Structural variation in generated health reports

Catalina Hallett and **Donia Scott**
Centre for Research in Computing
The Open University
Walton Hall
Milton Keynes MK7 6AA
{c.hallett,d.scott}@open.ac.uk

Abstract

We present a natural language generator that produces a range of medical reports on the clinical histories of cancer patients, and discuss the problem of conceptual restatement in generating various textual views of the same conceptual content. We focus on two features of our system: the demand for "loose paraphrases" between the various reports on a given patient, with a high degree of semantic overlap but some necessary amount of distinctive content; and the requirement for paraphrasing at primarily the discourse level.

1 Introduction

Patient records are typically large collections of documents that reflect the medical history of a patient over a period of time. On average, the electronic patient record of a cancer patient contains information from over 150 documents, representing consult notes, referral letters, letters to and from the patient's GP, hospital admission and discharge notes, laboratory test results, surgery and other treatment descriptions, and drug dispensing notes. Although each document in this collection will have a specified purpose, there tends to be a high degree of redundancy between documents, but the sheer volume of information makes access extremely difficult.

The work presented in this paper is part of the Clinical E-Science Framework project (CLEF), which aims at providing tools to facilitate easy access to a patient's medical history. In particular, we describe a natural language generation system that produces a range of summarised reports of patient records from data-encoded views of patient histories which we call *chronicles*. Although we are concentrating on cancer patients, we aim to produce good quality reports without the need to construct extensive domain models.

Our typical user is a GP or clinician who uses electronic patient records at the point of care to familiarise themselves with a patient's medical history and current situation. A number of specific requirements arise from this particular setting:

- Reports that provide a quick potted overview of the patient's history are essential; this type of report should not be too long (ideally they should fit entirely on a computer screen) and should take less than a minute to read;
- At the same time, a complete view of the medical history must always be available on demand;
- Clinicians often need to examine a patient's history from a particular perspective (e.g., tests administered, treatments undertaken, drugs prescribed), and having focussed reports is also a requirement;
- Reports should be formatted to enhance readability;
- The selection of events for inclusion in a report should follow some basic rules:

- Events that deviate from what is considered to be normal are more important than normal events (for example, an examination of the lymphnodes that reveals lymphadenopathy is more important than an examination that doesn't).
- Some events are more important than others and should not only be included in the report but also highlighted (e.g., through colour coding, graphical timelines or similar display features).
- Less important events should be available on a need-to-know basis

These requirements impose important restrictions on the content of the reports and implicitly on the variety of lexical and syntactical devices we can employ:
(a) the veracity of the report is essential, therefore we are not at liberty to employ synonymy or lexical paraphrasing that may alter (however slightly) the meaning of the original input,
(b) we are required to maintain a certain syntactical ordering throughout a report in order to allow the user to quickly scan through the report with ease, and
(c) we have to produce several types of reports from the same input data.

In this paper, we focus on this last requirement, describing the methods we employ for reformulating content according to the type and focus of the generated report.

2 Types of report

In the current implementation, the generator produces two main types of report. The first is a longitudinal report, which is intended to provide a quick historical overview of the patient's illness, whilst preserving the main events (such as diagnoses, investigations and interventions). It presents the events in the patient's history ordered chronologically and grouped according to type. In this type of report, events are fully described (i.e., an event description includes all the attributes of the event) and aggregation is minimal (events with common attributes are aggregated, but there is no aggregation through generalization, for example). The following example displays a fragment of a generated longitudinal report[1]:

```
Example 1
The patient is diagnosed with grade
9 invasive medullary carcinoma of the
breast.  She was 39 years old when
the first cell became malignant.  The
history covers 1517 weeks, from week
180 to week 1697.  During this time, the
patient attended 38 consults.
YEAR 3:
Week 183
• Radical mastectomy on the breast was
performed to treat primary cancer of the
left breast.
• Histopathology revealed primary cancer
of the left breast.
Week 191
• Examination of the abdomen revealed
no enlargement of the liver or of the
spleen.
• Examination of the axillary lymphnodes
revealed no lymphadenopathy of the left
axillary lymphnodes.
• Examination of the breast revealed no
recurrent cancer of the left breast.
• Testing of the blood revealed
no abnormality of the haemoglobin
concentration or of the leucocyte count.
• Radiotherapy was initiated to treat
primary cancer of the left breast.
Week 192
• First radiotherapy cycle was
performed.
• ...
```

