Generating Coreferential Descriptions from a Structured Model of the Context

Hélène Manuélian

Equipe Langue et Dialogue
LORIA, Campus Scientifique, BP 239, F - 54506 Vandoeuvre lès Nancy Cedex
helene.manuelian@loria.fr

Abstract

This paper shows on the basis of a corpus study how a model of the context should be structured for the generation of coreferring descriptions in French. We show that this way of structuring the context can help to generate more paraphrases and a particular kind of referring expressions used to add information about the referent.

1. Introduction

Generation algorithms for referring expressions (Dale and Reiter, 1995; Gardent and Striegnitz, 2000) usually generate definite coreferential descriptions which denote only given information about the referent. One way of extending these algorithms could be to make them generate definite and demonstrative coreferential descriptions which add information about the referent.

With this purpose in mind, we conducted a corpus study of demonstrative and definite coreferential noun phrases in French. The corpus used is the PAROLE corpus[1] which contains 65 000 words and about 10 000 definite and demonstrative noun phrases, of which 1771 are coreferential.

This study confirmed that definite and demonstrative coreferential noun phrases can be used both to repeat information and to introduce new information about their referent (Manuélian, 2003). In this paper, we will show how this result can be used to improve the generation of referring expressions by using a structured model of the context.

2. Goal of the paper

The main goal of this paper is to show how the model of the context used by algorithms should be structured for a better generation of referring expressions.

We first present existing algorithms for the generation of referring expressions, in particular the method presented by (Gardent and Striegnitz, 2000) for structuring the context for the generation of bridging descriptions. We will then summarise the results of the corpus study which led us to propose a new structure for the model of the context used by the algorithm. In the last part of the paper we present this structure, the required databases and a proposition for the extension of Gardent and Striegnitz's algorithm.

3. Generation algorithms

The algorithm proposed by Gardent and Striegnitz generates the content of definite descriptions however they are used (first mention, bridging, coreferential). The algorithm, like the standard algorithm for referring expressions of Dale and Reiter (1995) computes the distinguishing description for the referent using constraints of unicity (the referent must be the only one to fit the description).

The first difference with (Dale and Reiter, 1995) is that a **familiarity constraint** is added. The referent has to be *hearer-old*. To satisfy this constraint, a set of intended anchors I(A) is built, which contains all the entities already mentioned in the previous discourse which the speaker can relate either by identity or by a bridging relation to the intended referent. The intended anchor has to be included in the potential anchor set P(A) which includes all the entities that the hearer can relate to the intended referent.

The other difference with the standard algorithm is that it uses as an input (1) a **structured model** of the context, composed of the following components: (i) a knowledge base representing world knowledge formalised in first order logic (WKL) (ii) a speaker model representing the new information to be generated (SM) (iii) a discourse model representing all the information already given by encapsulating the previous discourse (DM); and (2) the referent to be described *t* (target entity).

The algorithm can be formalised as follows:

Input:
WKL (world knowledge): set of rules linking entities together
DM (discourse model): set of atomic formulas
SM (speaker model): set of atomic formulas
t: target entity, t is a term of SM and DM.

Initialisation:
1. goals ← stack with the element t
2. N ← syntactic structure with an empty place for the noun phrase

Check success:
3. If goal is empty, return <uniquely identifying, N>
4. current-goal ← top goal
5. If IA(current-goal) $\nsubseteq$ PA(current-goal, L(N)), then return <unfamiliar, N>
6. If PA(current-goal, L(N)) = IA(current-goal) and $\forall$ a $\in$IA(t) : t is unique in a given L(N) then top goal ; go to 4.

Extend description:

[1] This corpus is shared with the ATILF research unit (Analyse et Traitement Informatique de la Langue Franaise, UMR 7118) in the context of the regional collaboration "CPER Intelligence Logicielle".

```
7.  If current-goal ∈ terms(DM) then Ctxt
← DM else Ctxt ← DM + SM
8.  Chose an atomic formula p such Ctxt +
WKL ⊨ p
9.  If no such p exists then return <non
identifying, N>
10 For each o ∈ terms(p) - terms(L(N))
unstack(o, goals)
11.  N ← N' such as L(N') = L(N) ∪ {p}
12.  Go to 4.
```

The algorithm builds the semantic content and the syntactic tree of the referring expression simultaneously.

