A Bayesian approach combining surface clues and linguistic knowledge: Application to the anaphora resolution problem

Davy Weissenbacher, Adeline Nazarenko

A Bayesian approach combining surface clues and linguistic knowledge: Application to the anaphora resolution problem

Davy Weissenbacher, Adeline Nazarenko
Université Paris-Nord - Laboratoire d'Informatique de Paris-Nord.
99 av. J-B. Clément 93430 Villetaneuse, FRANCE
dw@lipn.univ-paris13.fr, *nazarenko@lipn.univ-paris13.fr*

Abstract

In NLP, A traditional distinction opposes the linguistically-based systems and the knowledge-poor ones which mainly rely on surface clues. Each approach has its drawbacks and its advantages. In this paper, we propose a new method which is based on Bayes Networks and allows to combine both types of information. As a case study, we focus on the specific task of pronominal anaphora resolution which is known as a difficult NLP problem. We show that our bayesian system performs better than state-of-the art anaphora resolution ones.

Keywords

Bayesian Network, Anaphora Resolution, linguistic knowledge, surface clue

1 Introduction

One often opposes knowledge based and knowledge poor Natural Language Processing (NLP) systems. The first ones exploit complex knowledge pieces which may be automatically or manually built and which are therefore not always reliable or available. The second ones rely on machine learning methods and take only surface clues into account. They give mitigated results on complex NLP tasks.

This paper proposes an approach that overcomes that opposition. It relies of the Bayesian Network formalism, a probabilistic model designed for reasoning on dubious, partial and lacking information, which is still little exploited in NLP.

This approach is tested on the resolution of the anaphoric pronoun *it*, which is a complex task involving different types of knowledge and for which there is a clear contrast between linguisically-based methods and methods based on surface clues. We designed a system that relies on a Bayesian Network for the classification of antecedent candidates and we compare its performances with that of a state-of-the-art system, MARS proposed by R. Mitkov [10], which can be considered as a knowledge-poor system.

The next section presents the opposition between rich and poor approaches in the case of anaphoric pronoun resolution. Section 3 describes the formalism of the Bayesian Networks, its advantages for NLP and we present our classifier for anaphora resolution. In Section 4, we compare its performances with several other ones. The last section discusses the results.

2 The opposition between linguistic knowledge and surface clues

Anaphora is a linguistic relation that holds between two textual units where one of them (the *anaphor*) cannot get interpreted as such but refers to the other, which usually occurs before (the *antecedent*). As the presence of anaphors significantly degrades the performances of NLP tasks such as information extraction or text synthesis, a lot of work has been devoted to the automatic resolution of these anaphoric relationships, *i.e.* the identification of the antecedents of anaphoric pronouns. In this paper, we focus on the pronoun *it* in English texts, which is a well-known and frequent type of anaphors.

2.1 The usefulness of surface clues

The traditional approach for anaphora resolution is composed of three steps: the distinction between anaphoric and impersonal occurrences of the pronoun (*it is known that...* vs. *it produced...*), the selection of antecedent candidates and the choice of the most plausible antecedent. For each of these steps, the first systems relied on complex linguistic knowledge that reflected the deep syntactic and semantic constraints of anaphoric relations. As these constraints seemed too complex to build automatically, the first systems relied on a set of manually designed rules, which required a thorough corpus analysis.

During the 1990's, several systems relying on surface clues were proposed to face the need for robust and less expensive anaphora resolution methods [14]. These systems got rid of the complex linguistic rules of the first ones and tried to approximate them by simple clues that are presumably more reliable and easier to compute.

For instance, [7] modifies the RAP algorithm initially proposed by [8]. Considering that a deep syntactic analysis cannot be achieved with state-of-the-art parsers, the authors implement a relaxed version of that algorithm based on shallow parsing. They show

that, even if full parses are not available, the performances of the new algorithm are comparable to that of the first one. Another example is given in [5], which proposes to approximate the semantic constraints by cooccurrence frequencies. The antecedent is supposed to belong to the same distributional subject or object class as the anaphoric pronoun and the reported experiments show that these distributional constraints can partially supply deeper semantic ones.

2.2 The limits of surface clues

The surface clues proposed during the 1990's enabled to build robust systems [10] but recent work has underlined their limits.

