Using Natural Language Generation Techniques to Produce Virtual Documents

Robert Dale†, Stephen J Green†, Maria Milosavljevic‡, Cécile Paris‡, Cornelia Verspoor† and Sandra Williams†

†MRI Language Technology Group
Macquarie University
Sydney NSW 2109 Australia
{rdale,sjgreen,kversp,swilliam}@mri.mq.edu.au

‡CSIRO Mathematical and Information Sciences
Locked Bag 17
North Ryde NSW 1670 Australia
{Maria.Milosavljevic,Cecile.Paris}@cmis.csiro.au

Abstract

With the increasing importance of Web publishing, there has been considerable interest in the production of virtual documents on demand. The bulk of this work has used existing documents annotated with meta-data as a source. We suggest that more flexibility and functionality can be obtained if virtual documents are generated instead from raw data. This capability can be achieved by using natural language generation techniques. In this paper, we describe a project concerned with automatically generating natural language descriptions of museum artefacts directly from a museum's Collection Information System.

Keywords Natural Language Generation, Databases

1 Introduction

Natural language generation (NLG) is concerned with the development of techniques for producing linguistic output, whether written or spoken, from some underlying information source. While some of the aims of NLG can be seen as similar to those of the work in document computing and management (e.g., the production of virtual documents on demand), the process and resources it uses are different. In document computing, the emphasis has been on controlling the authoring of a text in order to annotate it with the meta-data which will enable a system to later reason about that text, potentially creating a new text from it (e.g., [8]). Document computing thus deals with *existing* texts, reasoning about its meta-data to form a new text on demand. In contrast, an NLG system reasons about some *information* to be conveyed to the reader, the purpose for which it needs to be conveyed and the intended reader (listener) to construct a text from scratch. It does so by employing *linguistic resources*, which capture principles of communication, information about how to refer to some concept, and how to form sentences and paragraphs. The input to a generation system is not a *text* or *paragraphs*, but some "underlying" information, considered to be of a more conceptual or semantic nature. The linguistic resources then allow for the mapping of this semantic information to a text. Unlike a parser or a translation system, an NLG system does not need to understand the meaning of a text. Rather, it must reason about how to express information linguistically. By starting from information as opposed to text, and by constructing the text from the underlying information, NLG technology offers a number of important benefits, including:[1]

- description of internal data: given the appropriate linguistic resources, an NLG system can automatically produce text to describe a wide range of internal data. For example, an NLG system can be employed to explain the contents of a database, to produce reports describing numbers (e.g., stock reports, weather reports, etc.), to explain the reasoning of an expert system, or to provide documentation for a program.
- contextual tailoring: the generation process can make use of information only available at the point of use (such as characteristics of the particular reader, or information about

Proceedings of the Third Australian Document Computing Symposium, Sydney, Australia, August 21, 1998.

[1] Of course, it is important to recognise that NLG techniques are only appropriate when underlying information is available. If all that is available are existing texts, the technology may not be applicable.

the content of recent interactions the user has had with the system) to create texts that are tailored to specific requirements. Since the process is not constrained by existing text fragments, the range of texts that can be produced is potentially very large and not producible by other means.

- up-to-date reporting and documentation: if descriptions of the information source are created automatically and dynamically, there is no requirement to update such descriptions manually, with the attendant problems of errors and time lag.
- multilinguality: if the underlying information source is not expressed in terms of a particular natural language, then it is possible to generate descriptions of the same information in different languages automatically.

In NLG, a great deal of research has been carried out to explore the technical requirements that need to be met to provide these capabilities. It is clear from that research that it is possible to construct the appropriate engines and linguistic resources to allow for the automatic production of virtual documents, on demand, providing the benefits mentioned above. Little work, however, has been done to ensure that the input required by a generation system was available (when it is not simply a set of numbers). Indeed, typical NLG systems take as input an AI-style knowledge base, and, in much of the work, proof-of-concept prototypes were built using small knowledge base samples often manually constructed for experimental purposes—see, for example, [1, 3, and 7].[2]

Most digitally encoded information is not, however, available in such richly structured and annotated form. If this technology is to make a significant impact, and if we want to be able to exploit it to produce virtual documents, then we need ways of using it in conjunction with *existing* sources of information. In particular, it must be possible to exploit existing databases. This is the issue we address in our work. In particular, this paper presents some results from an experiments we have been pursuing in using a real database as a source of information for the production of virtual documents on demand. Our particular goal is the automatic description of the contents of a museum Collection Information System (CIS).