The second type of report focusses on a given type of event in a patient's history, such as the history of diagnoses, interventions, investigations or drug prescription. Under this category fall user-defined reports as well, where the user selects classes of interesting events (for example, *Investigations* of type *CT scan* and *Interventions* of type *surgery*).

A report of the diagnoses, for example, will focus on the *Problem* events that are recorded in the chronicle (e.g., cancer, anaemia, lymphadenopathy); other event types will only

[1] All the examples presented in this paper are extracted from summaries produced by our Report generator.

appear if they are directly related to a *Problem*. As it can be seen in Example 2, this type of report is necessarily more condensed, since the events do not have to appear chronologically and can be grouped in larger clusters. Secondary events are also more highly aggregated.

`Example 2`

• `In week 483, primary cancer of the right breast was revealed by the histopathology report. The cancer was treated with radical mastectomy on the breast.`

• `In week 491, no abnormality of the leucocyte count or of the haemoglobin concentration, no lymphadenopathy of the right axillary lymphnodes, no enlargement of the spleen or of the liver and no recurrent cancer of the right breast were found. Radiotherapy was initiated to treat primary cancer of the right breast.`

• `In the weeks 492 to 496, 5 radiotherapy cycles were performed.`

If the focus is on *Interventions*, the same information in the previous example will be presented as:

`Example 3`

• `In week 483, histopathology revealed primary cancer of the right breast. Radical mastectomy on the breast was performed to treat the cancer. Radiotherapy was initiated to treat primary cancer of the right breast.`

• `In the weeks 492 to 496, 5 radiotherapy cycles were performed.`

In an *Investigation*-focussed report, the intervention will be omitted, since they are not directly relevant:

`Example 4`

• `In week 483, histopathology revealed primary cancer of the right breast`

• `In week 491, examination revealed no abnormality of the leucocyte count or of the haemoglobin concentration, no lymphadenopathy of the right axillary lymphnodes, no enlargement of the spleen or of the liver and no recurrent cancer of the right breast.`

It is important to note that although the reports are generated from the same input content, they are not exact reformulations of each other, but rather different views of the same content with a large degree of overlap. This feature is a direct result of the report requirements.

3 Input

As mentioned earlier, the input to our Report Generator is a data-encoded *chronicle* of the patient's medical history. Technically, the chronicle is the partial result of information extraction applied on clinical narratives, combined with structured data (such as radiology results or demographic data), and supplemented with inferences. However, in developing our report generator, we are currently using a Chronicle Simulator, which constructs invented chronicles, allowing us to ignore for the time being some problems that can appear when using an information extraction system (being developed in parallel). Firstly, the resulting data is complete and correct, thus allowing us to concentrate on the design and testing of the generation and summarisation system without having to take into account at this point errors in the Information Extraction. Secondly, our data on cancer patients is highly confidential, which makes presentation of the output of the report generator (e.g., for evaluation with real subjects, or dissemination purposes) very difficult. Using a simulator also means that we can have instant access to a large number of randomly generated chronicles, which at this stage of the project are not yet available.

The Chronicle Simulator simulates the history of a patient's illness, and links the events in the history in a manner that closely resembles the expected output of the real Automatic Chronicler. The current output format of the simulator is a relational database that stores six types of event[2] (interventions, investigations, consults, drugs, problems and loci) and 14 types of relation between events (e.g., *Problem*

[2] The term event is loosely used to denote dynamic (such as interventions) as well as static concepts (such as problems).

