A bayesian classifier for the recognition of the impersonal occurrences of the 'it' pronoun

Davy Weissenbacher, Adeline Nazarenko

A bayesian classifier for the recognition of the impersonal occurrences of the *it* pronoun

Weissenbacher Davy*, Nazarenko Adeline*

* Laboratoire d'Informatique de Paris-Nord
Universite Paris-Nord
Villetaneuse, FRANCE
{davy.weissenbacher, adeline.nazarenko}@lipn.univ-paris13.fr

Abstract

This paper presents a new system that makes the distinction between the impersonal and anaphoric occurrences of the *it* pronoun. Compared with the state of the art methods, our system relies on the same types of linguistic knowledge but performs better. We argue that this is due to the bayesian model on which it is based: it enables to combine various pieces of knowledge and to exploit even unreliable ones in the process of pronoun occurrence classification.

1. Introduction

To solve the anaphoric relations of the pronoun *'it'*, it is necessary at first to make the distinction between the impersonal and the anaphoric occurrences of the pronoun. This task can be considered as a traditional classification problem. The main difficulty is to identify the relevant attributes, whose availability, reliability and usefulness vary with the type and domain of the corpus on which the classification is carried out. We argue that taking into account these attribute properties calls for a new classification approach in anaphora resolution.

This article describes in details the system that was first presented in (Weissenbacher, 2006).Our system is based on the formalism of the Bayesian Networks (BN). This probabilistic formalism, still little exploited for NLP, gives a method for integrating heterogeneous types of attributes as well as an elegant mechanism to exploit them and to *a priori* estimate the reliability of each attribute in the classification decision.

For sake of comparison, various methods presented in the state of the art for the distinction between the impersonal and anaphoric pronoun occurrences have been implemented as separated systems. We also integrated them in a unique classifier based on a bayesian network. The performance of the integrated system have been compared with that of the previous methods taken in isolation. These systems have been tested on a corpus made up of abtsracts of genomic articles. The results show that our integrated system performs significantly better than the state of the art ones.

The next section presents the recognition of impersonal occurrences as a classification task. Section 3 justifies our bayesian approach, whereas sections 4 and 5 describe our system and its results.

2. The classification of pronoun occurrences

In most systems, the anaphora resolution process starts with the distinction between the impersonal and the anaphoric occurrences of the pronouns.

2.1. The classification problem

This task can be considered as a classification problem in which each pronoun occurrence is tagged either as anaphoric or as impersonal according to various contextual clues.

Let be *Corpus* a set of texts from a given domain, *Training_Corpus* and *Test_Corpus* two disjoint subsets of *Corpus*, C_1 and C_2 the classes of the impersonal and anaphoric occurrences of *it*. Let e be an occurence of a pronoun in the *Test_Corpus*.

e can be represented as a vector $a = v_1, ...v_a$ of normalized attribute values defined over **R**, the set of clues used for the classification. For example, the fact that a pronoun occurs at the beginning of an abstract is considered as a relevant clue, since we know that such an occurrence is more likely to be impersonal than other ones. This is represented by the boolean attribute, which we will call *Start_Sentence* in the following.

Bayes' theorem indicates how to predict the best class for an unseen example e on the basis of the observations made on training data: the class i that maximizes the following probability must be chosen for e

$$P(C_i|e) = \frac{P(e|C_i)*P(C_i)}{P(e)}$$

where $C_i \in \{C_1, C_2\}$ and $P(e|C_i)$ is estimated from the *Training_Corpus*. If we consider that the attributes are independent of each other, the classifier is a "Naive Bayes Classifier" (NBC) and the probability $P(e|C_i)$ is expressed by the product $P(v_1|C_i) * ... * P(v_a|C_i)$. The probability to maximize can be reformulated as

$$P(C_i|e) = \frac{P(C_i)}{P(e)} \prod_{j=1}^{a} P(v_j|C_i)$$

The system described here is a bayesian classifier (BC) for pronoun ocurrences. We show that this new bayesian approach improves the quality of the recognition and we argue that, more generally, it is well suited for natural language processing tasks.