Since the predicate-arguments schemata that improve the candidate filtering [11], are seldom available, they have been approximated by concurrence frequencies [5]. However, [2] shows that these frequencies do not really enhance the performances of a system that is already based on morpho-syntactic knowledge. The contribution of frequencies seems to pertain more to hazard than to semantics.

Such a conclusion brings back to the initial problem. Anaphora resolution involves complex syntactic and semantic knowledge that is not always available and which is often not fully reliable. Previous works have tried to substitute linguistic knowledge by surface clues which are easier to compute and therefore more reliable. However these clues only partially reflect the linguistic constaints and may lead to erroneous decisions, when solving ambiguous cases.

2.3 Enriching the surface clues with linguistic information

The MARS system [10] relies on surface clues to identify the most salient element in the discourse fragment preceding a pronoun occurrence. This salient element is considered as the most probable pronoun antecedent. The system relies on a part-of-speech tagging (POS tagging) of the text and applies some simple grammar rules in order to list the noun phrases (NPs) of the two sentences preceding a given pronoun occurrence and the NPs preceding the pronoun occurrence in the same sentence. For each NP associated to the pronoun occurrence, a set of constraints and preferences is applied. The constraints filter out the impersonal pronoun occurrences and the NPs that cannot be antecedent. The preferences rank the remaining NP candidates. Each preference is associated with a score, either positive or negative, and the various scores of a candidate are summed up in a global score. The antecedent with the highest score is chosen. When two candidates end with the same score, additional heuristics are used to rank them[1].

We propose a new system exploiting all the surface clues of MARS but also integrating the linguistic constraints that the surface clues approximate, whenever some linguistic knowledge is available. We argue that combining both types of information is beneficial. For instance, the subject of a sentence is often the most salient element but, since the syntactic role analysis may be erroneous, it is useful to exploit in parallel the information relative to the NP location: the surface clue (the first NP of the sentence is very often the verb subject) corroborates the grammatical role hypothesis.

Our system is modeled thanks to a Bayesian Network. This type of representation has been designed to reason on dubious and incomplete knowledge. It offers a probabilistic approach that unifies in a single representation deep linguistic constraints and surface clues. This unification allows to corroborate linguistic constraints with the surface properties observed in corpora and to correct the errors made by the systems based on surface clues.

3 A unified approach: the Bayesian model

3.1 Classification problems

As many other NLP tasks, distinguishing anaphoric and impersonal pronoun occurrences and more generally solving anaphors can be considered as classification problems [3].

Let us consider for instance the choice of the antecedent among various candidates. Let *Corpus* be a set of texts belonging to the same domain, *Training_Corpus* and *Test_Corpus* two distinct subsets of *Corpus*, *Pronouns* and *NounPhases*, the sets of the pronoun and NP occurrences of *Corpus*. Let R be the set of potential anaphora relationships. Each relation $r_{i,j}$ is represented as a couple (p_i, np_j) of *Pronouns X NounPhrases*, where np_j is considered as a candidate antecedent of the pronoun p_i[2]. *Antecedents* and *Not_Antecedents* are two complementary subclasses of R. $r_{i,j}$ belongs to the class *Antecedent* if the candidate np_j is the antecedent of the pronoun occurrence p_i. It belongs to the class *Not_Antecedent* if the candidate np_j is not the antecedent or if the pronoun p_i is impersonal. Any couple $r_{i,j}$ is described by a vector $a = v_1, ..., v_a$ of attributes whose values are defined in $\mathbf{R}$. Each attribute v_k is selected on the basis of an analysis of *Training_Corpus* and corresponds to either a linguistic piece of knowledge or a surface clue.

The Bayes theorem states how to predict the best class for any new couple of candidate NP and pronoun occurrence of *Test_Corpus* on the basis of the regularities observed on the set of couples of *Training_Corpus*: select the class that maximises the probability

$$P(C|E) = \frac{P(E|C) * P(C)}{P(E)}$$

where $C \in \{Antecedent, Not\_Antecedent\}$, E is an example of *Test_Corpus* and $P(E|C)$ is the conditional probability that E belongs to the class C given the values of the attributes of E. That probability is estimated on the basis of the training examples.

