http://ansinet.com/itj

ITJ

ISSN 1812-5638

# INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 6 (8): 1152-1157, 2007
ISSN 1812-5638


# Toward the Resolution of French Anaphoric Definite Descriptions

Allaoua Refoufi
Department of Computer Science, Ferhat Abbas University, Setif 19000, Algeria

**Abstract:** Anaphora resolution attempts to determine the correct antecedent (entity that it is referred to) of an anaphor (the term pointing back). In what follows, we propose an algorithm for the resolution of anaphoric definite descriptions that relies on lexical, syntactic and semantic knowledge. The semantic knowledge is incorporated as a hand-coded ontology. The algorithm, based on several heuristics, returns an indefinite noun phrase as the antecedent when the latter and the anaphor (the definite expression) satisfy one of the following semantic relations: synonymy, hyperonymy (superclass), hyponymy (subclass), meronymy (part-whole) and the compatibility constraints. The compatibility constraints are applied to the modifiers of the antecedent and to those of the anaphor.

**Key words:** Anaphora, synonymy, hyperonymy, part-whole relation

## INTRODUCTION

Anaphora (from the Greek αναφορα meaning carrying back) describes the dependence of an expression on a previously mentioned one in a discourse segment. It is an important phenomenon required in almost every natural language application. Anaphora is used to explicitly exhibit relations between different linguistic units that designate the same entities, that is are co referential. The object that is being referred to is called the antecedent, the expression that refers to the antecedent is called the referring expression or the anaphor. The process of finding the proper antecedent for each anaphora in texts is termed anaphora resolution. In French, anaphoric expressions are signalled by two linguistic categories: pronominal anaphora and definite descriptions. Definite descriptions (noun phrases beginning with a definite article) that provide a recall can be identical to their antecedent (direct anaphora), or semantically related to it (indirect anaphora). Indirect anaphora, also known as associative anaphora, arises when a reference becomes part of the hearer's or reader's knowledge indirectly rather than by direct mention (Mitkov, 2002). For example (the anaphor is in bold, the correct antecedent in italics), the following sentences contain instances of direct and indirect anaphora.

(01) «J'ai adopté un chien. Le chien est très affectueux. ("I adopted a dog. The dog is very affectionate»)
(02) «Qu and le fuyard regagna son domicile, la porte était entrouverte. ("When the runaway arrived home, the door was wide open»)

The referring expression, "la porte", relates to the antecedent, "son domicile" through a part-whole relation.

Anaphoric noun phrases (NPs) are essentially signalled by definite or demonstrative determiners and also by other semantically related entities. They establish links between discourse elements and provide additional information that allows the discourse to evolve. Anaphora resolution is complex because it depends on several factors: the nature of the anaphor, its syntactic function and the distance between the anaphor and the antecedent. The interpretation of an anaphor must be correctly assured in order to perform the right associations to the entities or characters evoked in the discourse. In general several antecedents are possible for the same anaphor. This ambiguity constitutes the main challenge of the problem.

In any resolution method two stages are necessary: first locate the anaphoric expressions that designate entities mentioned elsewhere in the text, then find out the correct antecedent to each one. Indirect anaphora is especially challenging to resolve because the referring expression and the antecedent are related by unstated background knowledge.

The algorithm identifies antecedents (indefinite noun phrases, proper-names and named entities) of definite noun phrases in French and is applied to the parse tree generated by the parser. An apposition, a discourse segment deliminated by commas, usually brings additional information. Therefore definite descriptions occurring in appositions are treated as anaphoric on the preceding noun phrase; consequently, there is no need to search for an antecedent when the anaphor is an apposition, as in:

(03) Caesar, the roman emperor, died in 44.

The algorithm works as follows: during the parsing process, when we encounter a definite description all indefinite noun phrases appearing in the last four sentences, starting from the current sentence, are extracted. Each indefinite noun phrase is considered a potential candidate and is compared to the definite description. Based on the semantic knowledge provided, the algorithm returns a first response when one of the following conditions is met:

- The head of the indefinite noun phrase matches exactly the head of the definite description.
- The head of the indefinite noun phrase is a synonym of the head of the definite description, such as un livre and le bouquin.
- The head of the indefinite noun phrase is a hypernym (superclass) of the head of the definite description as un vehicule and la voiture.
- The head of the indefinite noun phrase is a hyponym (subclass) of the head of the definite description, such as un bras and le corps.
- The head of the indefinite noun phrase has a part-whole relation (meronymy) with the head of the definite expression, such as le volant and une voiture.

In the second stage, the modifiers of the value provided in stage one are checked for compatibility with the modifiers of the referring expression. If the compatibility checks succeed a correct antecedent is returned, otherwise the search continues with the next potential candidate.