2.2. Previous approaches for *it* classification

One of first *it* occurrences classification systems was proposed by (Husk and Paice, 1987). It relies on a set of first

order logical rules to make the distinction between the impersonal and anaphoric occurences of the pronoun.
It exploits the fact that impersonal sequences often have a similar form: they start with an *it* and end with a delimiter like *to, that, whether*... Paice's rules express these constraints (with slight variations from one delimiter to another). They specify the left context of the pronoun (it should not be immediately preceded by a preposition like *before, from, to*), the distance between the pronoun and the delimiter (no longer than 25 words), and the types of lexical items that may occur between the pronoun and the delimiter (*e.g. certain, known, unclear, etc.*).
The tests performed by Paice give some good results with 91.4%Accuracy[1] on a technical corpus. However the performances are degraded if one applies them on a corpus of a different domain. The attributes which are discriminating on a technical corpus may be less relevant for a different one. In order to avoid this problem, (Lappin and Leass, 1994) proposes some more constrained rules in the form of finite state automata, which exhaustively describes the sequences containing an impersonal pronoun.
Due to the noise produced by the attributes of (Lappin and Leass, 1994), (Evans, 2001) and (Litran et al., 2004) give up such complex properties and concentrate on more reliable and more accessible attributes. They focus on surface clues but a training phase reduces the estimation error by determining the relative weight of the attributes.
Despite their lack of reliability, the attributes of (Husk and Paice, 1987) and (Lappin and Leass, 1994) express linguistic pieces of knowledge which are relevant for our task. Our system combines them with surface clues.

2.3. The specificity of the NLP classification attributes

The performance of a classifier mainly depends on the quality of the attributes used to describe the data. Choosing and representing these attributes is a difficult task.
The first difficulty comes from the fact that NLP attributes are complex and heterogeneous. As shown above, for pronoun occurrence classification, the previous approaches have exploited rich linguistic information (such as syntactic automata or semantic classes) as well as very simple ones (such as word distance and sentence boundaries). The selection of the relevant attributes for a given task is based either on human expertise or on corpus evidence and machine learning. A language must be defined to represent these various classification attributes and the representation power of that formalism directly affects the *discrimination power* of the classifier.If a relevant attribute misses, it may become impossible to distinguish a positive anaphoric example from a negative one.
The classification algorithms are often based on the hypothesis of attribute independence. This raises a second problem for NLP tasks where independant attributes are difficult to isolate. The position of a word in the sentence and its syntactic role are often correlated, for instance.
A third problem affects the classifier reliability. Whatever attributes are chosen at the representation level, each pronoun occurrence must be described according to these attributes. Some attribute values may be easy to identify but others require a previous NLP computation (*e.g.* syntactic parsing or semantic tagging) and nothing guarantees that the resulting values are fully reliable.
Lastly, even when the attributes are relevant and reliable for the task, their respective weight in the classification vary from one corpus to another. An attribute which is a good indicator for a class *I* on a given corpus may rather be a class *J* indicator on a different corpus. This may lead to estimation errors in the classification process.

[1] Accuracy(Acc) is a classification measure: Acc=$\frac{TP+TN}{TP+TN+FP+FN}$ where FP is the number of anaphoric pronoun occurences tagged as impersonal, which we call the false positive cases, FN the number of impersonal pronoun ocurrences tagged as anaphoric, the false negative cases. TP and TN are the numbers of correctly tagged impersonal and anaphoric pronoun occurences, the true positive and true negative cases respectively.

3. A bayesian approach

Our pronoun occurrence classifier is based on a bayesian approach. We argue that this model gives an elegant solution to solve the previous problems related to the representation choice, the attribute dependancies, their lack of reliability and their variability. The bayesian network is based on a probabilist formalism, which makes it possible to exploit unreliable attributes. It graphically models the influence between uncertain and heterogeneous pieces of knowledge. The learning mechanism enables the training of the classifier on different corpora, if necessary.

