Dynamic Document Delivery: Generating Natural Language Texts on Demand

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

Research in natural language generation promises significant advances in the ways in which we can make available the contents of underlying information sources. Most work in the field relies on the existence of carefully constructed artificial intelligence knowledge bases; however, the reality is that most information currently stored on computers is not represented in this format. In this paper, we describe some work in progress where we attempt to generate large numbers of texts automatically from existing underlying databases. We focus here in particular on the automatic generation of descriptions of objects stored in a museum database, highlighting the difficulties that arise in using a real data source, and pointing to some possible solutions.

1 Introduction

Natural language generation is concerned with the development of techniques for producing linguistic output, whether written or spoken, from some underlying information source. By providing this capability, the technology offers a number of important benefits, including the following:

- 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;

- contextual tailoring: since the texts can be generated on-demand, 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 the content of recent interactions the user has had with the system) to create texts that are tailored to specific requirements.

A great deal of research has been carried out to explore the technical requirements that need to be met to provide these capabilities. Much of this work, however, has focussed on underlying representations in the form of AI-style knowledge bases, and often these are small samples which have been hand-constructed for experimental purposes— see, for example, [13, 5, 10 and 11]. Most digitally-encoded information is not, however, available in such richly structured and annotated form. Furthermore, even where the information is encoded using an AI knowledge representation language (as is the case, for example, in expert systems), it is still generally the case that this knowledge has to be augmented in various ways for it to be usable by an NLG system. If this technology is to make a significant impact in the medium-term, then we need ways of using it in conjunction with existing databases of information.

This paper presents some results from experiments we have been pursuing in using real databases as a source for the generation of natural language texts. Our particular goal

Figure 1. A text generated by Power

is the automatic description of the contents of a museum Collection Information System (CIS). In Section 2, we describe an early prototype system we developed to show the kinds of texts that might be generated in the museum domain given an appropriate encoding of the relevant knowledge. In Section 3, we present some work in progress which uses a knowledge source derived completely automatically from the museum's database. In Section 4, we highlight and discuss the problems that arise in achieving quality results from real data, and finally, in Section 5, we draw some conclusions and point to some ways forward.

2 Generating from a Hand-Crafted Knowledge Base

We took as our starting point the PEBA-II system [7], which describes and compares animals. Using a sophisticated underlying information source, PEBA-II explores how we might build interactive dialogues with databases using the Web as a delivery vehicle [9].

By taking the core components of the PEBA-II system and adding a small hand-constructed knowledge base of museum objects, within a relatively short time we were able to develop our initial prototype system for describing and comparing museum objects. Called Power, this enabled us to demonstrate to our partners at the Powerhouse Museum the potential of NLG technology.

The system begins with a discourse goal, which, in this scenario, 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 a discourse plan which can be used to satisfy the goal. These discourse plans are based on the notion of discourse schemas introduced by [7], but modified for use in a hyper-

text environment.

After selecting a discourse plan, the text planning component instantiates it with facts from the knowledge base. For this experiment, the knowledge base was hand-constructed.

A user model is used to keep specialised information about particular users in order to modify how descriptions and comparisons are presented. A record of the discourse is also maintained for each user, and is used in combination with the user model in order to improve the conceptual 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 used by the system in later descriptions where a comparison can be made with that entity (see [8]); Figure 1 shows a comparison between the Difference Engine and the Analytical Engine produced by this mechanism.

The discourse history is also utilised to improve the textual coherence of descriptions [3]. For example, the entity which is currently being described (the focused entity) can be related to the most recently described entity in order to smooth the transition from one description to the next. This functionality provides a more natural discourse between the user and the system (see [1] and [2]).

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, the filled discourse plan is passed to the surface 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. If a picture exists for the focused entity, the surface realisation component includes it in the hypertext page. 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.

3 Generating from a Real Database

Our next step was to see how this prototype could be used in conjunction with a real database. The knowledge base used in the prototype was, as we noted, hand-constructed. This allowed us to encode precisely the kinds of information we needed to generate the texts we were aiming for. However, constructing such a knowledge base by hand is unrealistic for large scale information sources.

