Bayesian Network, a model for NLP?

Davy Weissenbacher
Laboratoire d'Informatique de Paris-Nord
Universite Paris-Nord
Villetaneuse, FRANCE
davy.weissenbacher@lipn.univ-paris13.fr

Abstract

The NLP systems often have low performances because they rely on unreliable and heterogeneous knowledge. We show on the task of non-anaphoric *it* identification how to overcome these handicaps with the Bayesian Network (BN) formalism. The first results are very encouraging compared with the state-of-the-art systems.

1 Introduction

When a pronoun refers to a linguistic expression previously introduced in the text, it is anaphoric. In the sentence *Nonexpression of the locus even when it is present suggests that these chromosomes[...]*, the pronoun *it* refers to the referent designated as '*the locus*'. When it does not refer to any referent, as in the sentence *Thus, it is not unexpected that this versatile cellular...* the pronoun is semantically empty or non-anaphoric. Any anaphora resolution system starts by identifying the pronoun occurrences and distinguishing the anaphoric and non-anaphoric occurrences of *it*.

The first systems that tackled this classification problem were based either on manually written rules or on the automatic learning of relevant surface clues. Whatever strategy is used, these systems see their performances limited by the quality of knowledge they exploit, which is usually only partially reliable and heterogeneous.

This article describes a new approach to go beyond the limits of traditional systems. This approach stands on the formalism, still little exploited for NLP, of Bayesian Network (BN). As a probabilistic formalism, it offers a great expression capacity to integrate heterogeneous knowledge in a single representation (Peshkin, 2003) as well as an elegant mechanism to take into account an *a priori* estimation of their reliability in the classification decision (Roth, 2002). In order to validate our approach we carried out various experiments on a corpus made up of abtsracts of genomic articles.

Section 2 presents the state of the art for the automatic recognition of the non-anaphoric occurences of *it*. Our BN-based approach is exposed in section 3. The experiments are reported in section 4, and results are discussed in section 5.

2 Identification of Non-anaphoric *it* occurences

The decisions made by NLP systems depend on the available knowledge. However this information is often weakly reliable and leads to erroneous or incomplete results.

One of first pronoun classifier system is presented by (Paice, 1987). It relies on a set of logical first order rules to distinguish the non-anaphoric occurences of the pronoun *it*. Non-anaphoric sequences share remarkable forms (they start with an *it* and end with a delimiter like *to, that, whether...*). The rules expresses some constraints which vary according to the delimiter. They concern the left context of the pronoun (it should not be immediately preceded by certain words like *before, from to*), the distance between the pronoun and the delimiter (it must be shorter than 25 words long), and finally the lexical items occurring between the pronoun and the delimiter (the sequence must or must not contain certain words belonging to specific sets, such as words expressing modality over the sentence content, *e.g. certain, known, unclear...*). Tests performed by Paice show good results with

91.4%Accuracy[1] on a technical corpus. However the performances are degraded if one applies them to corpora of different natures: the number of false positive increases.

In order to avoid this pitfall, (Lappin, 1994) proposes some more constrained rules in the form of finite state automata. Based on linguistic knowledge the automata recognize specific sequences like *It is not/may be<Modaladj>*; *It is <Cogv-ed> that <Subject>* where *<Modaladj>* and *<Cogv>* are modal adjective and cognitive verbs classes known to introduce non-anaphoric *it* (*e.g. necessary, possible* and *recommend, think*). This system has a good precision (few false positive cases), but has a low recall (many false negative cases). Any sequence with a variation is ignored by the automata and it is difficult to get exhaustive adjective and verb semantic classes[2]. In the next paragraphs we refer to Lappin rules' as Highly Constraint rules (HC rules) and Paice rules' as Lightly Constraint rules (LC rules).

(Evans, 2001) gives up the constraints brought into play by these rules and proposes a machine learning approach based on surface clues. The training determines the relative weight of the various corpus clues. Evans considers 35 syntactic and contextual surface clues (*e.g.* pronoun position in the sentence, lemma of the following verb) on a manually annotated sample. The system classifies the new *it* occurences by the k-nearest neighbor method metric. The first tests achieve a satisfactory score: 71.31%Acc on a general language corpus. (Clement, 2004) carries out a similar test in the genomic domain. He reduces the number of Evans's surface clues to the 21 most relevant ones and classifies the new instances with a Support Vector Machine(SVM). It obtains 92.71%Acc to be compared with a 90.78%Acc score for the LC rules on the same corpus. The difficulty, however, comes from the fact that the information on which the systems are built is often diverse and heterogeneous. This system is based on atomic surface clues only and does not make use of the linguistic knowledge or the relational information that the constraints of the previous systems encode. We argue that these three types of knowledge that are the HC rules, the LC rules, and the surfaces clues are all relevant and complementary for the task and that they must be unified in a single representation.