3.1. NLP and bayesian classifiers

Until now, very few NLP systems have been based on this formalism despite its advantages. The system proposed by (Peshkin and Pfeffer, 2003) aims at extracting information from texts of seminar announcements in order to fill automatically the information fields (such as date, place, presenter) of a seminar announcement forms. The BN allows to integrate within a single representation the various linguistic or more generally textual pieces of information that play a role in the extraction information (IE) task. Compared with classical IE systems, this attribute integration increases the system's performance.
On another IE task, the Roth's system exploits a BN to reason on uncertain knowledge (Roth and Wen-tau, 2002). It recognizes entities and their relations at the same time, a method that proves to be more efficient than the traditional one operating first on entities and then on their relationships. Actually, the two steps are not independent: for example, in terrorism news stories, knowing that entities *X* and *Y* are persons reinforces the probability of a relation *X is the assassin of Y* and vice versa. In discovering the entities and their relations simultaneously, (Roth and Wen-tau, 2002) shows how the attribute dependancies compensate for the lack of reliability of the attribute values in the BN.

3.2. A simple exemple of a bayesian network for pronoun classification

The BN is a formalism designed to reason on uncertain pieces of information. It is defined by a qualitative description of the attributes and their dependances (each attribute is represented as a node of an acyclic and oriented

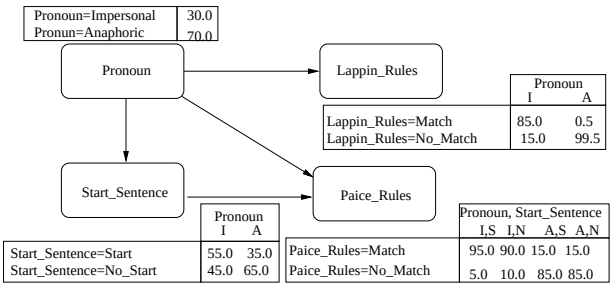

Figure 1: Exemple of a bayesian classifior modeled by a bayesian network

graph) and by a quantitative description that indicates their relative weights (a set conditional probability tables where each Random Variable (RV) is associated with a node of the graph).

The initial graph models the linguistic expertise. A node is associated to any piece of information that is supposed to play a role in the pronoun classification. Their dependancies are explicitly represented as edges in the graph.

The bayesian classification is a two-step process. The first parameterizing phase exploits a tagged training corpus and learns the relative weight of the attributes in the decision. The contribution of an attribute is represented as a conditional probability table. The second inferential phase classifies the new pronoun occurrences as impersonal or anaphoric. For each occurrence, some of the *a priori* probabilities are modified according to the actual values of the attributes which are observed in the context. These revised *a posteriori* probabilities propagate through the network edges to update the *a priori* values of the still unknown attributes.

Let us explain on a very simple example the mechanism of the BC. The figure 1 presents a network for the classification of the pronoun *it*. There are 4 different nodes: the decision node (`Pronoun`) and three attribute nodes which respectively represent the fact that the occurrence occurs at the beginning of the sentence (`Start_Sentence`) or that it is matched by a Paice or Lappin rule (`Lappin/Paice_Rules`). The links show the node dependancies.

Each node is associated with a probability table. During the parameterizing phase, the *a priori* probabilities are computed. From a training corpus analysis or an expert estimation, we assume *a priori* that approximately a third of the *it* occurrences are impersonal and we set P(Pronoum=impersonal)=0.3. A link connects the variables `Pronoun` and `Lappin_Rules`, indicating that a pronoun has more chance to be matched by a Lappin's rule if it is an impersonal one. In the same way, the links connecting (`Pronoun`,`Paice_Rules`) and (`Pronoun`,`Start_Sentence`) indicate respectively that a pronoun is more likely to be matched by a Paice's rule and to start a sentence if it is an impersonal one. Finally, the arc (`Start_Sentence,Paice_Rules`) indicates that the reliability of the Paice's rule is increased if the *it* occurrence begins the sentence. This influence is measured by the conditional probabilities table associated to the `Paice_Rule` node in the figure 1.