The Powerhouse Museum's CIS is a database of the 200,000 objects the museum owns, although 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). Because all of the parts of an object may not be on display,

```
<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 2. A database record

we have supplemented these 5,000 objects with any objects that are part of a display object, and any objects that have a display object as a part. The resulting database contains 15,483 records.

3.1 Inputs and Outputs

Figure 2 shows the database record corresponding to the Balmoral boots, one of the objects displayed in the Powerhouse Museum. Figure 3 shows a text generated from this database record. It is clearly of a less sophisticated nature than the text shown in Figure 1. 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 obtaining this amount of information directly from the database records was not a trivial task, as discussed in the following sections.

3.2 Processing 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.

We break down the processing of the data file down into a number of steps:

Figure 3. A text generated by PowerTNG

1. normalisation of the database;

2. extraction of dimensions;

3. extraction of thesaurus categories;

4. extraction of names, materials, makers, locations, and dates of construction; and

5. extraction of PART-OF and A-KIND-OF information.

Normalising the database consists simply of ensuring that each record is surrounded by an SGML-style rec tag and that each field of an entry is on a single line. In the second step, the normalised database is run through a Perl script that extracts the dimensions of the objects. This information resides in easily identifiable fields (e.g., the DIM field in Figure 2) and the information in that field is structured and can be decomposed into its subfields (e.g., length, height). The dimension information is output as a set of extra fields in a given record.

The next 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. This is a straightforward process, which simply involves a look-up table connecting object names to thesaural categories.

The next part of the processing 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. In reality, of course, what a human (in this case a museum curator) considers to be a rigorous specification is not rigorous enough to be fully exploited by a computer program, and each person entering information may use different methods. In theory, the OBS field contains (in a comma

3

separated list) the name of the object, what it is made of, a list of materials, who it was made by, where it is made and the year in which it was made. However, in practice, not all this information is present, or it is present in a different order, or format, from the norm.

To aid in the processing of the data, we constructed a set of about 280 materials from an earlier sample. To this we added a list of 205 country names acquired from the machine readable Macquarie Thesaurus. These resources formed the basis of our lexical resources for generation; but more importantly they allowed us to gain a foothold on the data, enabling identification of materials and location information in the OBS field.

This information is then used by a set of Perl scripts to identify the required elements of information in the OBS field. For example, the script tests whether a field looks like a date or a range of dates in order to determine when the object was made.

As with the previous stages, the information extracted at this stage is written as an additional set of database fields. The extracted information from the record in Figure 2 is as follows:

```
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
```

Finally, 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 2, the OID H4448-513 indicates that this object is the 513th part of the object with OID H4448 (in this case the Balmoral boots are part of a large collection of footware). According to the database specifications, an object may have parts, sub-parts, and sub-sub-parts.

The resulting expanded database is then used to generate a knowledge base and lexicon for use by the NLG system.

4 Issues

Our experiments so far have allowed us to identify a number of important issues.

4.1 Sparse data issues

Of the 15,483 records that we received in the database dump, only 9,887 (in other words, around 64% of the total) actually have an OBN field. Furthermore, of the 9,887 objects that have OBN fields, only 7,751 are valid object names (i.e., names which appear in the museum thesaurus). Thus,

about 50% of the database entries do not provide any information about the types of the objects. This clearly causes a problem in terms of automatically constructing a taxonomy to be exploited in generating object descriptions, as the type of an object plays a major role in descriptions.

Thus, sparsity of data in a real database has a big impact on how much information one can hope to extract directly from the database, not only to fit an object into a taxonomy, but also to be able to generate information about it in a description.

4.2 Data quality issues

We rapidly found that there is a tradeoff between extracting a limited quantity of high quality data and extracting a large quantity of poor quality data. It was our policy to always opt for the former to ensure our data is always of high quality. This is necessary to ensure that the text that can be generated from the data is sensible, albeit simple and short.

It is important to note that we have rigidly kept our extraction process entirely automatic throughout the whole experiment. It is easy to fall into the trap of hand-crafting special rules to fit in with the idiosyncratic nature of a database. However, if this is attempted, sets of special rules might clash with and undo the work of the general rules. Furthermore, large numbers of special rules can soon become unmanageable and make the system hard to maintain. Finally, there are no guarantees that the rules will be applicable for subsequent releases of the data.

4.3 Difficulty of data extraction

Much of the data that we have extracted so far is of a fairly mundane nature (such as dimensions of museum objects). It seems likely that the really interesting data is located in the free text fields of the Powerhouse database. For instance, in the MDN field of the database fragment in Figure 2, we note that the unique feature of the Balmoral boots is that they have a hinged device to increase flexibility. This kind of information is so diverse that it is nearly impossible to predict accurately enough for information extraction. One solution would be to extract the entire field for use as canned text (similar to the *stories* used in the *ilex* system; see [4]). However that would not be feasible in this case (and many others) because the free text in the MDN field is generally ungrammatical, and it may contain information that the museum does not wish to be on public view.

4.4 Database structure issues

From this experiment, we also learned some lessons about database structure, if databases are to be used as the source of information for natural language generation. As

4

we mentioned at the beginning of this paper, natural language generation offers a number of new prospects in terms of information delivery, such as its ability to tailor the output to a specific user and situation, and its potential for multilinguality. However, to take advantage of these potential benefits, some care has to be taken in designing and populating a database. In particular:

- Object attributes should be kept separate. In our data source, some attributes were grouped together. While, in some cases, we were able to take them apart (as in the case of dimensions), this was not the case in general. So, for example, the OBS field included a set of information which we were not always able to extract. By keeping distinct attributes (e.g., date of manufacture, place of manufacture, designer, and so on) in distinct fields, the task of information extraction would be simplified greatly, and a text produced from such database records would be more interesting.

- Items should be linked whenever possible. For example, part-of relationships should be explicitly stated, instead of being stated in free text. Similarly, given that there is often a thesaurus available, the thesaurus item should be included in the database record explicitly. Another example is to provide the appropriate link between the database record and the picture of the object, if one is available (or other multimedia information that relate to the object). While this may seem an obvious point, this link was not present in the data we were working with.

- Data should be kept consistent. This is of course important for any database, especially if it is to be processed by automatic means. Even simple inconsistencies greatly complicate the information extraction task: for example, we found a number of variations in the use of capitalisation, and grammatical incompatibility between field fillers.

To be able to exploit language technology and take advantage of the benefits it can offer, one must thus be careful from the outset, when a database is constructed for other purposes, to design it in the appropriate way. It is important to note that the features mentioned above do not necessarily impose more constraints on the end-users. Indeed, an interface to a database system can ensure that the interaction is not more difficult than it would have been, had the database been less structured with less consistent information. Finally, besides being able to support the exploitation of language technology, a more structured and consistent database can support a variety of other automatic processes (such as efficient search). It is thus not a real burden to add on the creation and population of a database, and yet it can create real benefits.

4.5 Linguistic resources required for generation

In the discussion above, we have focused on the issues related to automatically obtaining information from a database in order to form a knowledge base from which text can be automatically produced. However, the knowledge base is not the only source of information from which text is generated. A natural language system also needs a set of linguistic resources, in particular a grammar, which describes the syntax of the target language, a lexicon, which describes the vocabulary to be employed, and discourse information, which describes, for example, how a coherent text can be created to achieve a specific purpose in a specific domain.

In our system, we employ templates to represent the discourse and grammatical information, and a phrasal lexicon for the vocabulary. The templates are manually entered into the system. These are general, and can thus be re-used in many situations. They thus do not fall in the same category as the hand-crafting of a knowledge base.

The lexical information, on the other hand, is more problematic, especially when multilinguality is involved. In our system, we were able to obtain the English lexical information mostly from the database records themselves. Clearly, as the database was in English, this lexical information is only appropriate to produce English text. In order to produce text in other languages, we had to translate all the words into the other languages (e.g., *England* into *Angleterre* for French). While the data from which the text is produced remains the same and thus was obtained automatically, the lexicons had to be translated manually, at least for the purpose of this experiment.

5 Conclusions

We end by making some observations about the use of real data, and how the kinds of problems this presents might be surmounted.

We learned from this experiment that, while we were able to produce texts automatically from the database, these texts were of a mundane nature because of the scarcity and inconsistency of the underlying data as well as the lack of rich semantic content. To alleviate the problem of structure and consistency, we conclude that care must be taken from the outset to ensure that a database is appropriately designed and populated. The problems that arise from noisy data in our database are likely to be faced by any attempt to use a real database as an information source.

It is 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, [12]), so that database entry clerks do not have to acquire the skills of knowledge engineers in order to do their jobs. In the short-to-medium term, however, we are faced with the problem that the real data out there lives in more conventional forms, and that, as a result, the type of texts that we will be able to realistically generate from it is not as sophisticated and interesting as the texts whose production state-of-the-art generation techniques can support.

6. Acknowledgements

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, Maria Milosavljevic and Jon Oberlander [1997] The Web as Dialogue: The Role of Natural Language Generation in Hypertext. In *Proceedings of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*. Stanford University.

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

[4] Janet Hitzeman, Chris Mellish and Jon Oberlander [1997] Dynamic generation of museum web pages: The intelligent labelling explorer. *Archives and Museum Informatics*, **11**, 105–112.

[5] 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.

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

[7] Maria Milosavljevic, Adrian Tulloch and Robert Dale [1996] Text generation in a dynamic hypertext environment. In *Proceedings of the Nineteenth Australasian Computer Science Conference*, pp. 417–426. Melbourne, Australia.

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

[9] Maria Milosavljevic and Jon Oberlander [1998] Dynamic Hypertext Catalogues: Helping Users to Help Themselves. In *Proceedings of The 9th ACM Conference on Hypertext and Hypermedia* (HT'98). Pittsburgh, PA, USA. 20-24 June 1998.

[10] Cécile Paris, Keith Vander Linden, Markus Fischer, Anthony Hartley, Lyn Pemberton, Richard Power and Donia Scott. [1995] A Support Tool for Writing Multilingual Instructions. In Proceedings of *The International Joint Conference on Artificial Intelligence* (IJCAI'95). pp. 1398–1404.

[11] Cécile Paris and Keith Vander Linden [1996] DRAFTER: An Interactive Support Tool for Writing Multilingual Instructions. *IEEE Computer* **29(7)**:49–56. Special Issue on Interactive Natural Language Processing.

[12] Cécile Paris, Keith Vander Linden and Shijian Lu [1998] A Practical Approach to the Generation of On-Line Help. CSIRO Technical Note.

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