Pronominal anaphora resolution using XML tagged documents

Allaoua Refoufi
Computer Science Department Faculty of Sciences
University of Serif, Algeria

***Abstract*--Anaphora resolution has become a major issue in NLP systems; in this work we propose a resolution approach in which texts are parsed by a definite clause grammar and then converted into an XML tagged representation, where sentence elements are marked with discourse, syntactic, and semantic attributes. This extension was made primarily to test the viability of using XML tagged documents for anaphora resolution. The XML representation allows valuable text's enrichment with anaphoric information in an elegant and easy way. The system's performance arises primarily from the integration of multiple knowledge sources in a modular architecture and uses constraints and preferences to select the antecedent. The developed system proposes to resolve pronominal anaphora, namely personal pronouns.**

Keywords--- anaphora, constraints, preferences, XML, Definite Clause Grammars

I. INTRODUCTION

Anaphora is a complex phenomenon in natural language communication and no truly comprehensive computational approaches to anaphora resolution have been proposed. Finding the appropriate referent has long been recognized as a difficult problem requiring syntactic, semantic, as well as pragmatic knowledge. The interpretation of many natural language expressions depends on the context (the previous utterances and their content); in particular the interpretation of pronouns (personal, demonstrative and possessive) depends on entities introduced in the previous linguistic context. We will use the term anaphora ("act of carrying back") to indicate expressions that depend on the linguistic context, i.e., on objects explicitly mentioned or objects whose existence can be inferred from what has been said. The term reference is used throughout this paper to indicate the relation between an expression of the language and an object in the world.

Our approach to anaphora resolution consists of three main components: A parser based on Definite Clause Grammars, an XML generator which represents the output of the parser, and the anaphora resolution mechanism. Extensible Markup Language (XML) was chosen as the underlying representational mechanism, primarily because it provides a more natural vehicle for retaining the tree structure produced in parsed sentences. XML also provides a convenient mechanism for extracting, attributes and annotations attached to parsed tree nodes. In what follows we introduce the notion of anaphora and present our motivation or using XML representation.

There are various forms of referring pronouns, which differ primarily according to the rules that govern their anaphoric behavior, and their use. In this paper we will concentrate on methods for identifying the referent of the following types of pronouns (see Table1 for the complete set of pronouns treated):

- Subject pronouns (« il », « elle », « ils », « elles»)
- Direct Object pronouns (« le », « la », « les »)
- Indirect Object Pronoun (« lui », « leur »)
- Demonstrative pronouns (« celui », « celle », « ceux ») such as in the sentence : « Je vous donne celle de ma fille ».
- Relative pronouns (« qui », « que », « quoi », « dont », « où »).

In what follows we present and define anaphora resolution (also known in Computational Linguistics as coreference resolution) and discuss why XML tagging is useful to such analysis.

1. What is anaphora resolution?

"Anaphora, in discourse, is a device for making an abbreviated reference to some entity (or entities) in the expectation that the receiver of the discourse will be able to disabbreviate the reference and, thereby, determine the identity of the entity". (Hirst, [04]). Pronominal resolution refers to identification of a definite noun phrase (DNP) that is referred by a pronominal. A definite noun phrase (DNP) is a noun phrase preceded by a definite determinant (" la jeune femme"), a proper name ("Sarah"), or a named entity.

Non Pronominal resolution refers to identification of a noun phrase (NP) referring to other NPs. It is the aim of Natural Language Processing (NLP) to recognize those mechanisms, while producing intelligible and coherent information. The scope of this work, Anaphora Resolution, is the identification of a pronoun or a noun phrase that functions as a regular grammatical substitute for a preceding noun phrase. When the anaphor and its antecedent have the same referent in the real world, like the previous examples, they are termed coreferential. This form of anaphora occurs as definite noun phrases and proper names. Anaphora is a complex phenomenon in natural language communication and no truly comprehensive computational approaches to anaphora resolution have been proposed. Finding the appropriate referent has long been recognized as a difficult problem requiring syntactic, semantic, as well as pragmatic knowledge. It requires access and integration of all the knowledge sources necessary for dialog and text interpretation. These linguistic knowledge sources are brought to bear as constraints and preferences encoded in multiple resolution strategies. The Process of finding the antecedent for an anaphor is termed anaphora resolution. The concept anaphora relates to the reference that points to the previous item, antecedent is the entity to which the anaphor refers.