Once all *a priori* conditional probabilities have been determined, the inference phase begins. For example let us consider the sentence *It is well documented that treatment of serum-grown....* Contextual evidence leads to revise some probabilities: as no Lappin's rule matches the sequence, we set P(Lappin_Rules = No_Match)=1; as one Paice's rule matches the sequence, we set P(Paice_Rules = Match)=1; as the sequence starts the sentence, we set P(Start_Sentence = Start)=1.

On the basis of these observations, the pronoun type probability:

(1) P(Pronoun=Impersonal|Lappin_Rules=No_Match, Start_Sentence=Start, Paice_Rules=Match)

can be computed. According to the conditional probability definition, we get

$$P(Pronoun=I|Lappin\_Rules=N, Start\_Sentence=S, Paice\_Rules=M) = \frac{P(Pronoun=I, Lappin\_Rules=N, Start\_Sentence=S, Paice\_Rules=M)}{P(Lappin\_Rules=N, Start\_Sentence=S, Paice\_Rules=M)}$$

Relying on the inference links, we can compact the global probability law:

$$P(Pronoun, Lappin\_Rules, Start\_Sentence, Paice\_Rules) = P(Pronoun).P(Lappin\_Rules|Pronoun).P(Start\_Sentence|Pronoun).P(Paice\_Rules|Pronoun, Start\_Sentence)$$

The numerator of the equation (1) is computed on the basis of the *a priori* conditional probabilities and the denumerator is computed by marginalizing the global probability law (Pearl, 1998).

The network therefore infers that the pronoun is impersonal with a probability of 38.9% from the facts that a Paice's rule matches the sequence and the sequence starts the sentence. This initial network can be enriched with additional RV or by taking into account uncertain and missing information. For example we could indicate that the reliability of an observation is lower than 100% and set P(Lappin_Rules=No_Match)=0.9.

Note that a naive bayesian classifior (NBC) is a particular BN. If we delete the arc (`Start_Sentence,Paice_Rule`) then all attributes are considered as independent and we get the naive bayes classifior associated to our BN. More generaly a BN is a NBC if the graph with a length equal to 1, the root is the prediction node and no arc from one leaf to another.

4. Description of the bayesian classifier

4.1. The classification attributes

The structure of our BN is based on the linguistic expertise. We take into account all the attributes which are mentioned in the state of the art, what ever their importance may be. In the following together, the names of the attributes correspond to the network nodes of figure 2. The probabilities are computed on the training corpus[2].

[2]Because of the large number of conditional probabilities, we give only the simplified probabilities of a naive bayesian classifior which reflects the same proportions.

[Previous-Word] If the word immediately preceding the pronoun is a preposition, the pronoun is with no doubt anaphoric ($P(Previous\_Word = Match|Pronoun = Anaphoric) = 1$). It is the most discriminative attribute.

[Start-Clause, Start-Sentence, Start-Abstract] If the pronoun is one of the first 3 words of the abstract, one of the first 3 words of the sentence, or the first word of the clause, we consider that it begins respectively the abstract, the sentence or the clause. This position has an impact on the probability for the pronoun to be impersonal. For example, if the pronoun follows a comma or a period, the probability for the pronoun to be impersonal is reinforced ($P(Start\_Proposition = \{Comma, Mark\}|Pronoun = Impersonal) = \{0.35, 0.55\}$), whereas it is more likely to be anaphoric if it follows a word ($P(Start\_Proposition = Word|Pronoun = Anaphoric) = 0.4$).

[Grammatical_Role] The probability for the pronoun to be impersonal increases if the pronoun is a subject of the sentence and decreases in the other cases ($P(Grammatical\_Role = Subject|Pronoun = Impersonal) = 0.98$ and $P(Grammatical\_Role = \{Object, Preposition\}|Pronoun = Anaphoric) = \{0.11, 0.5\}$).

[Lappin-Rules] If the sequence containing the pronoun is matched by one of the Lappin's rules, the probability for the pronoun to be anaphoric is very low ($P(Lappin\_Rules = Match|Pronoun = Anaphoric) = 0.01$).

[Unknown-Words] In our automata, we have loosened the Lappin's rules, so that there can be at most three unmatched words between the pronoun and the delimiter, but the more unknown words there is, the less reliable the rule is.