## TYPES OF ANAPHORA

The following types of anaphora are treated:

- Direct or identity anaphora: the antecedent is referred to by an identical definite noun phrase:

(04) J'ai adopté *un chien*. Le chien est très affectueux

- Nominal anaphora: the antecedent to the definite expression is a named entity or a proper name:

(05) Abdelaziz Bouteflika visited the town of Setif. L'hôte du palais du peuple inaugura plusieurs réalisations.

- Indirect or associative anaphora: the antecedent is referred to by a semantically related definite noun phrase:

(06) J'ai acheté une belle voiture. Le moteur est très puissant, le prix abordable.

The expressions le moteur and le prix are related to the expression une belle voiture.

Direct or identity anaphora happens when the head of the noun phrase of the candidate matches exactly the head of the anaphor, it is also termed faithful anaphora. The anaphor's determiner may be a demonstrative pronoun. Examples are:

(07) Un homme entra dans la salle. L'homme semblait chercher quelqu' un.
(08) Une grande ferme se dessinait à l'horizon. Cette ferme appartenait à un ami.

Indirect, or associative anaphora, happens when the head of the antecedent and the head of the anaphor are different but semantically related. Examples are:

(09) Une voiture s'arrêta soudain. L'automobile était toute neuve.
(10) Une grande gazelle apparut aussitôt. L'animal paraissait exténué
(11) Le medecin nous montra le corps inerte. Le bras était sérieusement amoché

Indirect anaphora is a type of anaphora that may account for 15% of noun phrase anaphora, making it an important type (Vieira and Poesio, 2001). The semantic relation between the anaphor and the antecedent may take different forms; in this paper we explore four types: synonymy, hyperonymy, hyponymy and part-whole relations. Example (09) is a case of synonymy, (10) is an example of hyperonymy or generalization and example (11) is a part-whole association; that is le bras is a part of the object le corps. Part-whole (meronymy) relations are hard to implement mainly because a clear understanding requires a deep interaction of logic, semantics and pragmatics (Girgu and Budulescu, 2006). The relation hyponymy is the reverse of the hyperonymy relation. Other types of anaphora which are not treated in this paper are: pronominal anaphora and verb phrase anaphora. The former requires a different kind of treatment well documented in the literature (Refoufi, 2007), while verb phrase anaphora requires more complex processing (Mitkov, 2002). An example of verb phrase anaphora would be:

«Sarah essaya de convaincre Rachid de se reposer. La tentative fût vaine.» ("Sarah tried to convince Rachid to rest. The attempt was vain.")

## SOURCES OF KNOWLEDGE

Anaphora resolution requires considerable knowledge, ranging from morphology, syntax, semantics and pragmatics. Lexical and syntactic knowledge must be provided in an efficient manner for the parser to be able to identify parts of the sentences. It is crucial to clearly designate which modifier is attached to which noun in order to perform the operations needed by the algorithm (type of semantic relations between noun phrases, compatibility constraints).

Semantic knowledge is of particular importance when dealing with indirect anaphora. A strategy adopted is to search for conflicts between the semantic descriptions associated with the anaphoric noun phrase and those associated with the candidate noun phrase. A contradiction arises if the heads of the respective noun phrases are not in a synonymy, generalization, specialization or part-whole relation. A contradiction would also arise if the modifiers (premodifiers and postmodifiers) of the anaphor and the modifiers of the candidate noun phrase are semantically incompatible.

## DEFINITE AND INDEFINITE DESCRIPTIONS

We usually use a definite noun phrase to refresh the hearer's memory, since it is difficult to understand references that are out of focus. Therefore definite descriptions in natural language discourses are those noun phrases which, on one hand, introduce new entities and on the other, those that refer back to entities previously introduced. It is in general believed that pronouns are used to refer to entities in focus, whereas entities that are out of focus are referred to by definite descriptions. Definite descriptions do more than just refer, they convey more additional information (Salmon-Alt, 2004). However, not all definite descriptions are anaphoric, as in:

(12) Le tourisme s'est fortement développé avec l'aménagement du littoral.

It is reported in the literature that definite expression are anaphoric about half of the time (Salmont-Alt and Vieira, 2002), as a result of this claim techniques for resolving definite expression should include a two stage process in which the first stage identifies anaphoricity of the noun phrase and the second the antecedent for anaphoric noun phrases (Markert and Nissim, 2005). In this article, we focus on the second stage, namely, antecedent selection of a definite expression, thus making the assumption that all definite noun phrases are anaphoric.

Moreover, a definite expression might refer to another definite expression, as in:

(13) Le paludisme continue de faire des ravages en Afrique. Ce fléau n'est pas encore sérieusement pris au sérieux par la communauté internationale.

Another problem happens when an indefinite noun phrases refers back to a definite noun phrase in:

(14) Le professeur distribua les feuilles d'examen. Une copie manquait.