N is a partial syntactic tree with an empty place for the referring expressions; L is a function that gives all the properties denoted by N. The output of the algorithm is a syntactic tree and a classification of the description as *uniquely identifying, non-uniquely identifying, unfamiliar*.

The structure of the algorithm is the following: a goal stack contains the entities to be described. After the initialisation (1-2), they enter the main loop wich terminates successfully when the goal stack is empty (3). Otherwise, the algorithm examines the top entry of the goal stack (4-6) and if necessary extends the description (7-12).

The main strategy of this algorithm is to extend the description untill the unicity condition is satisfied. (i.e. the set P(A) must include the set I(A)) (6).

The algorithm fails if there is no uniquely identifying description (i.e. the unicity condition is not satisfied lines 8-9) or if the familiarity condition is not satisfied (the referent cannot be linked to a previously mentioned entity, i.e. the set of potential anchors does not include the intended anchor, line 5).

What is important to us is that *the model of the context comprises different knowledge bases used to build inferences* and to generate bridging descriptions. When a referent of the speaker model is not already in the discourse model, if the world knowledge says that it is linked with a previously mentioned entity, a bridging definite description is generated. We will show that this kind of structurating of the model of the context allows the use of the data given by the corpus study to propose an extension for definite and demonstrative coreferring expressions.

4. Summary of corpus study results

The corpus study was conducted on about 10 000 noun phrases, annotated by hand with the annotation tool MMAX (Muller, Strube, 2001). This number is significant when compared to similar studies such as the one by (Poesio, Vieira 1998) which was conducted on about 1500 NPs. The annotation was completely manual because we have no automated solution to annotate coreferring noun phrases with the information we were looking for. However, to facilitate the annotation, we took advantage of the morphosyntactic annotation contained in this corpus (Beaumont et al. 1998; Lecomte, 1997) for automatic recognition of the noun phrases with Gsearch, the chunker of (Corley et al., 2001).

The corpus study showed that the coreferential noun phrases can be classified into two categories : information repeating anaphors (IRA) and information adding anaphors (IAA).

IRA is the category of coreferential descriptions that repeat given information about the referent. This given information can come from different knowledge bases, some of them requiring inferences. Indeed, to produce the coreferring expression, the speaker can use some information already known by the hearer, and this information can come from several kind of knowledge and can be built via an inferential process. This category of coreferring noun phrases represents about 90% of the coreferring expressions of the corpus. We summarise the results in table 1.

The different knowledge bases identified through our corpus study are the following: the antecedent, the linguistic context of the referring expression, lexical knowledge and world knowledge. This lead us to construct five subcategories of IRA which are the following:[2]

The first subcategory of IRA uses explicit information given in the antecedent, i.e. the words used in the anaphoric noun phrase are the same as in the antecedent. This is the largest category of coreferential descriptions found in the corpus (40%).

(1) *It would have built a network of **ambiguous links in the police**. The trial of members of a neo-nazi organisation, (...), had highlighted **these links**.*

The second subcategory uses information explicitly given in the context and the antecedent. This means that the speaker uses inferences from the context to build the coreferring noun phrase.

(2) *The French bosses have strongly modified **their behaviour**. (...) **This new behaviour** is explained by two facts (...)*

The third subcategory uses information inferred from a lexical relation between the antecedent and the anaphoric noun phrase.

(3) *Every year, India suffers more and more from **drought** (...). **This phenomenon** became more marked because of incorrect economical choices.*

The fourth subcategory uses information inferred from both lexical relation and context.

(4) *The town council recently built a **splendid Concert Hall**. The ceremony took place in **this brand new and comfortalbe building**.*

The last subcategory uses information inferred from the world knowledge of the speaker and is quite sizeable (20%).

(5) *No report will be made about **M. Honecker visit to the graveyard of Neunkirchen, where his parents are buried**. It was decided because the chancelor asked for peace during **this private part of his trip to the Federal Republic.**.*

[2] All the examples given in this section are inspired by the corpus, (but translated and simplified).