3 A Bayesian Network Based System

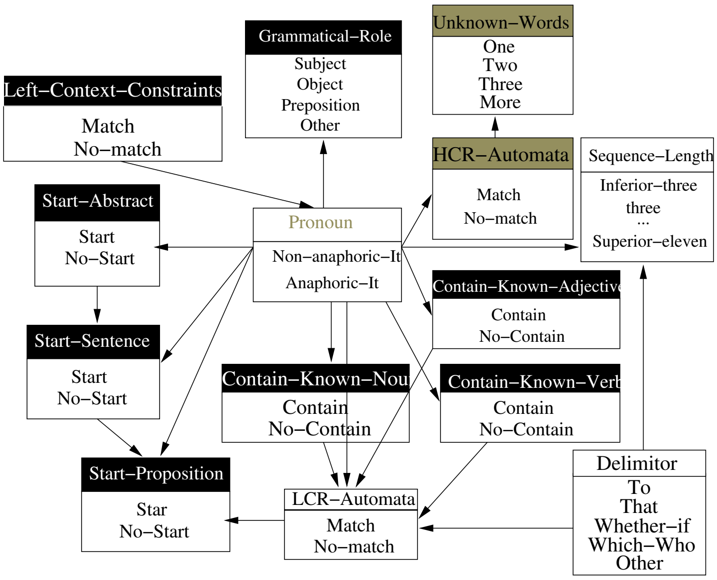

Figure 1: A Bayesian Network for identification ofnon-anaphoric *it* occurrences

Neither the surface clues nor the surface clues are reliable indicators of the pronoun status. They encode heterogeneous pieces of information and consequently produce different false negative and positive cases. The HC rules have a good precision but tag only few pronouns. On the opposite, the LC rules, which have a good recall, are not precise enough to be exploited as such and the additional surface clues must be checked. Our model combines these clues and take their respective reliability in to account. It obtains better results than those obtained from each clue exploited separately.

The BN is a model designed to deal with dubious pieces of information. It is based on a qualitative description of their dependancy relationships, a directed acyclic graph, and a set of conditionnal probablities, each node being represented as a Random Variable (RV). Parametrizing the BN associates an *a priori* probability distribution to

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

[2] For instance in the sentences *It is well documented that treatment of serum-grown...* and *It is generally accepted that Bcl-2 exerts...* the *it* occurences are not classified as non-anaphorics because *documented* does not belong to the original verb class *<Cogv>* and *generally* does not appear in the previous automaton.

the graph. Exploiting the BN (inference stage) consists in propagating new pieces of information through the network edges and updating them according to observations (*a posteriori* probabilities).

We integrated all the clues exploted by of the previous methods within the same BN. We use dependancy relationships to express the fact that two clues are combined. The BN is manually designed (choice of the RV values and graph structure). On the Figure1, the nodes associated with the HC rules method are marked in grey, white is for the LC rules method and black for the Clement's method[3]. The Pronoun node estimates the decision probability for a given *it* occurence to be non-anaphoric.

The parameterising stage establishes the *a priori* probability values for all possible RV by simple frequency counts in a training corpus. They express the weight of each piece of information in the decision, its *a priori* reliability in the classification decision[4]. The inference stage exploits the relationships for the propagation of the information and the BN operates by information reinforcement to label a pronoun. We applied all precedent rules and checked surface clues on the sequence containing the *it* occurrence and set observation values to the correspondant RV probabilities. A new probability is computed for the node's variable Pronoun: if it is superior or equal to 50% the pronoun is labeled non-anaphoric, anaphoric otherwise.

Let us consider the sentence extracted from our corpus: *It had previously been thought that ZEBRA's capacity to disrupt EBV latency...*. No HC rule recognizes the sequence even by tolerating 3 unknown words [5], but a LC rule matches it with 4 words between the pronoun and the delimiter *that*[6]. Among the surface clues, we checked that the sequence is at the beginning of the sentence (1) but that the sentence is not the first of the abstract (2). The sentence also contains the adverb *previously* (3) and the verb *think* (4), which words belong to our semantic classes[7]. The *a priori* probability for the pronoun to be non-anaphoric 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 classifies it as non-anaphoric.

Table 1: Prediction Results (Accuracy/False Positive Cases/False Negatives Cases)

Method	Results
Highly Constraint Rules	88.11% / 12.8 / 169.1
Lightly Constraint Rules	88.88% / 123.6 / 24.2
Support Vector Machine	92.71% / - / -
Naive Bayesian Classifier	92.58% / 74.1 / 19.5
Bayesan Network	95.91% / 21.0 / 38.2

4 Experiments and Discussion

Medline is a database specialized in genomic research articles. We extracted from it 11966 abstracts with keywords *bacillus subtilis, transcription factors, Human, blood cells, gene and fusion*. Among these abstracts, we isolated 3347 occurences of the pronoun *it* and two human annotators tagged *it* occurences as either anaphoric or non-anaphoric[8]. After discussion, the two annotators achieved a total agreement.