[1] The final ranking depends on the types of the preferences that have been used for each candidate and the most recent candidate is chosen, if nothing else applies.

[2] Actually, only the NPs occurring in the two sentences preceding the pronoun occurrence or before it in the same sentence are considered as candidates.

If the attributes are independent, the probability $P(E|C)$ can be decomposed into $P(v_1|C)*...*P(v_a|C)$ and the probability to maximise is

$$P(C|E) = \frac{P(C)}{P(E)} \Pi_{j=1}^{a} P(v_j|C)$$

In that case, the classifier is a Naive Bayes Classifier (NBC)[3].

For any pronoun occurrence p of *Test_ Corpus* and for each couple to which it belongs, the Bayesian classifier computes the probability for that couple to belong to the class *Antecedent*. If the pronoun occurrence is anaphoric, the candidate with the highest probability is chosen as antecedent.

3.2 Inferring from imperfect attributes

A Bayesian Network is a model designed for reasoning on dubious and incomplete attributes. It is composed of a qualitative description of the attribute dependancies, an oriented acyclic graph, and of a quantitative description, a set of conditional probability tables, each random variable (RV) being associated to a graph node. A first parameterising step associates *a priori* conditional probability tables to each RV. The second inferring step modifies the RV values on the basis of corpus evidence (it updates the *a priori* probabilities into *a posteriori* ones). The observations made in corpus are propagated through the network, which leads to update the *a priori* values even for some unobserved variables.

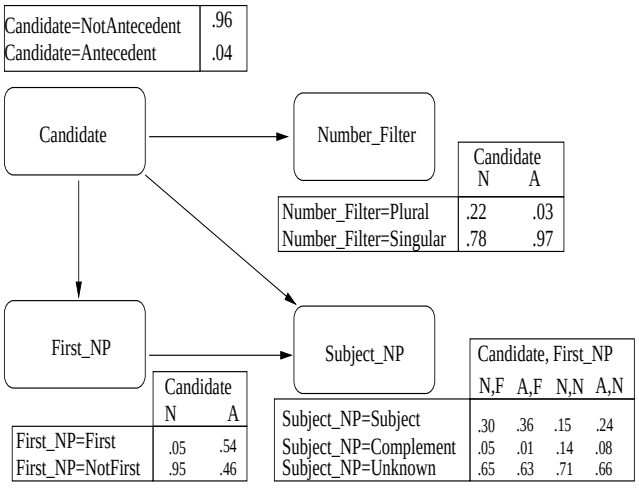

Fig. 1: *Example of a Bayesian classifier represented by a Bayesian Network*

Let us explain on a simplified example the inferring mechanism of the Bayesian Network represented on Figure 1. This network chooses the pronoun antecedent by ordering the various couples (p_i, np_j). It is composed of 4 nodes, which respectively represent the probability for the candidate np_j to be the antecedent of p_i(`Candidate`), to have some morphological properties regarding number (`Number_Filter`), to be the first NP(`First_NP`) or subject (`Subject_NP`) of the sentence.

The first prameterising step computes the *a priori* probability values. These probabilities are estimated on the basis of the frequencies computed on the set of couple examples extracted form a training corpus, for which all the attribute values are instantiated. From these observations, we state for instance that P(Candidate=Antecedent)=0.04 *i.e.* we consider that any candidate has *a priori* a probability of 4% to be the antecedent of an anaphoric pronoun occurrence[4].

The influence link between the variables `Candidate` and `Number_Filter` indicates that a candidate is less likely to be plural if it is the antecedent of the pronoun *it* (reversely, it is less likely to be its antecedent if it is a plural noun). In the same manner, the links between the variable `Candidate` and `First_NP` on the one hand, `Candidate` and `Subject_NP` on the other hand respectively indicate that the candidate is more likely to be the first NP of the preceding sentence and to be the subject of the verb if it is the pronoun antecedent. The link (`First_NP,Subject_NP`) connects two variables that are considered as dependant on each other on the basis of the training corpus and expert estimation. This means that the reliability of the subject syntactic role is increased if the candidate also occurs at the beginning of a sentence. This interdependency is measured through the table of conditional probabilities that is associated to the node `Subject_NP` on Figure 1. We also added a value *Unknown* to the RV of the `Subject_NP` node as the syntactic analysis quite often fails to associate a grammatical role to some NPs. This is a way to avoid to take into account incomplete data for the first evaluation of our system [4].