The category of coreferential descriptions used to add information (IAA) about the referent represents about 9% of the coreferential descriptions of the corpus. It is divided into four classes, with respect to the linguistic means used to add information. These include the following: hyponymy (the head of the coreferring noun phrase is an hyponym of the antecedent (example 6)), modifiers (the new information is contained in the modifiers, example 7), hyponymy and modifiers (example 8), and finally a noun phrase without explicit relation with the antecedent (i.e., the link between the two noun phrases is constructed by world knowledge and accommodation, example 9).

(6) *The previous night **torrential rains** fell on the capital(...). The inhabitants expressed their happiness : **The monsoon** had finally arrived !*

(7) *The Israeli air force carried out a raid over **the Palestinian refugee camp of Ain-el-Heloue, in the suburbs of Saida, the main town of south Lebanon**(...). The planes executed several attacks over **this camp which counts sixty thousands inhabitants**, (...).*

(8) *In Roubaix (...), **the employees** feels as a referee. La Lainire is supressing the shuttle service ! For **these female workers from the coalfield**, this new (...).*

(9) *And if **Carl Lewis** was condemned to fight against the pipe dream of modern sport ? **This little boy who suffered from growing pains** became an adult, a tremendously gifted athlete (...).*

Category	Number	Proportion
IRA	**1613**	**91,1%**
antecedent	734	41,4%
ant. + context	155	8,75%
lexical relation	297	16,8%
lex. rel. + context	82	4,6%
WKL	345	19,5%
IAA	**159**	**8,9%**
hyponym	9	0,5%
modifiers	41	2,3%
hypo + mod	5	0,3%
NP	98	5,5%
Total	1771	100%

Figure 1: Results of the annotation

5. Proposal for an extension

The corpus study showed that there was no difference in the choice of the content of a coreferential definite or a coreferential demonstrative description, and that the determiner is chosen once the content and the syntactic realisation of the content are produced. The comparative study of definite and demonstrative noun phrases shows that the functions (IRA and IAA) are used in the same proportion with both determiners, the preferred sources for inference are the same for both, and the preferred linguistic means of adding information are also the same for definites and demonstratives. This result shows that the content of a coreferring expression in terms of new or given information has no influence on the choice of the definite or demonstrative determiner in French.

As a consequence, our proposal for an extension of the algorithm is used for both types of noun phrases. Moreover, it takes into account two facts : first, we propose using a structured model of the context to rank the different possibilities of generating IRAs. Second, we introduce the possibility of generating coreferential description adding information about the referent.

5.1. Information-repeating anaphors

In order to generate IRAs, we use the various possibilies found in the corpus. Our results lead us to structure the context into more databases than Gardent and Striegnitz. These databases are: the discourse model, divided into the semantic representation of the antecedent and the global semantic representation of the previous discourse; lexical databases such as Wordnet (Fellbaum, 1998) and Framenet (Baker et al., 1998) which contain the standard lexical relations of hyperonymy, synonymy, hyponymy; and a knowledge base containing general world knowledge. As shown in previous sections, these knowledge bases can be combined to build the necessary inferences. In order to generate the coreferring expression, a property identifying the referent is chosen in one or more of these databases.

Then, we propose to generate different paraphrases for coreferential noun phrases based on the frequency of the use of the different sources of inference in our corpus.

The information repeated in the coreferential noun phrase will be searched for in the different knowledge bases of the generator in the following order:

1. semantic representation of the antecedent
2. world knowledge
3. lexical databases
4. semantic representation of the antecedent and discourse model
5. lexical databases and discourse model.

The fact that the world knowledge appears in second position might seem problematic at the first sight. In fact, this is linked to the fact that many entities are named by proper nouns in first mention and are then mentioned by their ontological type or their job (when they are humans) something which is easy to encode in a database.

5.2. Information adding anaphors

In order to generate IAAs, we will introduce a new condition in the algorithm formalising this possibility, and in the input of the algorithm, we add the function of the coreferential noun phrase (IRA or IAA). The realisation of new information is performed in another module of the generator, the extended algorithm treating only the content determination of the referring expression.

5.3. Proposal for an extension

We modify the input by adding new knowledge bases and new data. The function of the description which is either *repeat given information - IRA* or *add information - IAA*, the semantic content of the description (for the first category of IRA), and a lexical database to compute the lexical relations.