HAS-LOCUS *Locus*, *Intervention* CAUSED-BY *Problem*, *Intervention* SUBPART-OF *Intervention*, *Investigation* HAS-INDICATION *Problem*). Each event has a variable number of attributes, and each dynamic event is time-stamped with a start date and an end date[3]. A typical chronicle contains around 350 events and about 600 relations.

4 Architecture

The design of the Report Generator follows a classical pipeline architecture, with a content selector, content planner and syntactic realiser. The Content Planner is tightly coupled to the Content Selector, since part of the discourse structure is already determined in the event selection phase. Aggregation is mostly conceptual rather than syntactic, and thus it is performed in the content planning stage as well as during realisation (Reape and Mellish, 1999).

4.1 Content selection

The Content selection process represents the most important component of the Report Generator. Although in some contexts it may be useful to generate reports containing all the events in a chronicle, the most useful types of report are the focused, summarised ones, for which good selection of important events is essential.

The process of content selection is currently driven by two parameters of a report: *type* and *length*. We define the concept of *report spine* to represent a list of concepts that are essential to the construction of a given type of report. For example, in a report of the diagnoses, all events of type *Problem* will be part of the spine. Events linked to the spine through some kind of relation may or may not be included in the summary, depending on the type and length of the summary (see Figure 1). The design of the system does not restrict the spine to containing only events of the same type. In future extensions to the system where the user will be able to select facts they want in the summary, a spine could contain, for example, problems of type *cancer*, investigations of type *x-ray* and interventions of type *surgery*.

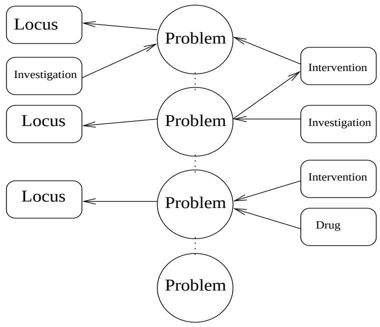

Figure 1: Example of a generated spine structure

Spines are not predefined templates, but structures that are constructed dynamically with each request and they depend on the type of request and on the length of the summary.

Important events are selected according to semantic relations. The first step in the selection process is to cluster related events based on the relations stored in the chronicle. A cluster of events may tell us, for example, that a patient was diagnosed with cancer following a clinical examination, for which she had a mastectomy to remove the tumour, was given a histopathological test of the removed tumour, which confirmed the cancer, and had a complete radiotherapy course to treat the cancer; the radiotherapy caused an ulcer, which in turn was treated with some drug. A typical chronicle contains a small number of clusters, typically one or two large clusters and several small ones. Smaller clusters are generally not related to the main thread of events.

The summarisation process starts with the removal of small clusters, which in the current implementation are defined as clusters containing at most three events[4]. This excludes some specified types of information that will be included in the report even when they only appear in short clusters; for example, all reports will contain essential information such as the initial diagnosis and the cause of death (if available).

The next step is the selection of important events, as defined by the type of report. Each cluster of events is a graph, with some nodes representing spine events. For each cluster, the spine events are selected, as well as all nodes that are at a distance of less than *n* from spine events,

[3] In the current implementation of the chronicle, time stamps are week numbers starting with the date of the first diagnosis.

[4] This threshold was set following a series of experiments.

where the depth *n* is a user-defined parameter used to adjust the size of the report. For example, in the cluster presented in Fig. 2, assuming a depth value of 1, the content selector will choose *cancer*, *left breast* and *radiotherapy* but not *radiotherapy cycle* or *ulcer*.

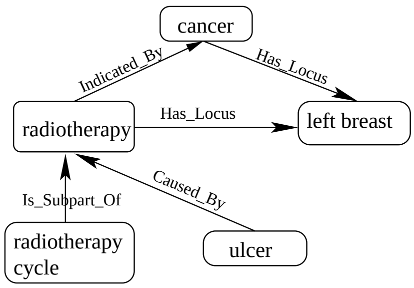

Figure 2: Example of a cluster

A document plan is typically a hierarchical structure that contains and combines the messages to be conveyed by the report generator. Technically, a document plan is an ordered collection of message clusters, where messages within a cluster are combined using rhetorical relations, while individual clusters are ordered and linked according to the type of report.