Once all the *a priori* conditional probabilites have been computed, the inferring step begins. Let's take as an example the couple (*citA transcription*, it_1) extracted from the sentence *In minimal medium, [citA transcription]$_1$ was about 6-fold lower when glucose was the sole carbon source than [it]$_1$ was when succinate was the carbon source.* Our system computes the values of the attributes of that couple. The candidate is not a plural NP but it is the first NP of the sentence. Since these observations are very reliable, we can state that P(Number_Filter=Singular)=1 and P(First_NP=First)=1 (strong evidence). Even if the parser has produced a dependancy analysis of that sentence in which the candidate is the subject of the verb, we know that this analysis may be erroneous and we consider that this third observation is only a soft-evidence: P(Subject_NP=Subject)=0.89

On the basis of these observations, the probability for the candidate to be the pronoun antecedent can be computed:

P(Candidate=Antecedent|Number_Filter=Singular, First_NP=First, Subject_NP=Subject) = 0.4

Our system similarly computes the probability for

[3] If this link is erased, the classifier becomes a naive Bayesian classifier. More generally, a Bayesian Network which structure, which structure is a tree of depth 1 and without any link between leaves is a Naive Bayesian classifier.

[4] Actually a part of human expertise is combined with corpus evidence in this probability estimation because the training data set, although complete, is not fully reliable (some values may be erroneous). To lower that noise effect, we integrate an expert estimation into the *a priori* probability computation, using the *Maximum A Posteriori* approach [13].

any other NP to be the antecedent of the pronoun it_1. If none of the other candidates has a probability higher than 40%, *citA transcription* is considered to be antecedent of the pronoun.

3.3 An extensive list of classification attributes

We keep all the attributes of MARS, except the C-command constraint that is mostly useful for demonstrative pronoun anaphors (*e.g. this*) and the preferences specifically designed for the technical type of corpora on which MARS has been initially tested[5]. We also enrich that list with some additional clues clues that are relevant for salience calculus and which are used in several other systems described in the state of the art.

The following list details the various properties that are used as attributes by our classifier. Each property is modelled as a node in our Bayesian Network (see Figure 2, where MARS attributes and the additional ones are distinguished. They are respectively coloured in black and grey):

- `Gender_Filter` and `Number_Filter`: the candidate must be morphologically compatible with the pronoun occurrence.
- `Impersonal_Filter`: the candidate cannot be the antecedent of an impersonal pronoun occurrence.
- `First_NP`: the first NP of the sentence is very often the verb subject.
- `Subject_NP`: a candidate is more likely to be the antecedent if it is the verb subject than if it holds in a different syntactic role.
- `Indicative verb`: the NPs immediately following the verbs that belong to the indicative class (*analyze, check...*) are supposed to be complement of these verbs and are more salient than others. For our experiments, this class has been manually acquired from a training corpus.
- `Repeated_NP`: an NP that is repeated several times in the same paragraph of the pronoun occurrence is more likely to be salient. These repetitions are computed by counting the number of occurrences of the NP head constituent (on the basis of a simple character string comparison).
- `Heading_Candidate`: NPs occuring in a title or at the beginning of a paragraph are emphasised and are more salient.
- `Collocation_Patterns`: our system exploit some collocation patterns with order constraints (<NP/pronoun verb> or <verb NP/pronoun>, in which we consider the lemmatised form of the verbs) but also with syntactic constraints (<Subject verb> and <verb complement>). Occurrence frequencies are computed for each candidate head in each type of collocation pattern.
- `Term`: the NPs belonging to the domain terminology are considered as salient discourse elements.
- `Definite_NP`: indefinite NPs are less salient than definite ones. We consider that an NP is indefinite if it does not follow a definite, possessive or demonstrative determinant.
- `Prepositional_NP` : if an NP belong to a prepositional complement, its salience score is decreased. The prepositional complements are identified through the text constituent analysis.
- `Distance`: the candidates that are closer to the pronoun occurrence are more likely to be the antecedent.
- `Proper_Name`: the proper names are discourse salient elements. We consider as proper names all the NPs tagged as such by the POS tagger or tagged as named entities.
- `Pronoun_NP`: if the candidate is itself an anaphoric pronoun, its own antecedent is considered as a salient candidate for the new pronoun.
- `Appositive_NP`: if a candidate occurs in an appositive clause, its salience is decreased. The appositive clauses are identified as textual segments that are preceded and followed by the same or symmetric punctuation marks[6] and which contain no verb occurrence.
- `Syntactic_Parallelism`: we check that the candidate has the same syntaxic role as the pronoun occurrence.
- `Semantic_Class`: some semantic classes are more salient than others. For instance, in biological corpora, the genes are more salient than persons.
- `Semantic_Consistence`: if the candidate is a named entity, we check that it is semantically coherent with the pronoun occurrence. We list the semantic classes of the NPs occurring in the same collocation patterns as the pronoun occurrence and we check that the candidate semantic class is one of those.