There are many aspects and types of anaphora, the most common type occur when the anaphor is a pronoun. Personal, possessive and demonstrative pronouns both singular and plural can also function as pronominal anaphora. First and second personal (“je”, “tu”, ”nous”, ”vous”) pronouns, singular or plural usually refer to the dialog interlocutors, thus these pronouns do not establish a coreference relation between elements present in the analyzed sentences; consequently there are not taken care of in this paper. Relative pronouns (**qui, que, quoi, dont, où)** on the other hand, have as their antecedent the immediate noun, being modified by the relative clause, for instance: “il y avait un enfant qui portait un gros bouquet ».

However all pronouns do not refer back to an entity. For example in the sentence “il pleut aujourd’hui”, the pronoun “il” does not refer to any entity. Such an instance of “il” is called pleonastic “il”. Before we do pronominal resolution, we need to identify and discard such pronouns. It is clear that non-anaphoric pronouns identification improves the precision of the resolution mechanism. Non anaphoric expressions are signaled by expressions of the form « il est {possible, évident, admis, normal, permis, logique, courant, etc.} que S » and indicate that the pronoun « il » is not anaphoric. In the appendix we provide more examples where the pronoun "il” is pleonastic. The identification of non anaphoric pronouns is based a simple pattern matching procedure. The number of template matching used to identify such constructions can be updated or augmented.

Because anaphors typically refer back to other constituents in the same sentence, or to constituents in earlier utterances in the discourse, they can be classified as intrasentential or intersentential, according to the location of the antecedent. If the anaphor and its antecedent occur in the same sentence, it is called intrasentential anaphora. If they are located in different sentences, it is called intersentential anaphora. Syntactic information plays a major role in establishing appropriate referents for the former case, intrasentential anaphora. However semantic and pragmatic information is absolutely required for the latter case, intersentential anaphora. Empirical studies have shown that intersentential anaphora is far more frequent and more crucial in the design of interactive natural language interfaces. This paper addresses the problem of both intrasentential and intersentential anaphora resolution integrating multiple knowledge sources.

2. Why XML tagged representation?

Recent research ([09]) presented results which suggested that exploiting Extensible Markup Language (XML) tagged documents could potentially be useful in such natural language processing applications as anaphora resolution. Moreover, the performance of computational linguistics Research’s Knowledge Management System in DUC 2005 corroborates the conjecture in [09] that XML-tagged documents provide a useful basis for anaphora resolution. Ideally annotating corpora, a mechanism of how to encode linguistic features in a text, is highly needed to conduct anaphora resolution. However, the lack of such corpora was our first motivation to use the XML representation.

Parting from these observations we plan to conceive an anaphora resolution system from the output of a parser which is converted into an XML tagged document. We believe that the XML tagged document offers more flexibility in the search for antecedents in the anaphora resolution system.

Our past experiences from NLP applications have shown that the XML structures representing the output of the parser allow an easy to handle examination and manipulation of low level tree nodes. While the output of the syntactic parser (the syntactic tree in list format) is more difficult to handle during the analysis process which follows. Furthermore the XML format of the text allows a more detailed examination of word frequency and the kinds of syntactic and semantic relations present in the documents. A valid XML document is a tree and the entire representation can readily be designed on this tree structure. The entire collection of the output of the parser can be represented as one tree.

II.APPROACHES TO ANAPHORA RESOLUTION

There are various approaches to anaphora resolution: syntax based approaches operate on the rules and principles that control sentence structure, usually represented by syntactic trees. In Syntax based models syntactic information plays an important role both in filtering certain types of interpretation (gender, binding constraints) and in determining preferred interpretations (subject assignment, parallelism). Several algorithms have been developed that incorporate these types of syntactic knowledge for anaphora resolution, in particular for the resolution of pronouns. On the other hand an approach based on shallow processing would have to approximate the syntax-based constraints based on the information in partial parses, and use a heuristic approach to reach full coverage for gender determination.

Machine learning approaches gave the possibility of acquiring this information automatically. They use a set of patterns to extract knowledge from raw or annotated corpora and use it to produce decision trees. Statistical approaches process large amounts of annotated corpora analyzing the occurrence of anaphors and its candidates, regarding its morpho-syntactic characteristics and semantic roles.