The construction of document plans is partly performed in the content selection phase, since the content is selected according to the relations between events, which in turn provide information about the structure of the target text. The actual document planner component is concerned with the construction of complete document plans, according to the type of report and cohesive relations identified in the previous stage. A report typically consists of three parts:
(a) a schematic description of the patient's demographic information (name, age, gender),
(b) a two sentence summary of the patient's record (presenting the time span of the illness, the number of consults the patient attended, and the number of investigations and interventions performed) and
(c) the actual report of the record produced from the events selected to be part of the content.
We focus here on this last part.

4.2 Document planning

The first stage in structuring the body of the report is to combine messages linked through attributive relations (e.g., combining messages of type *Problem* with messages of type *Locus* if the *Problem* has a HAS-LOCUS relation pointing to a *Locus*). In the second stage, messages are grouped according to specific rules, depending on the type of report. For longitudinal reports, the rules stipulate that events occurring in the same week should be grouped together, and further grouped into years. In event-specific reports, patterns of similar events are first identified and then grouped according to the week(s) they occur in. For example, if in week 1 the patient was examined for enlargement of the liver and of the spleen with negative results and in week 2 the patient was again examined with the same results and had a mastectomy, two groups of events will be constructed:

```
Example 5
• In weeks 1 and 2, examination of the
abdomen revealed no enlargement of the
liver or of the spleen.
• In week 2, the patient underwent a
mastectomy.
```

Within groups, messages are structured according to discourse relations that are either deduced from the input database or automatically inferred by applying domain specific rules. At the moment, the input provides three types of rhetorical relation: *Cause*, *Result* and *Sequence*. The domain specific rules specify the ordering of messages, and always introduce a *Sequence* relation. An example of such a rule is that a histopathology event has to follow a biopsy event, if both of them are present and they start and end at the same time. These rules facilitate the construction of a partial rhetorical structure tree. Messages that are not connected in the tree are by default assumed to be in a List relation to other messages in the group, and their position is set arbitrarily.

The document planner also applies aggregation rules between similar messages and employs ellipsis and conjunction in order to create a more fluent text. Simple aggregation rules state, for example, that two investigations with

```
Investigation
    [id: 01,
     name: examination,
     HAS−TARGET: abdomen]          Investigation
Investigation                         [id:03,
    [id:02,                 >          name:examination,
     name:examination,                 HAS−TARGET: {abdomen, breast}]
     HAS−TARGET: breast]
```

Figure 3: Aggregation of *Investigation* messages on the HAS-TARGET field

the same name and two different target loci can be collapsed into one investigation with two target loci (Fig.3). Aggregation rules of this type are designed to make the resulting text more fluent, however they do not always provide the degree of condensation required by the summary. For example, each clinical examination consists of examinations of the abdomen for enlargement of internal organs (liver and spleen) and examination of the lymphnodes. Thus, each clinical examination will typically consist of three independent *Investigation* events. When fully aggregated according to conceptual and syntactical rules, the three *Investigation* messages are collapsed into one structure such as:

```
Example 6
Examination revealed no enlargement
of the spleen or of the liver and no
lymphadenopathy of the axillary nodes.
```

However, this level of aggregation that only takes into account the semantics of individual messages may be not enough, since clinical examinations are performed repeatedly and consist of the same types of investigation. Two approaches have been implemented in the Report Generator, both of which make use of domain specific rules. The first is to report only events that deviate from the norm. In the case of investigations, for example, this equates to reporting only those that have abnormal results. The second, which produces larger reports, is to produce synthesised descriptions of events. In the case of clinical examination for example, we could describe a sequence of investigations such as the one in example (5) as *Clinical examination was normal*. If the examination deviates from the norm on a restricted numbers of parameters only, this can be described as *Clinical examination was normal, apart from an enlargement of the spleen*.