The paper is structured as follows. We first briefly describe in Section 2 the architecture of an NLG system and show how NLG can be used to produce virtual documents on the fly. In Section 3, we present our work on generating texts from a knowledge source derived completely automatically from a museum's database. Finally, in Section 4, we draw some conclusions.

2 Generating text with a generation system

Figure 1 shows the architecture of a conventional NLG system and how it can be straightforwardly adapted to produce virtual documents on the Web. The system begins with a *discourse goal*, which captures the purpose of the text to be generated. In our scenario, the discourse goal is a user request either to describe a single museum object or to compare two museum objects. Based on this discourse goal, the system selects from its plan library (which is part of the system's linguistic resources) a discourse plan to employ to satisfy the goal. These discourse plans are based on the notion of discourse schemas introduced by [4], but modified for use in a hypertext environment, by indicating in which circumstances a hyperlink is appropriate, and what discourse goal is to be given to the system when clicking on that link.

After selecting a discourse plan, the text planning component instantiates it with *facts* from the knowledge base. These facts constitute the information which will be contained in the resulting text. For a first experiment, the knowledge base was hand-constructed as done in other NLG systems.

A user model is used to keep specialised information about particular users. It is the user model which allows for the tailoring of texts depending on the user. In our current system, the user model is exploited in two ways: to provide two different views of the object hierarchy for naïve and expert users, and to modify how descriptions and comparisons are presented.

A record of the discourse is also maintained for each user, and is exploited in combination with the user model in order to improve the conceptual and textual coherence of descriptions. For example, if the user has knowledge of an entity (as recorded in the user model) or has been told about an entity (as recorded in the discourse history), then the entity can be referred to in later descriptions where a comparison can be made with that entity [5]; Figure 2 shows a comparison between the Difference Engine and the Analytical Engine produced by this mechanism.

Once the text planning component has pulled together all the information about the entity (or entities) to be described in the document according to the user's knowledge and the discourse history, the filled discourse plan is passed to the surface

[2] However, see [6] which also focuses on the automatic acquisition of the knowledge required to produce texts, in order to make language generation technology more practical. In particular, the authors present a system capable of obtaining the knowledge required to generate software documentation directly from the software specifications for that application.

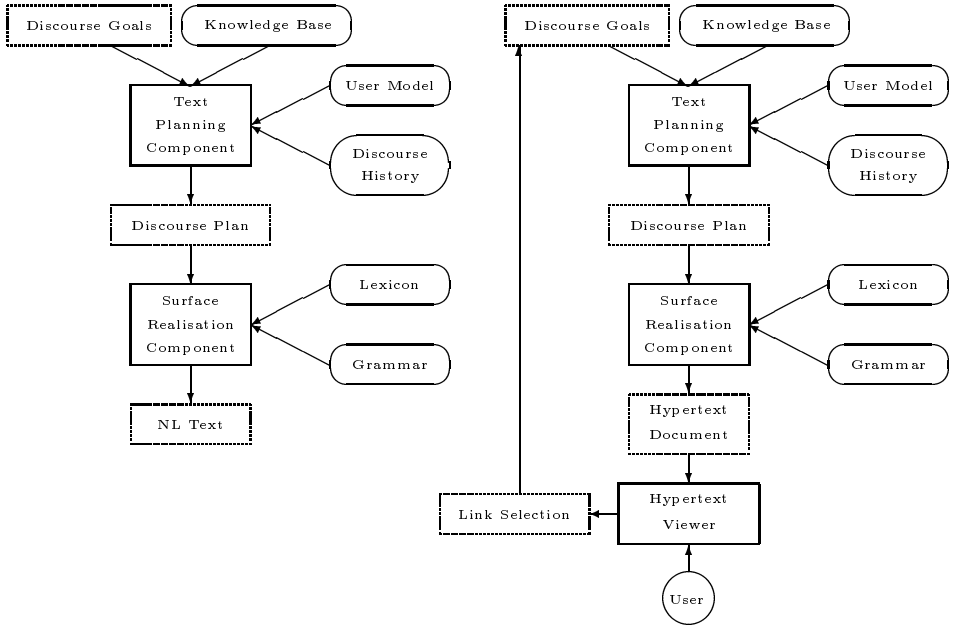

Figure 1: System architectures: (a) traditional NLG; (b) dynamic hypertext.