[Paice-Rules] If the sequence which contains the pronoun is matched by one of the Paice's rules, the probability for the pronoun to be anaphoric decreases ($P(Paice\_Rules = Match|Pronoun = Anaphoric) = 0.11$).

[Delimiter] This variable corresponds to the first delimiter following the *it* pronoun. The reliability of the Paice's rules depends on this delimiter type ($P(Delimitrr = \{To, That, Whether\_if, Which\_Who\}|Pronoun = Anaphoric) = \{0.09, 0.02, 0.005, 0.003\}$).

[Length-Pronoun-Delimiter] This variable corresponds to the number of words that occur between the pronoun and the delimiter. Based on corpus analysis, we consider that 10 words is the maximal length but the longer the sequence is, the less reliable the delimiter is. This dependence is represented by the arc (Delimiter,Length_Pronoun_Delimiter).

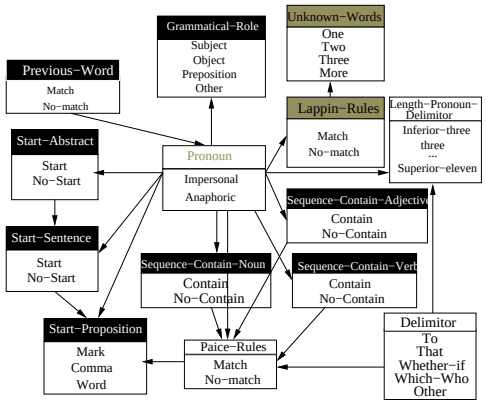

Figure 2: A Bayesian Network for impersonal *it* classification

[Sequence-Contain-{noun,verb,adjective}] These boolean variables check if a certain type of nouns (resp. verbs or adjectives) occurs between the pronoun and the delimiter. These nouns, verbs and adjectives are those that have already been found in the same position in a training training corpus. Their presence decreases the probability of the pronoun to be anaphoric ($P(Sequence\_Contain\_\{noun, verb, adjective\} = Contain|Pronoun = Anaphoric) = \{0.01, 0.03, 0.02\}$).

4.2. The attribute representation

Each pronoun occurrence to tag is therefore represented as an attribute vector. We could have decomposed the previous attributes in order to get an homogeneous vector of independent and elementary attributes but the vectors would have been more difficult to implement, interpret and update: adding any new attribute would lead to modify the full vector representation. Our approach is simpler: we consider the rules of (Husk and Paice, 1987) and of (Lappin and Leass, 1994) as attributes as such and we add the surface clues introduced by (Litran et al., 2004) and (Evans, 2001) if they are are not already included in the previous rules. Neither the rules nor the surface clues are fully reliable indicators of the pronoun status but they complement each other. They encode heterogeneous pieces of information and consequently produce different false negative and positive cases. The Lappin's rules have a good precision but tag only few pronouns. On the opposite, the Paice's rules, which have a good recall, are not precise enough to be exploited in isolation.

The figure 2 describes the Bayesian Network (BN) that we use to classify the impersonal occurrences. The attribute representing the fact that a rule of Lappin matches a sequence is marked in gray, in white (resp. in black) the attributes corresponding to the rules of Paice (resp. (Litran et al., 2004) and (Evans, 2001)). The prediction node is the Pronoun one in the middle. It estimates the probability for a given occurrence to be impersonal or anaphoric.

4.3. Implementation of the system

Our system is written in Perl. To compute the attribute values for a given pronoun occurrence, it integrates a set of finite state transducers (implemented with Unitex[3]) and exploits a Link Parser analysis of the corpus (Sleator and Temperley, 1991)[4]. For the classification process, it relies on a BN implemented in language C using the Netica[5] API. In the first parameterizing step, the system computes automatically from the training corpus frequencies, the conditional *a priori* probabilities for all possible Random Variables (RV) of our classifier. These probabilities express the weight of the various attributes in the decision, their *a priori* reliability in the classification task. Among the 2000 *it* occurences of a training corpus (see 5.1.), the Lappin's rules recognized 649 of the 727 impersonal occurrences and they have erroneously recognized 17 occurrences as impersonal, so we set the `Lappin_Rules` node probabilities as P(Lappin_Rules=Match|Pronoun=Impersonal)=89.2% and P(Lappin_Rules=Match|Pronoun=Anaphoric)=1.3%, which are the expected number of false negative cases and false positive cases produced by the Lappin's rules.