4 Experiments and results

4.1 Description of the classifieurs

We have used 6 different classifiers for the anaphora resolution.

Three of them are used as baseline systems: *Random* system, which randomly chooses the antecedent among the candidate list, *First_NP* system, which systematically selects the first NP of the preceding sentence as the pronoun antecedent, and *Bio_MARS*, which is our version of Mitkov's MARS system. The solving algorithm of Bio_MARS is the same as that MARS but our system is specifically designed for genomics. The preprocessing includes the following steps: the NP list

[5] Namely, the *immediate reference* and *sequential instruction* preferences.

[6] Except for parenthesis, which are often used for acronyms in biological corpora.

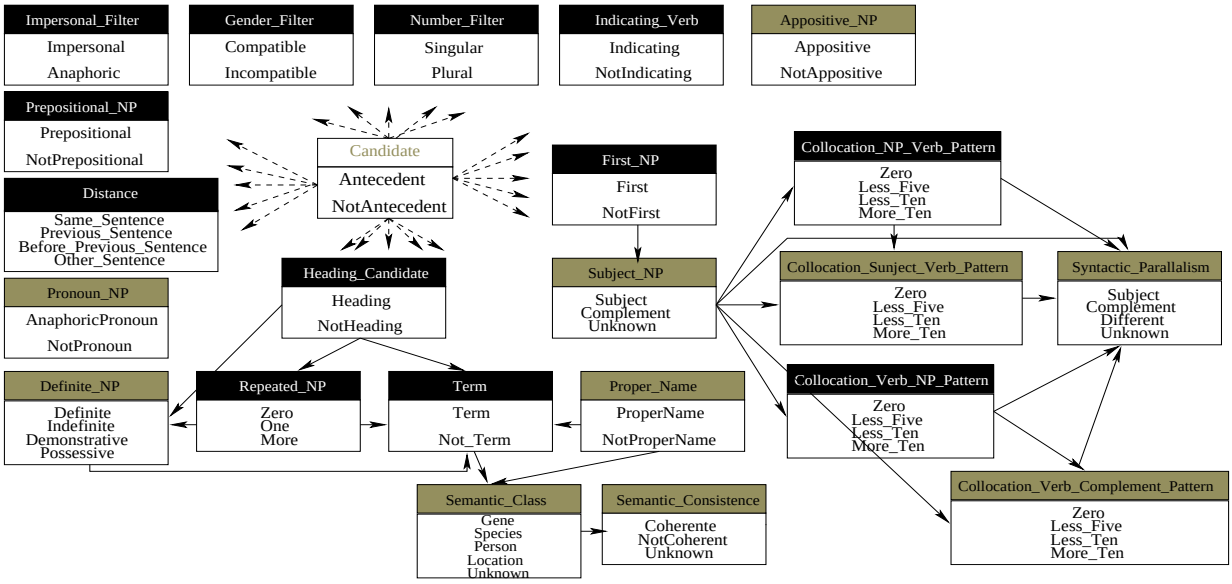

Fig. 2: *A Bayesian Network for the ranking of the antecedent candidates of the anaphoric occurrences of the pronoun* it. *The prediction node is the* `Candidat` *one, at the centre of the network. It gives the probability for a given candidate to be the antecedent of a given pronoun occurrence. It is linked to all the other network nodes.*

is extracted from a full constituent analysis of the corpus that is obtained thanks to a domain specific parser; for identifying the anaphoric occurrences, we exploit a filter that is based on a Bayesian Network and trained on a corpus of the same domain [16]; we rely on an extended and domain specific tagging of named entities and terms.