Figure 2: A text generated by Power

realisation component. Here, the discourse plan is realised as natural language sentences, and HTML tags are positioned within the text to allow the user to request follow-up questions by selecting them. When the user selects a hypertext link within the description, a new discourse goal is posted to the text planning component, and the cycle repeats.

We can readily see from this scenario how *nlg* allows for the production of virtual documents on demand, and the advantages it may offer compared to the creation of new documents from existing ones: by creating the text from raw data according to linguistic principles, and by making use of a user model and discourse history, it can ensure the generation of an appropriate text for the specific user and situation, and the smooth transition from one description to the next. This functionality thus provides a more natural discourse between the user and the system [1, 5]). This idea is the one we explored when building the PEBA-II system, a system which provide interactive dialogues with databases using the Web as a delivery vehicle [1, 2, 5]. Finally, NLG offers the capability of constructing the text in a different language, given the appropriate linguistic resources (e.g., grammar and vocabulary).

Having demonstrated the idea of using nlg to produce virtual documents on the web, we then examined the issue of doing it from an existing database, as opposed to a carefully crafted knowledge base, one in which we encoded precisely the kinds of information we needed to generate the texts we were aiming for. Our aim here was twofold. One the one hand, we want to see the types of texts that can be generated when using a real database. On the other hand, we also want to draw some conclusions as to how the databases should be constructed in order to take advantage of what NLG can offer in terms of the production of virtual documents on demand.

3 Generating from a Real Database

We are concerned with producing descriptions of the Powerhouse artefacts on the web, starting with their Collection Information System (CIS). The Powerhouse Museum's CIS is a database of the 200,000 objects the museum owns, although

```
<rec num=12798 id="H4448-513">
OID: H4448-513
INT: Part
LOC: TH2.STEP.6A
LOD: 27/11/1997
OBN: Boots
OBS: Balmoral boots, elastic sided, pair, women's,
patent/kid/leather/elasticise d fabric/wood,/brass prize work,
[Gundry & Sons], England, c.1851; 1862-1869. DES: Balmoral boots,
elastic sided, pair, women's, patent / kid / leather /
elasticised fabric / wood /brass, prize work, [Gundry & Sons],
England, c.1851; 1862-1869.  Pair of women's elastic sided
boots (Balmoral), with wooden filler, of welted construction with
rounded toes featuring peaked caps and stacked heels.  The
uppers consist of a patent golosh, seamed at the back, glace kid
leg, seamed at front and back, and elastic sides extending to the
golosh. The uppers are decorated with oval stitching at the edge
of caps and scallops at the throat of golosh.  The leather heel
is fine wheeled, featuring a top piece with brass nailed edge.
The black leather sole features a sueded forepart with brass
nails, as well as an internal clump and brass hinged section for
extra strength and a brown polished ridged waist with black edge.
Reputed to have been made by Gundry & Sons.  (See object file for
specialist report by June Swann)
MDE: Gundry & Sons; London, England
MDN: 1965 list says "made by Gundry & Sons, Soho Square."  Swann
says hinged device to increase flexibility is unusual. Similar
screws on H4448-515.  Note hinged sole in 1862 exhibition.  She
finds no information about Box in information she has about the
1851 exhibition, though William Walsh is mentioned in connection
with a pair of shoes.  Patent 558, 5 March 1861, granted to J.M.
Carter, a similar sole with 2 cuts across the tread and 4 rows
of screws "for soldiers, riflemen, sportsmen.  The inner sole is
whole and contains pitch."  It is not possible to confirm whether
these boots contain pitch.
DAT: c 1851 - 1869
MAR: Interior obscured by last, no marks on exterior
DIM: Length 248 mm Height 31 mm Overall Height 160 mm Width 58 mm
</rec>
```

Figure 3: A database record

for our pilot study we narrowed our focus to the approximately 5,000 objects that are actually on display on the museum floor (as with many museums, most of the collection is in storage).