4.3 Maintaining the thread of discourse

In producing multiple reports on the same patient from different perspectives, or of different types, we operate under the strong assumption that event-focussed reports should be organised in a way that emphasises the importance of the event in focus. From a document structure viewpoint, this equates to constructing rhetorical structures where the focus event (i.e., the spine event) is expressed in a nuclear unit, and skeleton events are preferably in sattelite units.

Within sentences, spine events are assigned salient syntactical roles that allows them to be kept in focus. For example, a relation such as

Problem CAUSED-BY *Intervention*

is more likely to be expressed as :

The patient developed a *Problem* as a result of an *Intervention*.

when the focus is on *Problem* events, and, when the focus is on *Interventions* as:

An *Intervention* caused a *Problem*.

This kind of variation reflects the different emphasis that is placed on spine events, although the wording in the actual report may be different. Rhetorical relations holding between simple event descriptions are most often realised as a single sentence (as in the examples above). Complex individual events are realised in individual clauses or sentences which are connected to other accompanying events through the appropriate rhetorical relation.

For example, a *Problem* event has a large number of attributes, consisting of *name, status, existence, number of nodes counted, number of nodes involved, clinical course, tumour size, genotype, grade, tumour marker* and *histology*, as well as the usual time stamp. The selection of attributes that are going to be included in a *Problem* description depends on a number of factors, including whether the *Problem* is a spine or a skeleton event, and whether the event is mentioned for the first time or is a subsequent mention. Aditionally, the number of attributes included in the description of a *Problem* is a decisive factor in realising the *Problem* as a phrase, a sentence or a group of sentences. In the following two examples, there are two *Problem* events (*cancer* and *lymphnode count*) linked through an *Investigation* event (*excision biopsy*, which is indicated by the first problem and has as a finding the second problem. In Example 7, the problems are first mentioned spine events, while in Example 8, the problems are skeleton events (the cancer is a subsequent mention and the lymphnode count is a first mention), with the *Investigation* being the spine event.

```
Example 7
A 10mm, EGFR +ve, HER-2/neu +ve,
oestrogen receptor positive cancer was
found in the left breast (histology:
invasive tubular adenocarcinoma).
Consequently, an excision biopsy was
performed which revealed no metastatic
involvement in the 5 nodes sampled.
```

```
Example 8
An excision biopsy on the left breast
was performed because of cancer.  It
revealed no metastatic involvement in
the 5 nodes sampled.
```

As can be seen from the examples above, the same basic rhetorical structure consisting of three nodes and two relations (causality and consequence) is realised differently in a *Problem*-focussed report compared to an *Investigation*-based report. The conceptual reformulation is guided by the type of report, which in turn has consequences at syntactical level.

5 Evaluation

Automatic evaluation of the generated reports is not possible, as there is no gold standard for such documents. Additionally, a full-blown quantitative evaluation is not yet feasible, since our users are cancer specialists who cannot easily dedicate time to evaluating large numbers of reports. However, we have conducted an informal survey with two cancer clinicians to gain feedback on the quality of the current output of the Report Generator. To do this, we showed them three patient records encoded as chronicles, and, for each patient, two types of report produced from that record: a longitudinal report, and a summarised report of diagnoses. The three patient records were selected to display a variety of events and sizes (a 6-year history containing 621 events, a 12-year history with 1418 events, and a 9-year history with 717 events).

Although they were (unusually) familiar with the coding scheme of the chronicles, the clinicians found it very difficult to extract a useful overview of the patients' histories from the three chronicles we showed them. In contrast, they found the generated reports to be much more useful and the quality of the text to be very good. The clinicians commended the reports for their ability to provide a quick and clear view of data that would be otherwise difficult to access and process. Most importantly, the various report types were judged to be highly appropriate for use in clinical care.

