Paper 7

Issues for the Generation of Document Deixis

IVANDRÉ PARABONI AND KEES VAN DEEMTER
ITRI, UNIVERSITY OF BRIGHTON, UK
{Ivandre.Paraboni,Kees.van.Deemter}@itri.brighton.ac.uk

ABSTRACT. In this paper we describe issues for the computational generation of references to document parts such as sections, pictures, paragraphs etc, which we call *document deixis*. We discuss the circumstances under which these references need to be used in place of ordinary referring expressions and a number of problems in their generation. We also sketch a preliminary proposal for document deixis generation and point out the impact of generating such references on a traditional natural language generation system architecture.

1 Introduction

Consider a written document dealing with a certain domain of discourse. Typically, most referring expressions will refer to entities in this domain. The noun phrases in medicinal patient information leaflets, for example, talk about the medicine and the patient, as well as about such more abstract entities as actions, side effects, and remedies. But, in addition to this, the leaflets contain many noun phrases referring to entities of a very different kind: sections, paragraphs, pictures, etc., which do not 'live' in the domain of discourse. Instead, they are document parts. Examples include expressions like 'this leaflet', 'the above', 'the previous picture', and 'the section entitled How to Use Your Medicine', which are taken from the patient information leaflets in (cf. ABPI, 1997), as most of the examples in this paper.

Documents may of course refer to *other* documents. For example, a scientific paper usually refers to a number of other papers. We will discuss the phenomenon of documents referring to *their own* parts. More specifically, we will ask how such references may be generated and what kind of knowledge would be necessary to generate them. Remarkably, this problem seems to have received little attention so far, even though the related problem of how to generate *cross-media* referring expressions (as when the text refers to a picture) has been given a fair amount of attention (e.g., McKeown et al, 1992; André & Rist, 94).

Expressions like 'the previous section' are clearly deictic, since they involve an implicit or explicit reference to the immediate situation of the reader ('the section before the section you are reading now'). Accordingly, (Lyons, 1977) spoke of 'pure textual deixis', denoting references to linguistic forms, and (Webber, 1991) addressed instances of 'discourse deixis', denoting references to propositions, events, facts etc. (e.g., expressions like 'that problem', referring to a previously described situation). We will focus on references to textual and nontextual document parts such as sections, pictures, tables, lists etc., and we call such phenomena *document deixis* (DD).

Document deixis can be employed as a means to either describe the document itself in *meta-level* DDs like 'read section 5' or as part of a larger expression that serves to identify a domain entity in *object-level* DDs like 'the problem discussed in section 5'. Both cases involve a number of interesting theoretical problems stemming from the unusual nature of their referents.

A problem that has attracted much attention in the philosophical literature (e.g., Barwise & Etchemendy, 1987), is that of self-reference, which can cause a sentence to be *self-affirming* ('This sentence contains five words'), *self-defeating* ('This sentence contains exactly five words'), or *paradoxical* ('This sentence is false'). A variant of this problem is directly relevant for document generation. For example, consider a document D containing the sentence S saying 'Section 1.1 describes the side effects of your medicine'.

Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief at ESSLLI XI
E. André, M. Poesio & H. Rieser (editors).

There are two possibilities: Either this section does describe the side effects or it does not. In the first case, the truth of S follows from $D - S$ and, as a result, D contains a redundancy. In the second case, D becomes self-defeating. Avoiding self-defeating DDs while, on the other hand, allowing certain redundancies is one of the problems document generation has to face.

2 Document Deixis in Natural Language Generation

The two main types of document deixis, i.e., meta-level and object-level DDs, raise different issues for their generation by a Natural Language Generation (NLG) system. These issues are discussed individually in the next sections.

2.1 Meta-level Document Deixis

Meta-level DDs (e.g.,'Section 1.1 describes the side effects of your medicine') can be used to make statements about the document itself for various purposes. Such references may help the readers to find their way through structurally complex documents, link text and corresponding pictures, suggest additional information, emphasize relevant parts of the document *etc*. Meta-level DDs can be generated from a *Document Description* base which describes the document structure in terms of document entities such as sections (sec) and the domain entities they represent, such as symptoms (symp) or side effects (sfxs), which are both typical domain entities in patient information leaflets (cf. ABPI, 1997), as follows:

describe(sec1.1, sfxs1)
describe(sec1.1, symp1)
describe(sec2.1, sfxs2)
describe(sec2.3, sfxs2)
describe(sec2.4, symp2)
describe(sec2.4, symp3)

In a document structure with multiple hierarchical levels (e.g., sections and subsections), a Document Description base records only the lowest relevant level of description, e.g., an entity described in section 2.1 would not be represented separately as being described in section 2, although references to a higher level entity can be still produced, and may be even preferred under circumstances discussed later in this paper. A typical a Document Description base could be much more complex, including further domain entities (e.g., problems, discussions etc), new document entities (tables, pictures etc) and new relations (summarize, explain etc). Although in this paper we assume the existence of a Document Description base, the question of whether it can be generated automatically by an NLG system remains to be investigated. One way in which this might be done is by systematically keeping track of the correspondence between the input to the generator (i.e., semantic structures) and its output (i.e., the document structure).