Input:

WKL: world knowledge
DM: discourse model
SM: speaker model
t: target entity.
F(L): description function (IRA or IAA).
ϕ: semantic content of the antecedent.
LEX: lexical database.

Our extension can be described as follows : If the referent is in the discourse model, then generate a coreferential distinguishing description. If the speaker's goal is to repeat given information, infer the distinguishing description from the databases in this order : (1) semantic representation of the antecedent (2) world knowledge (3) lexical databases (4) semantic representation of the antecedent and discourse model (5) lexical databases and discourse model. If the speaker's goal is to add information, then generate a distinguishing description inferred from the speaker model.

More formally, this can be expressed as follows (we reuse the algorithm of Gardent and Striegnitz (2000) and modify lines 7 and 8):

Extend description:

7. If current-goal $\in$ terms(DM) then R $\leftarrow$ DM
8'. If F(L) = IRA, then choose an atomic formula p such that:
  $\phi \models$ p
  WKL $\models$ p
  LEX $\models$ p
  ϕ + DM $\models$ p
  LEX + DM $\models$ p
8". If F(L) = IAA, then choose an atomic formula p such that: Ctxt+WKL $\not\models$ p and SM $\models$ p
7'. If Ctxt $\leftarrow$ DM + SM [generation of bridging description]
9. If no such p exists then return <non identifying, N>
10 For each o $\in$ terms(p) - terms(L(N)) unstack(o, goals)
11. N $\leftarrow$ N' such that L(N') = L(N) $\cup$ {p}
12. Go to 4.

6. Conclusion and future work

In conclusion, this paper shows how the knowledge bases should be structured for a better generation of referring expressions. Of course, not all of the knowledge bases are already available, but some of them exist. For lexical relations, we have Wordnet and Framenet. For world knowledge, it seems that we have to build the databases with respect to the application for which the generator is used.

This study also shows how important inference is for the generation of coreferring expressions and the necessity of building and structuring resources to process the necessary inferences.

The main points for future work are the following:

First, we need to think about how we can build and put together the different knowledge bases used as inference sources. This is not a trivial problem, because for the moment, we can consult the existing databases but we cannot use them directly and in a coordinate fashion in computational applications.

Second, we need to test and evaluate the algorithm something that can be done only when the databases are available.

7. References

Baker C.F., Fillmore C.J., Lowe J.B. (1998), The Berkeley Framenet Project in *Proceedings of the thirty-sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*

Beaumont C., Lecomte J., and Hatout N., (1998) *Etiquetage morpho-syntaxique du corpus "Le Monde" pour les besoins du projet PAROLE*, Technical Report, INALF, Nancy.

Corley S., Corley M., Keller F., Crocker M.W., et Trewin S., (2001) Finding Syntactic Structure in Unparsed Corpora : The Gsearch corpus query system, *Computer and Humanities*, 35(2), pp81-94.

Dale R. Reiter E., (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions *Cognitive Sciences* 19(2), pp233-263.

Fellbaum C., *Wordnet. An electronic lexical database*, MIT Press, Cambridge, Mass.

Gardent C., Striegnitz K., (2000), Generating Indirect Anaphora, proceedings of *IWCS'00 (International Workshop on Computational Semantics)*.

Gardent C., Manuélian H., Kow E., (2003), Which Bridges for Bridging Descriptions, in proceedings of *EACL Workshop on Linguistically Interpreted Corpora*.

Lecomte J., (1997) *Codage Multext - GRACE pour l'action GRACE / Multitag*, Technical Report, INALF, Nancy.

Manuélian H. (2003), Coreferential Uses of Definite and Demonstrative Descriptions in French : A Corpus Study, proceedings of *European Summer School in Language, Logic and Information (ESSLLI) Student Session 2003*.

Muller C., Strube M., (2001) Annotating Anaphoric and Bridging Relations with MMAX, proceedings of *2nd SIGDial Workshop on Discourse and Dialogue*, pp 90-95.

Poesio M., Vieira R., (1998), A Corpus Based Investigation of Definite Description Use, *Computational Linguistics*, 24-2 pp183-216.