Whilst this preliminary evaluation was conducted with the aim of finding early shortcomings of the Report Generator and receiving feedback from potential users, we are now embarking on a more extensive formal evaluation with cancer clinicians and medical researchers with specialist knowledge in the area of cancer. We believe, however, that the true test of utility will be the actual use of the Report Generator in practice.

6 Conclusions

We have described a system that generates a range of health reports on individual cancer patients. At present, our intended readership is composed of clinicians and medical researchers, and the type of report will depend on his or her stated needs. Reports that are required at the point of care (e.g., for a doctor interviewing a newly referred patient, or a team of medics on ward rounds) are likely to be short "30-second" potted histories. At other times longer, more detailed reports will be required, as will reports that focus on particular aspects of the patient's "journey" through their disease (e.g., from the perspective of the diagnoses that have been made, the drugs they have been prescribed, or surgery they have undergone). The system is fully implemented in Java and currently generates this full range of reports on-the-fly. A summarised report based on about 1000 input events is constructed in less than 2 seconds, a speed which is highly appropriate to the demands of clinical practice.

While the various types of generated report all share the same input (i.e., the patient's chronicle), and thus will have a large degree of conceptual overlap, clearly there will be occassions when information that is included in some reports will not be in others.

The range of reports for any given patient at any given point in their illness thus present a special class of paraphrase, with a looser adherance to semantic equivalence between versions than is typically found in other paraphrase generators, for example Kozlowski et al (2003), McKeown et al (1994), Power, Scott and Bouyaad-Agha (2003), Rosner and Stede (1994),(1996), and Scott and Souza (1990). In this sense, our Report Generator is rather closer in spirit to Hovy's PAULINE system, which generates descriptions of given news events from different perspectives and with different stylistic goals (Hovy, 1988). However, we achieve our goal with less reliance on terminological variation and more on structural variation at the discourse level. Syntactic variation, where it does occur, is almost always simply a side-effect of an earlier discourse choice. Terminological variation is deliberately avoided to prevent false implicatures; however, we are about to introduce a further class of readership, namely patients, at which stage we will make fuller use of our lexical resources.

7 Acknowledgments

CLEF is supported in part by grant G0100852 under the E-Science Initiative. Thanks are due its clinical collaborators at the Royal Marsden and Royal Free hospitals, to colleagues at the National Cancer Research Institute (NCRI) and NTRAC and to its industrial collaborators. Special thanks to Dr. Jeremy Rogers who provided us with the automated Chronicle Simulator that we have used in all our experiments.

References

Eduard H. Hovy. 1988. *Generating natural language under pragmatic constraints*. Lawrence Erlbaum, Hillsdale, New Jersey.

Raymond Kozlowski, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In Kentaro Inui and Ulf Hermjakob, editors, *Proceedings of the Second International Workshop on Paraphrasing*, pages 1–8.

Kathleen McKeown, Karen Kukich, and James Shaw. 1994. Practical issues in automatic document generation. In *Proceedings of the Fourth Conference on Applied Natural-Language Processing (ANLP-1994)*, pages 7–14, Stuttgart, Germany.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Mike Reape and Chris Mellish. 1999. Just what is aggregation anyway? In *Proceedings of the 7th European Workshop on Natural Language Generation (EWNLG'99)*, pages 20–29, Toulouse, France.

Dietmar Rösner and Manfred Stede. 1994. Generating multilingual documents from a knowledge base: the TECHDOC project. In *Proceedings of the 15th conference on Computational Linguistics (Coling'94)*, pages 339–343, Kyoto, Japan.

Donia Scott and Clarisse de Souza. 1990. Getting the message across in RST-based text generation. In R. Dale C. Mellish and M. Zock, editors, *Current Research in Natural Language Generation*, pages 31 – 56. Academic Press.

Manfred Stede. 1996. Lexical paraphrases in multilingual sentence generation. *Machine Translation*, 11:75–107.