## PREVIOUS WORK

Much research has been performed in the field of anaphora resolution and especially in the field of pronominal resolution however, less attention has been paid to resolving proper nouns and definite noun phrases. Anaphoric definite noun phrases are typically harder to resolve. Recent studies can be divided into WordNet based systems and machine-learning systems. The WordNet-based systems (Markert *et al*., 2003) and (Vieira, 2001) use WordNet as the knowledge base to identify semantic relationships between the anaphor and the antecedent. These systems are reported to perform poorly on definite noun phrases, moderately on proper names and quite good on personal pronouns and possessive pronouns.

Recent work on the interpretation of definite noun phrases is that of Vieira and Poesio (2001); their work led to the design of a shallow processing system relying on structural information and on information provided by existing lexical resources such as WordNet. As a result of the knowledge-poor approach adopted, the system is not really equipped to handle definite descriptions which require complex reasoning.

Another work of significant importance is the automatic discovery of part-whole relations (Girgu and Badulescu, 2006). They present an algorithm for the discovery of lexico-syntactic patterns using extensive corpus and WordNet. The domain independent approach relies heavily on a set of classification rules, although the method presented is highly complex, it raises nevertheless important issues on how to identify the main tasks for this linguistic problem. However, WordNet does not contain various semantic relations and is not available for most languages.

## THE HEURISTICS

In the problem we are considering, the relation between anaphor and antecedent is implicitly expressed,

that is, anaphor and antecedent do not stand in a structural or grammatical relationship. However, they are linked by some sort of semantic relation. Exhibiting the relation explicitly is a subtask of the algorithm.

The algorithm works as follows: During the parsing process, when we encounter a definite description all indefinite noun phrases appearing in the last four sentences, starting from the current sentence, are extracted. The number of sentences processed is chosen to be 4 in order to keep the complexity of the heuristics easy to handle, on one hand; on the other hand expressions which are quite far are considered to be out of focus.

Step 0: Check whether the definite noun phrase occurs in an appositive construction, if so there is no need to find an antecedent for it (example 03).

Step 1: Each indefinite noun phrase is considered a potential candidate and is compared to the current definite description. Based on the semantic knowledge provided, the algorithm checks whether one of the following conditions is met:

- The head of the indefinite noun phrase matches the head of the definite description.
- The head of the indefinite noun phrase is a synonym of the head of the definite description.
- The head of the indefinite noun phrase is a hypernym (superclass) of the head of the definite description.
- The head of the indefinite noun phrase is a hyponym (subclass) of the head of the definite description.
- The head of the indefinite noun phrase has a part-whole relation (meronymy) with head of the definite expression.

In the second stage of step1, the corresponding modifiers are checked for compatibility. If the compatibility checks succeed a correct antecedent is returned, otherwise the search continues in the set containing all indefinite NPs appearing in the last four sentences.

## IMPLEMENTING THE HEURISTICS

To track the potential antecedents, the system keeps track of noun phrase index, noun phrase structure, head noun and noun phrase type (definite, indefinite and other), as illustrated by the predicate description1 below (Refoufi, 2006):

Description1(Order,np(NP_Structure),head(H), premodifier(Prem), postmodifier(Postm), type(T))

the main components are:

```
description1(          */predicateforpotential
                    antecedent /*
Order,              */ order number /*
np(NP_Structure),   */ a part of the parse tree /*
head(H),            */ head of the noun phrase /*
premodifier(Prem),  */ noun phrase premodifier /*
postmodifier(Postm), */noun phrase postmodifier /*
type(T)             */type of the noun phrase (
                    definite, indefinite) /*
)
```

**Checking for the semantic relation:** We state that there is a semantic relation between two entities (an indefinite NP and a definite NP) if the following conditions hold together:

- The corresponding head nouns are identical, synonymous, hyperonymous, hyponymous, or in a part-whole relation.
- The corresponding premodifiers are compatible
- The corresponding postmodifiers are compatible

We explain in further detail each condition:

- Checking for direct anaphora means that the head noun of the antecedent matches exactly the head noun of the anaphor (a demonstrative article may be attached to the anaphor).

(15) Une maison paisible se dessinait à l'horizon. La maison servait de refuge pour les gens de passage.

- Checking for synonymous entities is easily taken care of by a table of nouns and their corresponding lists. Examples of such lists are:

synonymes-list1([maison, demeure, foyer, appartement, logis, foyer])
synonymes-list2([voiture, automobile, caisse])

- Checking for hyperonymy relations is implemented through a hand coded ontology. One way to do it is to write clauses like: is-a(Object, Class) as in:

is-a(chien, animal); is-a(lion, animal);is-a(bureau,meuble).