State of the art anaphora resolution systems all make use of some kind of knowledge to quantify the degree of semantic compatibility holding between the anaphor and the antecedent. Much of the early linguistic work on anaphora focused on the identification of morphological and syntactic constraints on the interpretation of anaphoric expressions. Among these constrains the better known are agreement constraints (syntactic and semantic) and binding constraints. In order to apply different strategies we need to make a distinction between constraints (which must not be violated) and preferences (which discriminate among candidates satisfying all constraints). Preferences are applied in a predefined order.

2.1 Linguistic constraints

Linguistic constraints are divided into morphological and syntactic constraints; we give a brief description in what follows.

2.1.1 Morphological constraints (also known as agreement constraints): this constraint states that the anaphor and the antecedent must agree in person, number and gender. This is a strong constraint in the French language due to the high number of flectional forms of adjectives and verbs.

Syntactic Constraints: many existing systems use Chomsky's Binding Theory to define syntactic constraints for filtering invalid candidates with respects to c-command and local domain. The traditional definition of a local domain of a constituent C is defined as the set of constituents contained in the closest S (sentence) or NP (noun phrase) that contains C ([01], [15]). Syntactic constraints state that a pronoun cannot refer to a c-commanded noun phrase within the same local domain. For example in the sentence " la jeune femme l'admire", the noun phrase "la jeune femme" cannot be a candidate for the pronoun "l". This constraint has been used in anaphora resolution to narrow down the search scope of candidates for antecedents.

2.1.2 Linguistic and syntactic constraints by themselves do not completely eliminate anaphoric ambiguity. When ambiguities persist we add some other factors termed preferences to discriminate among them.

Preferences

Obvious factors playing a role in anaphora resolution are the position of the pronoun as well as the antecedent in the syntactic structure of the sentence. Corpus statistics suggest that in most French corpora, about 65-75% of pronouns occur in subject position, and of these, around 70% have an antecedent also realized in subject position. This preference for pronouns in subject position to refer to antecedents in subject position has been called grammatical role. Another factor that clearly plays a role in anaphora resolution is salience, at least in its simplest form known as recency: generally speaking, more recently introduced entities are more likely antecedents (Miltsakaki, [10]).

Yet other factors may include first mention advantage: a preference to refer to first mentioned entities in a sentence, and focus. Focusing mechanisms exist and play an important role in the choice of an antecedent for anaphoric expressions (the global focus specifying the articulation of a discourse into segments, and the local focus of salience specifying how utterance by utterance the relative salience of entities changes). In chapter 6 we will present and discuss the preferences treated in our work.

RELATED WORK

Much research has been performed in the field of anaphora resolution and especially in the field of pronominal resolution. Works with significant importance include (Hobbs,[05], [06]), (Lappin & Leass, [08]), (Kennedy & Boguraev, [07]), (Mitkov, [11]), and (Trouilleux, [17]). Each method relies on various knowledge sources; we distinguish those that uses full syntactic parsers ([08],[18]) from those that use only poor knowledge sources ([07], [11]), which means only an output of a part of speech tagger that identifies only noun phrases and pronouns to be resolved. The first three algorithms we review are for English texts, and the fourth one for French.

A pioneer work reported by Hobbs ([05], [06]) uses a syntactic tree to search the input sentence for antecedents. It solves personal and possessive pronouns with noun phrases as antecedents.The algorithm is a left to right breadth first search on the syntactic parse tree of the input sentence. Given the usual order of syntactic categories in the English language (the subject is generally followed by the verb) the algorithm expresses a preference for the noun phrases subjects.

Probably one the well known algorithm for pronoun resolution was proposed by Lappin & Leass ([08]). The algorithm exploits salience factors and their associated weights such as sentence recency, subject emphasis, head noun emphasis), and so on to perform pronominal resolution. The salience value is simply the sum of the associated weights. Once salience values have been calculated for each referent, the algorithm can be applied to resolve the pronouns. The entity with the highest salience value is declared to be the most likely referent. If there are no pronouns to be resolved in a sentence, the next sentence is processed and the weights that contribute to an entity's salience are halved (to account for sentence recency). The weights used in the salience algorithm are ad hoc. Lappin & Leass use deep linguistic information in three places: firstly, to determine binding-based incompatibility and restrictions on the resolution of reflexives; secondly, to assign salience weights based on grammatical functions; thirdly, they use the parser's lexicon to assign the gender of full noun phrases.