There are many questions that cannot be discussed here. What, for instance, is the required level of detail for the Document Description base? (Is it necessary to record, for every domain entity mentioned in the document, at what exact places it is mentioned?) Perhaps the main challenge for generating these references lies in determining under what circumstances they are needed. After all, most of them are semantically redundant (see section 1).

2.2 Object-level Document Deixis

Object-level DDs are abbreviation devices not unlike anaphora, except that they seem to be immune from any of the known constraints on anaphora. For example, if a list of symptoms of some illness has been described somewhere in the document but no name has been attached to it, *recency* constraints prohibit the use of anaphora (e.g. 'they', or 'these symptoms') in case these symptoms become relevant at a much later place in the document. In such a case, a DD can be used as an indirect way of referring to the symptoms, *via* the document part that describes them: e.g., 'the symptoms described in Section 1.1'. The circumstances under which such indirect references are felicitous are largely unknown.

Ordinary referring expressions can be generated using the generic algorithm proposed in (Dale & Reiter, 1995), by increasingly aggregating properties to the head noun of the expression, until a uniquely identifiable description is obtained. Although in theory almost anything can be referred to via ordinary descriptions in this way, in practice long descriptions like

> 'the differences between medicines A, B and C on the one hand and the medicines D and E on the other',

where referring to each individual medicine may require a long description, can be infelicitous if the intended referent of the description (i.e., the differences) has already been mentioned in the document. In these cases, a document deixis like 'the differences mentioned in section 1' would be strongly preferred.

Conversely, we may also refer to almost anything via DD, although in practice long descriptions like 'the side effects in the second paragraph of the third subsection of the previous section' are hardly acceptable. Our current hypothesis about the choice between DD and ordinary referring expressions is that the choice is motivated by a Gricean maxim of Manner ('be brief') (Grice, 1975) which leads a writer to favor the shortest possible description (unless a special implicature is intended). In the setting of an NLG system, this could be implemented by generating both alternatives and then using a cost function to determine the choice between an ordinary description (making use of a noun and attached properties, e.g. 'the side effects of using medicine A and medicine B together') and an equivalent object-level DD (making use of a short description and layout features e.g., 'the side effects discussed in section 1.1'). In its simplest form, the cost function selects the description containing the smallest number of words.

Our preliminary corpus investigation suggests that simple domain entities such as 'the medicine' are preferably referred to via pronouns or definite descriptions, and so are complex domain entities in short-distance reference. DDs are preferably employed in order to refer to complex domain entities which to not have a single description (e.g., an arbitrary list of symptoms of a medicine without a distinctive shared property) or entities made explicitly salient in the document structure (e.g., a table identified by caption like in 'the medicines listed in table 1') for reference purposes.

Identifying a document entity which describes the intended referent is a nontrivial task, involving criteria such as the preference for optimally specific references. Note that (1) the most appropriate document part is not always the *smallest* document part containing the relevant information since this part may not be sufficiently salient, e.g. it may not have a header or number. Note also that (2) obviously, if a domain entity is described through all subsections of a document section, we may refer to either the whole set of subsections (e.g., 'the symptoms described in section 3.1 and 3.2') or to the main section as a whole (e.g., 'the symptoms described in section 3'), but even in cases where a given entity is described only in certain parts of a section, we may still refer, rather vaguely, to the section as a whole.

Simple object-level DDs could be generated by using a Document Description base (cf. the previous section). The information provided by a Document Description base can produce either a) a single description; b) multiple descriptions or c) no description at all. In the example, a reference to the side effects (sxfs1) could be made via (sec1.1), since (sec1.1) describes this entity and the fact that (sec1.1) also describes (symp1) does not affect the reference, which can be safely realized as 'the side effects described in section 1.1'. On the other hand, a reference to the side effects (sfxs2) produces two results, since the referred entity is described (perhaps redundantly) in both (sec2.1) and (sec2.3), and in this case we could refer to both sections or perhaps to a section at a higher level. Finally, references to the symptoms (symp2) or (symp3) cannot be easily produced, since (sec2.4) describes both entities. Such ambiguity may suggest that the referent is not sufficiently salient in the document structure (i.e., the referent lacks layout features which could make it distinguishable in the document and hence suitable for reference), necessitating a revision of the document.