During the second inference step, for each sequence containing an occurence of the pronoun *it*, we apply the Paices's and Lappin's rules and we determine the values of the remaining attributes (see 4.1.). The values of the RVs are updated according to these observations and a new probability is computed for the `Pronoun` node: if it is higher or equal to 50% the occurrence of the pronoun is classified as impersonal; it is anaphoric otherwise.

Let us consider the following sentence extracted from our corpus:*It had previously been thought that ZEBRA's capacity to disrupt EBV latency....* As no Lappin's rule recognizes the sequence – even by tolerating 3 unknown words–, we set P(Lappin_Rules=No_Match)=1 and P(Unknown_Words=More)=1. As a Paice's rule matches the sequence with 4 words between the pronoun and the delimiter *that*, we set P(Paice_Rules=Match)=1, P(Length_Pronoun_Delimiter=4)=1 and P(Delimiter=That)=1. We check the boolean attributes: the sequence is at the beginning of the sentence but the sentence is not the first of the abstract; it contains the adverb *previously* and the verb *think*, which words belong to our semantic classes. Others node values are set in the same manner. The *a priori* probability for an occurrence to be impersonal is 36.2%. After modifying the probabilities of the nodes of the BN according to the corpus observations, the *a posteriori* probability computed for this occurence is 99.9% and the system considers it as impersonal.

5. System validation

Our working corpus is made of genomic research articles extracted from the database *Medline*[6] on the basis of some keywords such as *bacillus subtilis*, *transcription factors*, *Human*, etc. We extracted 11 966 abstracts (approximately 5 million words), in which we identified 3.347 occurrences of the pronoun *it*. Two human annotators tagged them either as anaphoric or as impersonal. After discussion, the two annotators achieved a total agreement.

Method	Results(Acc/FP/FN)		
Lappin's Automata	88.11%	12.8	169.1
Paice's Rules	88.88%	123.6	24.2
Support Vector Machine	92.71%	-	-
Naive Bayesian Classifier	92.58%	74.1	19.5
Bayesian Classifier	**95.91%**	**21.0**	**38.2**

Figure 3: Prediction Results (Accuracy/False Positive Cases/False Negatives Cases)

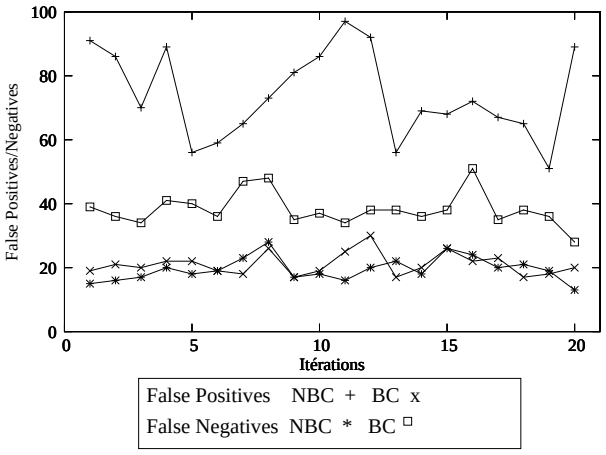

Figure 4: False Positives/Negatives of the classifiors for each iteration

5.1. Results

Since the size of our corpus is relatively small, we performed a 20-cross validation. We considered a third of the corpus for training and the remaining for testing.

Table 3 summarizes the average results (in Acc) of the state-of-the-art methods described above[7] and of our two classifiers, our Bayesian Classifier (BC) and the Naive Bayesian Classifier (NBC) associated to it. The results show that the BC achieves a better classification than other systems, in particular the rule-based ones. The BC exploits all the relevant attributes in such a way that they compensate for each other, whereas rule-based systems fully depend on the reliability of their attributes. These results confirm our initial analysis 2.2.: a low recall for Lappin's automata and a bad precision for Paice's rules.