Kennedy and Boguraev ([07]) describe a variant that does not require in-depth, full syntactic parsing of text. Instead, with minimal compromise in output quality, the modifications enable the resolution process to work from the output of a part of a speech tagger, enriched only with annotations of grammatical function of lexical items in the input text stream. Their method has been applied to personal pronouns, reflexives and possessives. The general idea is to construct co reference equivalence classes that have an associated value based on a set of ten factors. An attempt is then made to resolve every pronoun to one of the previous introduced discourse referents by taking into account the salience value of the class to which each possible antecedent belongs. Based upon their own evaluation of the results of their implementation they state that accurate anaphora resolution can be realized within natural language processing frameworks which do not, or cannot, employ robust and reliable parsing components.

Mitkov's algorithm ([11], [12], [13]) is another knowledge poor approach to pronominal resolution, which means that it uses only the output of a part of speech tagger with minimal syntactic information. The linguistic analysis for anaphora resolution includes the output of a part of speech tagger, augmented with syntactic function annotations for each input token. Mitkov's algorithm does not use any parsing of the input sentence and relies heavily on various boosting indicators (First Noun Phrases, Indicating Verbs, Lexical Reiteration, Section Heading Preference, etc.) as well as on some impeding indicators (Indefiniteness and NPs appearing in Prepositional Noun Phrases). The boosting indicators assign a positive score to a noun phrase, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding ones apply a negative score to a noun phrase, reflecting a lack of confidence that it is the antecedent of the current pronoun. A score is calculated based on these indicators and the discourse referent with the highest aggregate value is selected as antecedent. The author reports success rate of 89.7% on a corpus of technical manuals.

Trouilleux's algorithm ([17]) is a rule based pronoun resolution for French that uses full syntactic parsing. The method restrains the list of potential antecedents by using the notion of an insertion. He defines an insertion to be either a sequence between parentheses, a sequence delimited by commas between a verb and its subject or object, an opposition to the right of a noun phrase, or an opposition to the left of a subject noun phrase. The algorithm is reported to have a good rate of success (74.8 %). The corpus used is newspaper articles in the domain of finance. A similar resolution system is described in [02]

More recently, Vieira and Poesio ([19]) and Harabagiu et al. ([04]) explored the use of WordNet for different coreference resolution subtasks, such as resolving bridging references, other and definite NP anaphora, and MUC-style coreference resolution. All of them present systems which infer coreference relations by means of WordNet search to check whether the referring expressions are synonyms or in a hyponymy or hypernymy relation with each other.

Approaches using manually built knowledge bases rely on high-quality knowledge manually inputted by human experts at the cost of a limited coverage, whereas proposals making use of information automatically extracted from corpora achieve a higher coverage for a quality lower than that of humans.

Other research is concerned with implementing a pronominal resolution system and adjusting these weights empirically using machine-learning techniques based on real corpus data. Recently some authors use genetic algorithms to adjust the salience weights. Various corpora differ in the number of words, the domain of the texts, the types and the distribution of pronouns, types and distribution of named entities, the complexity of resolving certain constructs. An algorithm performing very well on a corpus of technical manuals may fail for a corpus of news articles or dialogs containing quoted speech. It is therefore important that the implementation considers specifics of the target texts upon which it is intended to operate. From 2005 to the present date more sophisticated machine learning techniques and richer features, especially semantic information begins to be incorporated.

GENERAL OUTLINE OF THE SYSTEM

The overall system design is solely based on the principle that psycholinguistic research supports the claim that listeners process utterances one word at a time, so when they hear a pronoun they will try to resolve it immediately. If new information comes into play which makes the resolution incorrect (such as a violation of constraints), the listener will go back and find a correct antecedent. In our system prototype we choose the rule based approach because parsing with definite clause grammars gives valuable morphological, syntactic, semantic and discourse information that can be exploited by the anaphora resolution module. This chapter describes a solution for the Anaphora Resolution for French based on previous systems and augmented with an XML generator that helps to use linguistic knowledge (syntactic, semantic, and discourse) more efficiently.

Our system is a knowledge rich algorithm; it relies on the parsing process, based on Definite Clause Grammars (DCGs), and the XML representation to collect valuable information (morphological, syntactic, semantic, and discourse structure) to be used by the anaphora resolution module. Moreover parsing with DCGs captures the essence of the French language to be analyzed. The anaphor identification module identifies as possible anaphors every pronoun in the text that belongs to the set {il, elle, ils, elles, le, la, les, me, te, lui, leur, leurs }. In order to apply different strategies we need to make a distinction between constraints (which must not be violated) and preferences (which discriminate among candidates satisfying all constraints). Preferences are applied in a predefined order.