We implemented the HC rules, LC rules and surface clues using finite transducers and extracted the pronoun syntactic role from the results of the Link Parser analysis of the corpus (Aubin, 2005). As a working approximation, we automaticaly generated the verb, adjective and noun classes from the training corpus: among all *it* occurences tagged as non-anaphoric, we selected the verbs, adjectives and nouns occurring between the delimiter and the pronoun. We considered a third of the corpus for training and the remaining for testing. Our experiment was performed using 20-cross validation.

Table1 summarizes the average results reached

[3] Only significant surface clues for our modelisation have been added to the BN.

[4] Among the 2000 *it* occurences of our training corpus (see next section), the HC rules recognized 649 of the 727 non-anaphoric pronouns and they have erroneously recognized as non-anaphoric 17 pronouns, so we set the HCR-rules node probabilities as P(HCR-rules=Match|Pronoun=Non-Anaphoric)=89.2% and P(HCR-rules=Match|Pronoun=Anaphoric)=1.3% which expresses the expected value for the false negative cases and the false positive cases produced by the HC rules respectively.

[5] So we set P(HC-rules = No-match)=1 and P(Unknown-Words = More)=1.

[6] We set P(LC-rules = Match)=1, P(Sequence-Length = four)=1 and P(Delimitor = That)=1.

[7] Others node values are set consequently.

[8] Corpus is available at http://www-lipn.univ-paris13.fr/˜weissenbacher/

by the state-of-the-art methods described above[9]. The BN system achieved a better classification than other methods.

In order to neutralize and comparatively quantify the contribution in the decision of the dependancy relationships between the factors, we have implemented a Naive Bayesian Classifier (NBC) which exploits the same pieces of knowledge and the same parameters as the BN but it does not profit from reinforcement mechanism, which leads to a rise in the number of false positive cases.

Our BN, which has a good precision, nevertheless tags as non-anaphoric some occurrences which are not. The most recurrent error corresponds to the sequences ending with a delimiter *to* recognized by some LC rules. Although none HC rule matches the sequence, its minimal length and the fact that it contains particular adjectives or verbs like *assumed* or *shown*, makes this configuration caracteristic enough to tag the pronoun as non-anaphoric. When the delimiter is *that*, this classification is correct [10] but it is always incorrect when the delimiter is *to*[11]. For the delimiter *to*, the rules must be more carefully designed.

Three different factors explain the false negative cases. Firstly, some sequences were ignored because the delimiter remained implicit[12]. Secondly, the presence of apposition clauses increases the sequence length and decreases the confidence. Dedicated algorithms taking advantage of a deeper syntactic analysis could resolve these cases. The last cause is the non-exhaustiveness of the verb, adjective and noun classes. It should be possible to enrich them automatically. In our experiments we have noticed that if a LC rule matches a sequence in the first clause of the first sentence in the abstract then the pronoun is non-anaphoric. 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.

5 Conclusion

Our system can of course be enhanced along the previous axes. However, it is interesting to note that it achieves better results than the comparable state-of-the art systems, although it relies on the same set of rules and surface clues. This comparison confirms the fact that the BN model proposes an interesting way to combine the various clues, some of then being only partially reliable. We are continuing our work and expect to confirm the contribution of BN to NLP problems on a task which is more complex than the classification of *it* occurences: the resolution of anaphora.

[9] We have completed the Clement's SVM score for the same biological corpus to compare its results with ours.

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

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

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

References

S. Aubin, A. Nazarenko and C. Nedellec. 2005. *Adapting a General Parser to a Sublanguage. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, 1:89–93.

L. Clemente, K. Satou and K. Torisawa. 2004. *Improving the Identification of Non-anaphoric It Using Support Vector Machines. Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 1:58–61.

I. Dagan and A. Itai. 1990. *Automatic Processing of Large Corpora for the Resolution of Anaphora References. Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, 3:1–3.

R. Evans. 2001. *Applying Machine Learning Toward an Automatic Classification of it. Literary and linguistic computing*, 16:45–57.

S. Lappin and H.J. Leass. 1994. *An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics*, 20(4):535–561.

C.D. Paice and G.D. Husk. 1987. *Towards the Automatic Recognition of Anaphoric Features in English Text: the Impersonal Pronoun It. Computer Speech and Language*, 2:109–132.

L. Peshkin and A. Pfeffer 2003. *Bayesian Information Extraction Network. In Proc.18th Int. Joint Conf. Artifical Intelligence*, 421–426.

D. Roth and Y. Wen-tau. 2002. *Probalistic Reasoning for Entity and Relation Recognition. Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, 1:1–7.