing the data heterogeneity by pre-clustering the examples; one classifier would then be learned per example cluster. From an IE point of view, the assumption that relevant sentences contain at least two gene or protein names should be relaxed. The attribute ranking will be used to identify automatically other potentially relevant sentences. Finally learning extraction rules requires semantic class acquisition. The attribute ranking will be also used to select the most relevant syntagms in the training corpora for learning semantic classes. Learning will thus focus on the potentially most relevant concepts with respect to the extraction task.

Acknowledgement

This work is financially supported by CNRS, INRA, INRIA and INSERM through *Caderige* contract. The authors thank V. Pillet, C. Brun and B. Jacq for the *Dro* and *HM* sets.

References

1. Blaschke C., Andrade M. A., Ouzounis C. and Valencia A., "Automatic Extraction of biological information from scientific text: protein-protein interactions", in Proc. of *ISMB'99*, 1999.
2. Collier N., Nobata C. and Tsujii, "Extracting the names of genes and gene products with a hidden Markov model. In Proc. *COLING'2000*, Saarbrück,, July-August 2000.
3. Craven M. and Kumlien J., "Constructing Biological Knowledge Bases by Extracting Information from Text Sources.", In Proc. of *ISMB'99*, 1999.
4. Domingos P. and Pazzani M., "Beyond independence: conditions for the optimality of the simple Bayesian classifier", in Proc. of *ICML'96*, Saitta L. (ed.), pp. 105-112, 1996.
5. Fukuda K., Tsunoda T., Tamura A. and Takagi T., "Toward Information Extraction: Identifying protein names from biological papers". In Proc. *PSB'98*, 1998.
6. Humphreys K., Demetriou G, and Gaizauskas R., "Two applications of information extraction to biological science article: enzyme interaction and protein structure". In Proc. of *PSB'2000*, vol.5, pp. 502-513, Honolulu, 2000.
7. John G. and Kohavi R., "Wrappers for feature subset selection", in *Artificial Intelligence Journal*, 1997.
8. Langley P. and Sage S., "Induction of selective Bayesian classifiers", in Proc. of UAI'94, Lopez de Mantaras R. (Ed.), pp. 399-406, Morgan Kaufmann, 1994.
9. Mitchell, T. M., *Machine Learning*, Mac Graw Hill, 1997.
10. Proceedings of the *Message Understanding Conference* (MUC-4-7), Morgan Kaufman, San Mateo, USA, 1992-98.
11. Ono T., Hishigaki H., Tanigami A., and Takagi T., "Automated extraction of information on protein-protein interactions from the biological literature". In Bioinformatics, vol 17 no 2 2001, pp. 155-161, 2001
12. Pillet V., Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information, thèse de l'Université de droit, d'économie et des sciences d'Aix-Marseille, 2000.
13. Proux, D., Rechenmann, F., Julliard, L., Pillet, V., Jacq, B., "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction". In *Genome Informatics 1998*, S. Miyano and T. Takagi, (Eds), Universal Academy Press, Inc, Tokyo, Japan, pp. 72 - 80, 1998.
14. Quinlan J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1992.
15. Riloff E., "Automatically constructing a Dictionary for Information Extraction Tasks". In Proc. of *AAAI-93*, pp. 811-816, AAAI Press / The MIT Press, 1993.
16. Soderland S., "Learning Information Extraction Rules for Semi-Structured and Free Text" in *Machine Learning Journal*, vol 34, 1999.
17. Stapley B. J. and Benoit G., "Bibliometrics: Information Retrieval and Visualization from co-occurrence of gene names in MedLine abstracts". In Proc. of *PSB'2000*, 2000.
18. Thomas, J., Milward, D., Ouzounis C., Pulman S. and Caroll M., "Automatic Extraction of Protein Interactions from Scientific Abstracts". In Proc. of *PSB'2000*, vol.5, p. 502-513, Honolulu, 2000.
19. Yang Y. and Pedersen J., "A comparative study on feature selection in text categorization.", in Proc. of *ICML'97*, 1997.