Despite the good performances of the NBC (see table 3), the BC obtains better results at each iteration and this difference is statistically significant (*w.r.t.* a test-t). Based on these 20 couples of exactitude values, the figure 4 details the FP and FN rates of each classifior for each iteration.

[3]URL: http://www-igm.univ-mlv.fr/ unitex/

[4]As our system was tested on a biological corpus, we exploited a version of the Link Parser that has been tuned for biology (Aubin et al., 2005).

[5]URL: http://www.norsys.com/netica.html

[6]http://www.ncbi.nlm.nih.gov/entrez/

[7]The Clement's SVM score have been computed on similar biological corpus as ours. The FP and FN values were not published.

5.2. Error analysis

Our BC, which has a good precision, nevertheless tags as impersonal some occurrences which are not (false positive cases). The most recurrent error corresponds to the sequences ending with a delimiter *to* that are recognized by some Paice's rules. Even if none Lappin's rules matches the sequence, its minimal length and the fact that it contains some specific words like *assumed* or *shown* makes this configuration caracteristic enough to tag the pronoun as non-anaphoric. When the delimiter is *that*, this decision is a good one [8] but it is always incorrect when the delimiter is *to*[9]. In that latter case, the rules should be more carefully designed.

Three different factors explain the false negative cases. (1) Some sequences are ignored because the delimiter remained implicit[10] and this is still an unresolved a problem. (2) The presence of apposition clauses increases the sequence length and decreases its reliability, but this should be fixed by exploiting a deeper syntactic analysis. (3) Our specific verb, adjective and noun classes are not exhaustive but we plan to enrich them automatically[11].

6. Conclusion

The distinction of the impersonal pronouns can be considered as a classification problem. The main difficulty deals with the selection and representation of the classifier attributes. The complexity and the variation of the natural language make it difficult to isolate the relevant attributes for a given task and the computation of the attribute values may be noisy. Lastly, the relative importance of the different attributes varies from corpus to another, which calls for a corpus-based approach.

In this article, we have proposed a classifior based on the formalism of the bayesian networks, a formalism adapted to classify data described by these types of attributes. Integrating within a single model heterogeneous and complementary pieced of knowledge increases the discrimination power. Representing the reliability of each attribute with *a priori* conditional probabilities decreases the classification errors caused by the noisiest attributes. Taking the attribute dependancies into account gives better results than the state of the art systems.

Based on this first encouraging result, we are currently extending our BN to tackle the more complex task of anaphora resolution.

7. References

S. Aubin, A. Nazarenko, and C. Nedellec. 2005. Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 89–93.

R. Evans. 2001. Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, 16:45–57.

G.D. Husk and C.D. Paice. 1987. Towards the automatic recognition of anaphoric features in english text: the impersonal pronoun it. *Computer Speech and Language*, 2:109–132.

S. Lappin and H.J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

J.C. Clemente Litran, K. Satou, and K. Torisawa. 2004. Improving the identification of non-anaphoric it using support vector machines. In *Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 58–61.

J. Pearl. 1998. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman.

L. Peshkin and A. Pfeffer. 2003. Bayesian information extraction network. In *In Proc.18th Int. Joint Conf. Artifical Intelligence*.

D. Roth and Y. Wen-tau. 2002. Probalistic reasoning for entity and relation recognition. In *Colling'02*.

D. Sleator and D. Temperley. 1991. *Parsing English with a Link Grammar*. Technical report.

D. Weissenbacher. 2006. Bayesian network, a model for nlp? In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'06)*.

[8] Like in the sentence *It is assumed that the SecY protein of B. subtilis has multiple roles...*

[9] Like in the sentence *It is assumed to play a role in ...*

[10] For example *Thus, it appears T3SO4 has no intrinsic...*

[11] In our experiments, we noticed that if a Paice's rule matches a sequence in the first clause of the first sentence in the abstract then the pronoun is impersonal. We could automatically extract from Medline a large number of such sentences and extend our classes by selecting the verbs, adjectives and nouns occuring between the pronoun and the delimiter in these sentences.