The three other systems have been designed to test various configurations of the Bayesian model. *NB_Mars* system exploits the same attributes as Bio_MARS but the final decision is based on a Naive Bayes classifier rather than on a global score. The fourth system is the Bayesian Network classifier itself (*BNC*): the choice of the attributes and the network structure are based on a linguistic analysis of a training corpus. The last system is the Naive Bayesian classifier (*NBC*), which has the same attributes as BNC but a simplified tree structure where the attributes are considered as independant of each other.

4.2 Experimental protocol

We tested our systems on a specialised corpus, *Transcript*. It is a collection of 2209 abstracts of scientific papers that have been retrieved by querying the *Medline* bibliographical base with the keywords *bacillus subtilis, transcription*[1]. 697 occurrences of *it* have been identified in Transcript (around 800,000 words). Two different annotators have tagged each of these occurrences as either anaphoric or impersonal and have identified the corefering antecedent of the anaphoric pronoun occurrences[7].

In order to determine the attribute values of each candidate/pronoun couple, we have exploited the *Ogmios* platform [6] to analyse our corpus. Ogmios integrates TagEN, a named entity tagger specifically designed for genomics, to identify the biological named entities, and BioLG, a version of Link Grammar Parser adapted for biology [12], for the dependancy and constituent syntactic analysis. It also exploits large specialised terminology. For our first experiments, we have manually built the class of indicative verbs out of our training corpus.

Since our working corpus is relatively small, we have validated our results using a cross validation method. We have randomly selected 2/3 of our corpus to compute the *a priori* conditional probabilities and we have applied the resulting parameterised systems to the remaing part of the corpus. We iterated these operations 20 times and we analysed the average performance of each classifier on our corpus.

4.3 Results

Table 1 summarises the performances of each system measured as a success rate (proportion of anaphors that have been correctly solved by the systems).

Two different measures are given for the last 6 lines: the strict and partial success rate which correspond to two different definitions of what a "correct" antecedent is. The strict success rate counts an anaphor as correctly solved only if the proposed NP exactly matches the phrase tagged as antecedent by the human annotators in the test corpus. The partial success score counts as correct an anaphor where the proposed NP only partially matches the phrases tagged as antecedent in the test corpus as soon as it can be substituted to the anaphoric pronoun without semantic inconsistence. For instance, in the sentence *[beta-Galactosidase expression from the spl-lacZ fusion] was silent during vegetative growth and was not DNA damage inducible, but [it] was activated at morphological stage III...* our system gives only *beta-Galactosidase expression* as antecedent instead of the whole NP but it can nevertheless be substitued to *it* : it is consid-

[7] Since the second annotator has not finished, no agreement rate can be given yet.

ered as a correct partial resolution but not as a correct strict one.

Since there are some errors in the input NP list[8], the anaphora resolution performance cannot reach 100% and the last row (MAX) gives the highest reachable resolution score for comparison.

System	Results	
	Strict	*Partial*
Random	6%	-
First_NP	36.3%	51%
Bio_MARS	26.7%	43%
NB_MARS	39.9%	56%
Naive Bayes Classifior	43.1%	59%
Bayesian Network Classifior	44.0%	61%
MAX	93.3%	97.8%

Table 1: *Anaphora Resolution Results (Success rate)*

5 Discussion

5.1 The importance of corpus specificity

The first striking observation that can be drawn from Table 1 is taht Bio_MARS performance that is significantly lower than the success rate of First_NP system on our corpus and also lower than the 45.81% score obtained by MARS on a different corpus made of technical manuals [10]. Most of the cases that are correctly solved by First_NP system and not by Bio_MARS involve the terminological and collocation pattern attributes that are not sufficiently discriminating in our domain[9]: our platform tags many terms which are not all salient elements (e.g. *use*, *work*) and the collocation patterns have a weight to high to be compensated by other observations. In the probabilistic version of Bio_MARS (NB_MARS), the parameterising step adapts these scores for our corpus and therefore avoids the previous errors.