Our approach to anaphora resolution consists of three main components: A parser, an XML generator, and the anaphora resolution mechanism.

- The first component of our system parses and processes the input text using a Definite Clause Grammar implemented in Prolog for French. The output of this component is a set of parse trees containing the constituents of the input sentences.
- An XML generator that uses the lists developed in the previous phase to tag each element of each sentence in creating the XML-tagged version of the parse trees generated by the parser. This step is of greater importance because we have observed that the XML structures representing the output of the parser allow an easy to handle examination and manipulation of low level tree nodes.
- Designing an interface allowing anaphora resolution.

In what follows we present the parser, the XML generator and the anaphora resolution module.

The Parser

Text processing begins by splitting the text into sentences. The splitter is very efficient and accurate, particularly dealing with abbreviations and initials that frequently result in sentences being improperly split. After splitting, each sentence is submitted to the parser. The use of definite clause grammars has proved very efficient for the processing of French sentences. The grammar which covers a substantial set of the language was designed in previous research projects in machine translation and has been improved over the years (Refoufi, [14]). The parser implemented in Prolog is based on a substantial set of definite clause grammars rules, which are grammar rules augmented with arguments used to capture features and build structures.

After parsing and converting the output of the parser into an XML representation, the resulting text is traversed by the algorithm. On encountering a possible antecedent (currently most NP's), a special structure (SS) is created. The SS has a semantic part with natural agreement values, human hood and an index, and a syntactic part with grammatical agreement, syntactic function and category. The SS is put on a stack A of possible antecedents. On encountering a pronoun P from the target category, the system is triggered, with P as the input. The details of the algorithm are given in the subsection below. The output of the system is the antecedent for P.

A full SS for P is constructed by combining the pronoun's syntactic information with the semantic features of the antecedent. The result is put on A, to serve as a possible antecedent itself. On encountering a pronoun, the algorithm looks through the list of discourse entities it has come across in the current sentence. The first entity that meets agreement and binding constraints is selected; otherwise we apply the preferences in order.

The XML generator

Converting the output of the parser to an XML representation allows an easy to handle mechanism to tackle anaphora resolution. This component is a key aspect of our system because it permits to correct and improve the representation, especially adding information (syntactic as well as semantic) to the structure. In fact the parsing process may lack helpful data, due to the complexity of natural language processing. (Recognition of named entities, wrong delimitation or misidentification of noun phrases, the peculiar use of reflexives and possessives in the French language). This component may include WordNet lookup, it can also be hand refined if needed. Annotations, such as number and tense information and attachments points of noun and prepositional phrases, may be included at any node. This step is the most time consuming part of our project.

The anaphora resolution mechanism

This difference between constraints and preferences plays an important role in many computational models of anaphora resolution and is also followed in standard expositions such as (Mitkov, [12]) so we'll follow it here even though there is not conclusive evidence about the existence of two distinct mechanisms. In most cases, the combination of constraints and preferences is sufficient to ensure that anaphoric expressions have a single most preferred interpretation in context. The actual resolution mechanism is an improvement of an earlier version (Refoufi, [14]).

The preprocessing step begins by invoking the procedure that identifies pleonastic pronouns; these pronouns are discarded in subsequent steps. Our resolution method works by applying the constraints first to reduce the number of candidates, then the preferences are applied to each of the remaining candidate. If the solution is not unique, we apply the preferences in the order given until a solution is selected. The algorithm stops as soon as a unique candidate for antecedence is found. If more than one preference applies, and each suggests a different candidate for the anaphor in question, then we consider the anaphor to have an ambiguous referent. In that case, we may reduce the space of possible referents to those that are accepted by the constraints.

The constraints are applied in order and the candidate must meet the conditions. Constraints are easy to use: reject all hypotheses which violate the hard constraints (if you can accurately detect the constraints), however preferences are more difficult to implement. After each sentence is parsed, its parse tree is traversed in a depth-first recursive function and converted to an XML-tagged representation. During this traversal, each node (terminal and non terminal) is analyzed, making use of parse tree annotations and other lexical resources. The focal points in the traversal of the parse tree are the noun phrases. When a noun phrase (discourse entity) is encountered, its constituents are examined and its relationship to other sentence constituents is determined. Each noun phrase is added to a list of discourse entities for the entire text, i.e., a "history" list. As each noun phrase is encountered, it is compared to discourse entities already on the history list. This comparison first looks for a prior mention, in whole or in part, to determine whether the new entity is a coreferent of a previous entity (particularly valuable for named entities). If the new entity is an anaphor, the anaphoric resolution module is invoked to establish the antecedent.