Figure 3 shows a typical record from the database. This record contains information about the Balmoral boots.

As this figure shows, a lot of the information contained in this record is of a *textual* nature. As mentioned in the introduction, an NLG system does not start with text an input, as it is not capable of understanding the meaning of a text fragment. An NLG system needs *facts* as input, as *semantic* units of information. Given records like the one shown in the figure, then, we must do some processing to obtain these facts.[3] They can then serve as input to the generation process. We thus studied how much semantic information could be automatically acquired, and applied NLG techniques to that information base. Figure 4 shows a text generated from the information obtained automatically from this database record. It is clearly of a less sophisticated nature than the text shown in Figure 2. This is largely because the knowledge base created automatically from the database record is not as sophisticated, structured or rich as the knowledge base created by hand for the purpose of generating descriptions. Yet, even with only that amount of information, some of the benefits of producing texts on demand using NLG techniques can be seen. For example, the text can be generated in different languages on demand, as shown in Figure refpowertng-output-sp-fr, provided the appropriate linguistic resources are available.[4]

[3] While some of this processing may be considered as *parsing*, we do not claim that the system *understands* the text fragments.

Figure 4: A text generated by PowerTNG

'Balmoral boots'
Estos son 'balmoral boots'. Fueron hechos entre los años 1870 y 1875. Fue produdido en London, Inglaterra. Estàn hechos de cuero, 'patent leather', 'glace kid', lino y madera. Los 'Balmoral boots' tienen 45 mm de altura, 255 mm de largo, 55 mm de anchura total, 150 mm de altura total y 30 mm de ancho.

'Balmoral boots'
Ces objects sont des 'Balmoral boots'. Ils ont été fabriqués en entre 1870 et 1875. Ils ont été fabriqués à 'London'. Ils sont en 'leather', 'patent leather', 'glace kid', 'linen' et 'wood'. Les 'Balmoral boots' ont 45 mm de hauteur, 255 mm de longeur, 55 mm de largeur totale, 150 mm de hauteur totale et 30 mm de largeur.

Figure 5: Texts in different languages: Spanish and French

We now look at how we obtained information from the database. The Powerhouse museum provided us with a dump of their database in ASCII format with the fields in the database records indicated by tags at the beginning of each field. They also provided us with a thesaurus of object types. This is the only information we had available to produce a structured knowledge base from which to produce text. To be able to obtain a knowledge base that can serve as input to the generation process, we processed the data file as follows:

1. normalisation of the database: This is to ensure that each record is surrounded by an

[4] In this figure, the words in single quotes indicate that the word is in English, as the appropriate lexical item for the language of the generated text has not yet been provided.

SGML-style `rec` tag and that each field of an entry is on a single line.

2. extraction of dimensions: In this step, a Perl script extracts the dimensions of the objects. This information resides in easily identifiable fields (e.g., the DIM field in Figure 3) and the information in that field is structured and can be decomposed into its subfields (e.g., length, height).
3. extraction of thesaurus categories: This step involves trying to identify the thesaurus category that applies to each of the objects in the database. This is normally found in the OBN (Object Name) field and corresponds to an entry in the Powerhouse's thesaurus.
4. extraction of names, materials, makers, locations, and dates of construction: This involves extracting information from the textual information contained in the database records. Most of our work here so far has focussed on the OBS (Object Statement) field. This field is supposed to include information encoded in a standardised and rigorous way. However, in practice, not all the information that is supposed to be included is present, or it is present in a different order, or format, from the norm. Yet, with the help of information from the thesaurus, we were able to identify information such as date of manufacturing or purchasing, materials and location.
5. extraction of PART-OF and A-KIND-OF information: We use the OID (Object ID) field to determine the PART-OF hierarchy for the database. For example, in the database record shown in Figure 3, the OID `H4448-513` indicates that this object is the 513^{th} part of the object with OID `H4448` (in this case the Balmoral boots are part of a large collection of footwear). According to the database specifications, an object may have parts, sub-parts, and sub-sub-parts.