5.2 The complementarity between linguistic constraints and surface clues

Comparing the systems NB_MARS and BNC shows the importance of the complex linguistic constraints in the resolution process, even if the corresponding attributes are not fully reliable. These additional attributes help to distinguish among various candidates. Let us consider for instance the following sentences extracted from our corpus *[A grpE heat-shock gene]$_1$ was found by sequencing in [the genome of the methanogenic archaeon Methanosarcina mazei S-6]$_2$. [It]$_1$ is the first example of grpE from the phylogenetic domain Archaea.* NB_MARS gives the same probability for the candidates 1 and 2 and finally chooses the candidate 2, using the heuristic of the most recent candidate. BNC classifier avoids this error: it exploits the syntactic role of the candidate 1 (subject) and its semantic type (*gene*), which increases the candidate probability to 0.73 and solves the ambiguity.

If surface clues are not always sufficient to decide between the candidates, their role is nevertheless important to correct the imperfectness of linguistic information. For instance, the syntactic and named entity information are not reliable enough to be used in isolation. BioLG parser has a fairly good precision (86%) but a low recall (55%) and the results of the named entity tagging are noisy (71% of gene names are identified but only 68% of the tagged entities are really gene names, due to ambiguous gene names such as *not, All, similar*).

It is important to understand how the linguistic properties and the surface clues complement each other. In BNC system, these complementarity is represented and measured by the interdependency links taht hold between two network nodes. These links express a set of reinforcement or invalidation constraints. NBC system, which does have such constraints, overestimates the attribute weights. NBC often puts the correct antecedent in the second or third position in the candidate list, whereas BNC chooses the correct candidate.

5.3 The limits of the salience-based approach

A detailed manual analysis of the BNC errors shows the limits of the salience-based approach. 47% of the errors produced by BNC are due to an erroneous calculus of the salient element. BNC not finds the element that is intuitively identified as the most salient by the human judge because a less salient element ends with a higher salient score than the actual antedent.

In 21% of the cases, BNC actually finds the salient element but it is not the pronoun antecedent. For instance, in the sentence *[Amino acid sequence analysis]$_1$ of [the 33-kDa protein]$_2$ revealed that it is a sigma factor, sigma E.*, the most salient element is candidate 1 which is erroneously preferred to the candidate 2. Solving such anaphors would call for more complex semantic and domain knowledge that would help to analyse the semantic compatibility between the candidate 2 and the pronoun occurrence [9].

The remaining errors are due to the corpus imperfect preprocessing (word segmentation errors and unidentified NPs) rather than to the resolution strategy itself.

5.4 The role of the various types of knowledge

In order to evaluate the contribution of the various attributes in the resolution process, we set up additional experiments, based on the same protocol.

[8] BioLG does not parse sentences that are more than 70 words long or that do not contain any verb. When there is no parse available, we create a list of NPs on the basis of the POS-Tagging.

[9] Our model allows to quantify this fact: $P(Term = Term|Candidate = Antecedent) = 0.16$, $P(Collocation\_NP\_Verb\_Pattern = Less\_Five, Less\_Ten, More\_Ten|Antecedent) = 0.08$, $P(Collocation\_Verb\_NP\_Pattern = Less\_Five, Less\_Ten, More\_Ten|Antecedent) = 0.01$.

Systems	Results	
BNC without...	*Strict*	*Partial*
syntactic information	42.8	59%
terminological information	44.0	61%
collocation information	42.1	58%
semantic information	41.9	58%
any surface clue	23.0	31%
Complete BNC	44.0%	61%

Table 2: *Role of various types of knowledge in various variants of BNC system (Success rate)*

Table 2 compares the performances of various variants of BNC system. The complete BNC is the one that has been described above. The five first variants are identical to that one except that a specific set of variables has been omitted in each variant to test the respective role of various types of knowledge[10].

As expected in the previous subsections, the low score of the classifier without any surface clue shows that it is impossible to exploit the linguistic variables alone. Among linguistic variables, the semantic ones have the stronger impact on the classifier decisions. A precise semantic tagging is an important factor of success for solving anaphors in a corpus such as ours. As opposed to the conclusions of [2], we observe that collocation patterns are useful indicators when they are corroborated by other clues. The terminological variable is the only one that has no impact on resolution (the BNC has the same score with or without terminological information). Finally, it is interesting to note that the contribution of syntactic information is relatively low although additional experiments on a more completely parsed corpus are necessary to really evaluate the impact of syntax in anaphora resolution.