The feature structure used to collect possible antecedent can be one of the two forms. For NPs we use the structure antecedent(det(G,N,P), adj(G,N), noun(G, N), GR, GC, RM). For named entities we use the structure : antecedent(first_name(), last_name(), GR, GC, RM) where GR stands for grammatical role, PA for parallelism, RM for repeated mention, and finally RE for recency.

Resolution procedure

In what follows we outline thoroughly the anaphora resolution strategy.

01 : read the next sentence

Collect all NPs and named entities according to the structures defined above. Call this set S

Identify and remove non anaphoric pronouns

Locate the next pronoun in the current sentence; call it p

Apply the heuristic ana-proc (S,p)

Go to 01

Procedure ana-proc(S,p)

Let N be the number of possible candidates (the size of the set S).

Apply the morphological and the syntactic constraints. Let S_1 be the set of remaining candidates and N_1 its size.

If $N_1 = 0$ report failure

If $N_1 = 1$ report success, output (S_1,p)

If $N_1 > 1$ then apply the Grammatical Role preference. Let S_2 be the set of remaining candidates and N_2 its size.

If $N_2 = 1$ then report success and output (S_2,p),

Else if $N_2 \# 1$ then apply the parallelism preference. Let S_3 be the set of remaining candidates and N_3 its size If $N_3 = 1$ then report success and output (S_3,p),

Else if $N_3 \# 1$ then apply the Repeated mention preference. Let S_4 be the set of remaining candidates and N_4 its size If $N_4 = 1$ then report success and output (S_4,p)

Else if $N_4 \# 1$ apply the Recency preference. Let S_5 be the set of remaining candidates and N_5 its size. If $N_5 = 1$ then report success and output $\{S_5,p\}$

Else report failure and stop.

Thus far we have been focusing on the problem of selecting the best anaphoric referent among several candidates all from a previous sentence. When prior context contains many sentences the question naturally arises of how far back to search for the anaphoric referent, and how to design that search. At the paragraph level we advocate searching sentences in reverse chronological order, applying all the constraints and preferences to select among candidates within each sentence. In practice the number of sentences that we explore never exceeds four (4) otherwise the search space becomes too complex to handle.

CONSTRAINTS AND PREFERENCES

As mentioned earlier, in order to apply different strategies we need to make a distinction between constraints (which must not be violated) and preferences (which discriminate among candidates satisfying all constraints). Preferences are applied in a predefined order. The constraints and the preferences that we propose are traditional resolution factors that in some cases have been adapted for French. The constraints that we use are:

1. Morphological agreement: both the anaphor and the antecedent must agree in person, number and gender.
2. Syntactic Agreement on local domain (see section.2.1.)

The preferences are:

1. Grammatical role: entities in subject/object position are more salient.
2. Parallelism: the pronoun and the antecedent share the same syntactic category.
3. Repeated mention: entities mentioned more frequently are more salient.
4. Recency: the most recent antecedent mentioned in the discourse. The recency preference favours the candidate which is nearest to the anaphor; that is the one evoked recently.

Preferences are applied in the above order. Some preferences are considered to be more important than others, mainly due to the analyzed language characteristics. This heuristic, consisting on the set of preferences and their relative priority derives from our past experience with French. The use of several anaphora resolution factors in combination allows greater confidence in the anaphor antecedent identification.

DISCUSSION

In this paper, we have described the fundamental components of a new anaphora resolution framework that can process French short texts from scholar manuals. Manual texts are generally well written, with a correct style, without metaphors or humor; the presence of anaphors is dense and is mainly pronominal. The resolution mechanism is derived from the XML representation of the parser's output. The parser is based on a set of definite clause grammars that define valid constructions of French sentences. The XML-tagged representation of the parser's output has proven valuable in the exploitation of syntactic and semantic features used by the anaphora resolver. Moreover the XML tagged representation is much more structured and easy to use than the direct raw output of the parser. On the linguistic side, developing the program and error analysis have already given interesting hints for further research especially the peculiar behaviour of possessives. The evaluation of

preliminaries version of our system's architecture shows promising results.