DDs can be realized in very different ways, potentially involving spatial relationships (above / below), numbers, captions, etc. Here are some examples (cf. ABPI, 1997):

implicit reference 'as shown in the picture';

explicit reference 'the side effects described in section 2';

relative reference 'the following side effects';

mixed references 'the section 3 above'.

The choice for a specific realization depends on a number of factors such as saliency of the document referent and the document style itself. The surface realization of other document parts can also be relevant. For example, one can only refer to something by means of the expression 'the second picture' if it is known that it is realized as the second picture in the document. DDs may even be *media*-dependent, as in '(...) are shown on the next page', which may lead to inconsistencies if the realization of the reference itself moves the referred entity to another page.

3 Architectural Issues

Document deixis generation gives rise to some challenges to the architecture of an NLG system, mainly with respect to circularity and order of generation. The most noticeable of these issues is the dependency between layout and content of the text, since document deixis makes assumptions about the structure of the document. For example, in a pipeline NLG architecture as suggested in (Reiter, 1994), the realization has to wait until the later stages of the generation process, and the document structure (and the existent document entities) may not yet be available to be referred to. Thus, even the generation of simple references like 'read the following', which imply both the existence of the referred information and a particular realization (in this case, text), can be difficult.

Document deixis is largely dependent on layout features of the document, and might even be motivated by their existence. For example, the division of a document into sections may facilitate the use of meta-level references like '(See Section 3)' to give additional information on a specific topic. On the other hand, some layout features of the document seem to be specifically tailored for reference purposes, like the use of numbers to identify examples in scientific papers, which allow the use of references like 'Example 1.2.1. shows that...' etc. In both cases (layout motivated by the wish to refer and vice-versa) several architectural issues are involved, which are yet to be investigated.

Depending on how much layout information a DD expression conveys, it is possible that a two-phase generation process become necessary. (Compare (Inui et al., 1992) for a different motivation for a two-phase generation architecture.) In a first generation phase, the document and its DD expressions would be tentatively generated, and in a revision phase, the DDs are adjusted so as to correspond to the actual realization of the document. A two-phase procedure might be necessary, for example, in the generation of references to pages of the document or 'small' document entities such as sentences and words.

4 Conclusion

This paper has sketched the results of a preliminary investigation into the generation of document deictic expressions. We have distinguished between references to document entities *per se* (meta-level DDs) and references to domain entities *via* document entities (object-level DDs). Although each of these raises specific issues, there are also important commonalities between them. Both, for example, could make use of what we have called a Document Description base (Section 2.1), as we have argued, and both pose some common problems for traditional NLG architectures, mainly with respect to the dependency between document content and layout.

Following the current work, we will propose an algorithm for generating selected classes of the phenomena found in a number of language corpora, including the Patient Information Leaflets (cf. ABPI, 1997) which are used in a number of current NLG projects, including the GNOME project (funded by the UK EPSRC under grant reference GR/L51126) which focuses on the Generation of Nominal Expressions.

5 Acknowledgments

This work is part of an ongoing Phd project of the first author started in September 1998, and it has been supported by the CNPq, the Brazilian Research Council.

Bibliography

The Association of the British Pharmaceutical Industry: 1997, *1996-1997 ABPI Compendium of Patient Information Leaflets*

Elisabeth André and Thomas Rist: 1994, *Referring to World Objects with text and Pictures.*, Proceedings of Coling'94

Nicholas Asher: 1993, *Reference to Abstract Objects*, Kluwer Academic Publishers, Dordrecht, The Netherlands

Jon Barwise and John Etchemendy: 1987, *The Liar; An Essay on Truth and Circularity*, Cambridge University Press

Robert Dale and Ehud Reiter: 1995, *Computational interpretations of the Gricean maxims in the generation of referring expressions*, Cognitive Science (19)

Steven K. Feiner and Kathleen R. McKeown: 1990, *Coordinating Text and Graphics in Explanation Generation*, Proceedings of the AAAI-90

H.P. Grice: 1975, *Logic and Conversation*, Cole and Morgan (Eds.) Syntax and Semantics, Vol.iii: Speech Acts. New York, Academic Press

K. Inui, T. Tokunaga, and H. Tanaka: 1992, *Text Revision: A Model and its Implementation*, 6th International Workshop on NLG, Trento

John Lyons: 1977, *Semantics*, Cambridge University Press

Kathleen McKeown et al: 1992, *Generating Cross-References for Multimedia Explanation*, Proceedings of the AAAI-92

Ehud Reiter: 1994, *Has a consensus NL generation architecture appeared and is it psycholinguistically plausible?*, Proceedings of the 7th International Natural Language Generation Workshop, pp.163-170, Maine U.S.A.

Bonnie Lynn Webber: 1991, *Structure and Ostension in the Interpretation of Discourse deixis*, Language and Cognitive Processes 6(2) May 1991, pp. 107-135