- Checking for a hyponymy is performed through the reverse relation is-a, that is: hyponym(X,Y):- is-a(Y,X)

For example hyponym(félin,chat):-is-a(chat,félin)

- Part-whole relations are taken care of by the use of the predicate part-of, as in:

part-of(moteur,[voiture, bateau, avion, etc.]).
Examples include:

(16) Une grande muraille surplombait la vallée. Le somptueux édifice fut construit il y a 5000 ans.

The compatibility condition certifies that the arguments are consistant in the usual sense, as in:

(17) Une belle femme prit la parole. La jeune dame était concernée par le problème.

In French the premodifiers are termed épithètes and the postmodiiers are termed adjectives qualificatifs or attributs.

The premodifiers (épithètes) belle and jeune are of course compatible.

The main predicate that compares the indefinite noun phrase (antecedent) and the definite noun phrase (anaphoric expression) is:

```
compare1(X, Y):- compare2(H1, H2),
compatible(Prem1, Prem2),
compatible(Postm1, Postm2),
T1<>T2.
Where X = description_1(Order1, np(NP_Structure1),head(H1), premodifier(Prem1), postmodifier(Postm1), type(T1))
And Y = description_1(Order2, np(NP_Structure2),head(H2), premodifier(Prem2), postmodifier(Postm2), type(T2))
compare2(H1,H2):-match(H1,H2),!.
    */direct anaphora /*
compare2(H1,H2):-synonyme(H1,H2),!.
compare2(H1,H2):-hyperonyme(H1,H2),!.
compare2(H1,H2):-hyponyme(H1,H2),!.
compare2(H1,H2):-part-whole(H1,H2),!
compatible(A,B):-not antynome(A,B).
```

## DISCUSSION

Since we are mainly dealing with synonyms and hyperonyms the agreement constraint (gender) between the antecedent's head noun and the anaphor's head noun is not required. However the imposition on the number constraint might help in the filtering of candidates.

The external mistakes that have negative impact on the anaphora resolution module performance are: not recognized named entities and sentences not properly parsed. In general to account for robustness, deficient parsing is not analyzed further.

Proper names are regarded as potentially anaphoric to preceding proper names that match in terms of first or last names.

In general parsing involves difficult problems like modifier attachments, scope quantifiers etc. To overcome these difficulties we have implemented the principle of minimal attachment to reduce the number of ambiguous parse trees. The minimal attachment principle states that we favour the parse tree with the minimum number of nodes.

Some relations between entities are hard to identify as in:

(18) une collation fut organisée par la direction. L'ambiance était festive

## CONCLUSION

In this study we are concerned with the resolution of noun phrases that refer directly or indirectly to an antecedent within the discourse. The main idea consists of the establishment of a link between definite descriptions and entities mentioned in the text that are semantically related. The resolution relies heavily on a robust parser and a hand-coded ontology. Preliminary test are quite good, however a real evaluation of the algorithm is undertaken and will be reported on future work. This work is conducted as a laboratory project by the natural language processing group at the computer science department of the university of Setif.

Improvements must be realized on diverse directions, namely:

- Design a corpus on which evaluation must be performed and compare the performance of the algorithm with similar resolution methods.
- The work has to be complemented by a module which identifies whether a definite noun phrase is anaphoric or not.
- Define precisely what synonymy, hyperonym and part-whole relations really mean.
- Identify precisely what knowledge should be incorporated in the ontology used and how knowledge should be encoded.
- Evaluate the costs of building, maintaining and searching ontologies.

- Try some other implementation languages and compare the efficiency.

## REFERENCES

Girgu, R. and A. Badulescu, 2006. Automatic discovery of part-whole relations. Comput. Linguistics, 32: 83-123.

Markert, K., M. Nissim and N. Modjeska, 2003. Using the web for nominal anaphora resolution. In: Proceedings of the EACL Workshop on the Computational Treatment of Anaphora, pp: 39-46.

Markert, K. and M. Nissim, 2005. Comparing knowledge sources for nominal anaphora resolution. Comput. Linguistics, 31: 367-402.

Mitkov, R., 2002. Anaphora Resolution. Longman, London.

Refoufi, A., 2006. A multiple knowledge source algorithm for anaphora resolution. Asian J. Inform. Technol., 5: 48-53.

Refoufi, A., 2007. A modular architecture for anaphora resolution. J. Comp. Sci., 3: 199-203.

Salmont-Alt, S. and R. Viera, 2002. Nominal expressions in multilingual corpora: Definites and demonstratives. Proc. of LREC 2002. Las Palmas, Spain.

Salmon-Alt, S., 2004. Automatic resolution of indirect anaphors in French: What resources for what contribution. TALN 2004, Poster Session, Fès, pp: 19-21.

Vieira, R. and M. Poesio, 2001. An empirically based system for processing definite descriptions. Comput. Linguistics, 26: 539-593.