The most common types of errors that influence the resolution process and introduce errors are misidentification and wrong delimitation of noun phrases. Other factors that have negative impact on the performance of the anaphora resolution module arise from deficient parses. For the time being the parser has a wide linguistic coverage of French syntax, and uses a dictionary of about 2000 words. Proper names are only accepted if they belong to a special list, which can be updated easily. In particular named entities or unknown words still provoke failures. We also encountered the case where the gender feature of the pronoun "elle" is of little help in the determination of the correct antecedent, as in the example "Le docteur Bouzidi est un spécialiste en chirurgie. Elle travaille dur nuit et jour". (Doctor Bouzidi is a specialist in surgery. She works hard night and day).

Our presentation shows that previous research efforts have made it possible to achieve a fair understanding of the phenomenon of anaphora and the main factors affecting it. However, from a computational perspective we argue that there are still a variety of open issues to achieve robust models of anaphora resolution.

TABLE1 TYPES OF ANAPHORA TREATED

Category of pronouns	Anaphoric item
Personal pronouns subjects	il, elle, ils, elles
Personal pronouns objects)	le, la, les
Personal pronouns indirect object	me, te, lui, leur, leurs

In its actual implementation the pronoun « il» is pleonastic in the following expressions : il pleut ; il neige ; il gèle ; il fait froid ; il fait chaud ; il y a ; il fait nuit ; il fait jour ; il est tôt ; il est tard ; il est midi ; il est minuit ; il est certain ; il est sûr ; il est probable ; il faut ; il est nécessaire ; il est possible.

REFERENCES

[01] Chomsky, N. 1981. Lectures on Government and Binding. Dordrecht: Foris.

[02] Depain-Delmotte F. 1999 La Sélection de l'Antécédent du Pronom dans les Systèmes de Traitement Automatique des Langues Naturelles, Proceedings of Vextal'99 Venezia San Servolo.

[03] Harabagiu, Sanda M., Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, Penn., 2–7 June, pages 55–62.

[04] Hirst, G. 1981. Anaphora in Natural Language Understanding: A Survey Lecture Notes in Computer Science 119. Berlin: Springer-Verlag.

[05] Hobbs,J.R. 1976. Pronoun resolution. Technical Report 76-1, Department of Computer Science, City College, City University of New York,1976.

[06] Hobbs, J.R. 1978 "Resolving pronoun references ".Lingua, 44, 339-352.

[07] Kennedy, C.& Boguraev B., 1996."Anaphora for everyone: pronominal anaphora resolution without a parser". Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), 113-118. Copenhagen, Denmark

[08] Lappin S.; Leass H.J. An Algorithm for Pronominal Anaphoric Resolution, Computational linguistics 1994, Vol 20(4) pp535-561

[09] Litkowski KC, 2002 Explorations in Disambiguation Using XML Text Representation Research www.clres.com/online-papers

[10] Miltsakaki E. (2007) A rethink of the relationship between salience and anaphora resolution. Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium, Lagos, Portugal, pp. 91-96.

[11] Mitkov, R., Robust Pronoun Resolution with Limited Knowledge, Proceedings of the 18P^{thP} International Conference on Computational Linguistics, Montreal, Canada, 1998

[12] Mitkov, R. 2001 « Outstanding Issues in Anaphora Resolution, in Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Mexico City.

[13] Mitkov, R. et al. (2007) Anaphora resolution: to what extent does it help NLP applications? In: Anaphora: Analysis, Algorithms and Applications, Springer-Verlag, Berlin Heidelberg, pp. 179-190.

[14] Refoufi A. 2007 "A modular architecture for anaphora resolution" Journal of Computer Science 3(4) 199-203.

[15] Reinhart, T. 1976. The Syntactic Domain of Anaphora. PhD thesis, MIT, Cambridge Mass., 1976.

[16] Steinberger J. 2009 Measures for Text Summarization, Computing and Informatics Vol 28,1001-1026

[17] Trouilleux, F., 2002. Insertions et interprétations des expressions pronominales. In Actes de l'Atelier « Chaînes de référence et résolveurs d'anaphores » TALN 2002, Nancy.

[18] V. Ng and C. Cardie, 2002 "Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution, " in Proceedings of the 19th International Conference on Computational Linguistics, vol. 1, pp. 1–7.