The result of this processing is the construction of an information record such as the one shown in Figure 6, from which a knowledge base suitable for input to an NLG system can be produced.

4 Discussion

From our experiments, we conclude that NLG techniques are appropriate to produce virtual documents on demand, and that in fact they offer a number of advantages over techniques dealing mostly with already existing documents, in specific situations where underlying information is available. We have briefly presented some proof-of-concept prototypes to this effect. In order for that technology to be appropriate, however, one must address issues related to the source of information. We have presented some work concerned precisely with these issues. In that work, we learnt that, to obtain enough information to allow for the production of sophisticated texts from existing databases, special care must be taken when the database is designed and populated. Lack of structure and consistency in the database, and possibility to include large amount of unstructured textual information in the fields all contribute not being able to obtain much information from a database to serve as input to the generation process. Yet, we believe this obstacle not to be insurmountable in many cases. Indeed, the features required (e.g., consistency and structure) are important for well-designed databases, and it is important to note that these features do not necessarily impose more constraints on the end-users, given appropriate interfaces and tools.

```
OBS.original: Balmoral boots, elastic sided, pair, women's,
patent/kid/leather/elasticised fabric/wood,/brass prize work,
[Gundry & Sons], England, c.1851; 1862-1869.
OBS.object: Balmoral boots
OBS.object.number: plural
OBS.material.1: patent
OBS.material.2: kid
OBS.material.3: leather
OBS.material.4: elasticised fabric
OBS.material.5: wood
OBS.production.country: England
OBS.create: 1851
OBS.create.inexact: 1
```

Figure 6: Data extracted from an Object Statement field.

It is also quite possible that there will be fewer problems of this kind in the future: as application programs become more sophisticated, it is likely that their underlying representations will have the characteristics required and that their content will move closer to the kinds of rich symbolic structures expected in AI systems. It is also possible that increasingly sophisticated data input tools will be developed to enable the construction of such knowledge bases (see for example, [6]), so that database entry clerks do not have to acquire the skills of knowledge engineers in order to do their jobs.

5 Acknowledgments

Thanks are due to Matthew Connell and Kevin Sumption from the Powerhouse Museum for their enthusiasm in supporting this project and to Des Beechey for supplying the Powerhouse Museum's databases.

References

[1] Robert Dale and Maria Milosavljevic [1996] Authoring on Demand: Natural Language Generation of Hypermedia Documents. In *Proceedings of the First Australian Document Computing Symposium (ADCS'96)*. Melbourne, Australia.

[2] Robert Dale, Jon Oberlander, Maria Milosavljevic and Alistair Knott [in press] Integrating natural language generation and hypertext to produce dynamic documents. *Interacting with Computers.*

[3] Leila Kosseim and Guy Lapalme [1994] Content and Rhetorical Status Selection in Instructional Texts. In *Proceedings of The International Workshop on Natural Language Generation*, pp. 53–60.

[4] Kathy McKeown [1985] Discourse strategies for generating natural-language text. *Artificial Intelligence* **27**:1–41.

[5] Maria Milosavljevic [1997] Augmenting the User's Knowledge via Comparison. In *Proceedings of the 6th International Conference on User Modelling.* Sardinia.

[6] Cécile Paris and Keith Vander Linden [1996] Building Knowledge Bases for the Generation of Software Documentation. In *Proceedings of the 1996 Meeting of the International Association for Computational Linguistics (COLING-96).*

[7] Dietmar Rösner and Manfred Stede [1992] TECHDOC: A system for the automatic production of multilingual technical documents. In *Proceedings of KONVENS-92.* Springer. Berlin. Also available as technical report FAW-TR-92021, FAW, Ulm, Germany.

[8] Anne-Marie Vercoustre, Jon Dell'Oro, Brendan Hills [1997] Reuse of Information through Virtual Documents. In *Proceedings of the Second Australian Document Computing Symposium*, Melbourne Australia, pp55-64.