6 Conclusion

In this paper, we have tried to show how interesting the Bayesian Network formalism is for NLP tasks, taking the complex problem of pronominal anaphora resolution as an example. This model allows to overcome the traditional opposition between systems based on linguistic knowledge and knowledge-poor systems. It appears that both approaches should rather be combined than opposed: linguistic knowledge is necessary but often lacking and usually not fully reliable; surface clues are easy to measure but fail to solve some ambiguities. By unifying both types of knowledge in a single representation, the Bayesian Network approach enables to exploit some information pieces to reinforce, invalidate or supplement others. This gives interesting results on the anaphora resolution task, in comparison with a state of the art system.

[10] The *Subject_NP, Collocation_Subject_Verb_Pattern and Collocation_Verb_Complement_Pattern* variables are omitted in the first variant (without syntactic information). The *Term* variable is omitted for the second one. The 4 collocational variables are erased for the third variant. The *Semantic_Class and Semantic_Coherence* variables are omitted for the fourth variant (without semantic information). The fifth one does not take any surface clue variable into account.

Our system can be further improved. We want to extend the set of clues that are exploited for anaphora resolution. For the moment, it only relies on the search of the most salient element to choose the pronoun antecedent and we have shown that this strategy sometimes fails. Our Bayesian Network can be enriched by integrating focused-based information [15]. We also want to test the possibility to learn the network structure from a training corpus, instead of relying of linguistic expertise. Our first tests show that some nodes seem to be useless. Finally, we would like to take into account the fact that the various candidate scores are not independent of each others. Actually, the choice of a candidate not only depends on the intrinsic properties of that candidate but also of alternative ones. This should lead us to exploit a specific extension of Bayesian Networks, the dynamic Bayesian Networks.

References

[1] E. Alphonse, S. Aubin, P. Bessières, G. Bisson, T. Hamon, S. Lagarrigue, A. Nazarenko, A.-P. Manine, C. Nédellec, M. Vetah, T. Poibeau, and D. Weissenbacher. Event-based information extraction for the biomedical domain: the caderige project. In *Proceedings on International Joint Workshop on Natural Language Processing in Biomedecine and its Applications (BioNLP/LNPBA), COLING'04*, pages 43–49, 2004.

[2] L. T. Andrew. Kehler, Douglas. Appelt and A. Simma. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference*, pages 289–296, 2001.

[3] R. Bouckaert. Low level information extraction, a bayesian network based approach. In *Workshop on Text Learning (TextML-2002)*, 2002.

[4] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Experts Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[5] I. Dagan and A. Itai. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of COLING'90*, volume 3, pages 330–332, 1990.

[6] J. Derivière, T. Hamon, and A. Nazarenko. A scalable and distributed nlp architecture for web document annotation. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, pages 56–67, 2006.

[7] C. Kennedy and B. Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of COLING'96*, 1996.

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

[9] Y.-H. Lin and T. Liang. Pronominal and sortal anaphora resolution for biomedical litterature. In *Proceedings of ROCLING XVI*, pages 101–110, 204.

[10] R. Mitkov. *Anaphora Resolution*. Longman, 2002.

[11] S. Ponzetto and M. Strube. Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of EACL'06*, pages 143–146, 2006.

[12] S. Pyysalo, T. Salakoski, S. Aubin, and A. Nazarenko. Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl 3), November 2006.

[13] C. Robert. *The bayesian Choice: a decision-theoretic motivation*. Springer, 1994.

[14] S. L. Ruslan. Mitkov and B. Boguraev. Introduction to the special issue on computational anaphora resolution. *Computational Linguistics*, 27(4):473–477, 2001.

[15] M. Strube. Never look back: An alternative to centering. In *COLING-ACL*, pages 1251–1257, 1998.

[16] D. Weissenbacher. Bayesian network, a model for nlp? In *Companion Volume of the Proceedings of EACL'06*, pages 195–198, 2006.