A Two-stage Model for Content Determination

Somayajulu G. Sripada
Dept. of Comp. Sc.
Univ. of Aberdeen,
Aberdeen, UK
ssripada@csd.abdn.ac.uk

Ehud Reiter
Dept. of Comp. Sc.
Univ. of Aberdeen,
Aberdeen, UK
ereiter@csd.abdn.ac.uk

Jim Hunter
Dept. of Comp. Sc.
Univ. of Aberdeen,
Aberdeen, UK
jhunter@csd.abdn.ac.uk

Jin Yu
Dept. of Comp. Sc.
Univ. of Aberdeen,
Aberdeen, UK
jyu@csd.abdn.ac.uk

Abstract

In this paper we describe a two-stage model for content determination in systems that summarise time series data. The first stage involves building a qualitative overview of the data set, and the second involves using this overview, together with the actual data, to produce summaries of the time-series data. This model is based on our observations of how human experts summarise time-series data.

1 Introduction

This paper addresses the problem of content determination in data summarisation. Content determination as the name indicates is the process responsible for determining the content of the texts generated by an NLG system (Reiter and Dale 2000). Although content-determination is probably the most important part of an NLG system from the end-user's perspective, there is little agreement in the NLG community as to how content-determination should be done, with different systems adapting widely varying approaches. Also, algorithms and architectures for content-determination seem to often be based on the intuitions of system developers, instead of on empirical observations, although detailed content determination rules are often based on corpus analysis and interaction with experts.

In this paper we propose a general architecture for content determination in data summarisation systems which assumes that content determination happens in two stages: first a qualitative overview of the data is formed, and second the content of the actual summaries is decided upon. This model is based on extensive knowledge acquisition (KA) activies that we have carried out in the SUMTIME project (Sripada, 2001), and also matches observations made during KA activities carried out in the STOP project (Reiter *et al* 2000). We have not yet implemented this model, and indeed one of the issues that we need to think about is to what degree a content-determination strategy used by human experts is also an appropriate one for a computer NLG system.

2 Content Determination

Content determination is the task of deciding on the information content of a generated text. In the three-stage pipeline model of Reiter and Dale (2000), content determination is part of the first stage, document planning, along with document structuring (determining the textual and rhetorical structure of a text). Content determination is extremely important to end users; in most applications users probably prefer a text which poorly expresses appropriate content to a text which nicely expresses inappropriate content. From a theoretical perspective content determination should probably be based on deep reasoning about the system's communicative goal, the user's intentions, and the current context (Allen and Perrault 1980), but this requires an enormous amount of knowledge and reasoning, and is difficult to do robustly in real applications.

In recent years many new content determination strategies have been proposed, ranging from the use of sophisticated signal-processing techniques (Boyd 1997) to complex planning algorithms (Mittal *et al* 1998) to systems which exploit cognitive models of the user (Fiedler 1998). However, most of these strategies have only been demonstrated in one application. Furthermore, as far as we can tell these strategies are usually based on the intuition and experiences of the developers. While realisation, microplanning, and document structuring techniques are increasingly based on analyses of how humans perform these tasks (including corpus analysis, psycholinguistic studies, and KA activities), most papers on content determination make little reference to how human experts determine the content of a text. Human experts are often consulted with regard to the details of content rules, especially when schemas are used for content determination (Goldberg *et al* 1994, McKeown *et al* 1994, Reiter *et al* 2000); but they rarely seem to be consulted (as far as we can tell) when deciding on the general algorithm or strategy to use for content determination.

3 Summarising Time-Series Data

3.1 Text summaries of Time-Series Data

Time-series data is a collection of values of a set of parameters over time. Such data is very common in the modern world, with its proliferation of databases and sensors, and humans frequently need to examine and make inferences from time-series data.

Currently, human examination of time-series data is generally done either by direct inspection of the data (for small data sets), by graphical visualisation, or by statistical analyses. However, in some cases textual summaries of time-series data are also useful. For example, newspapers regularly publish textual summaries of weather predictions, the results of polls and surveys, and stock market activity, instead of just showing numbers and graphs. This may be because graphical depictions of time-series data require time and skill to interpret, which is not always available. A doctor rushing to the side of a patient who is suffering from a heart attack, for example, may not have time to examine a set of graphs of time-series data, and a newspaper reader may not have the statistical knowledge necessary to interpret raw poll results.

Perhaps the major problem today with textual descriptions of time-series data is that they must be produced manually, which makes them expensive and also means they can not be produced instantly. Graphical depictions of data, in contrast, can be produced quickly and cheaply using off-the-shelf computer software; this may be one reason why they are so popular. If textual summaries of time-series data could be automatically produced by software as cheaply and as quickly as graphical depictions, then they might be more widely used.

3.2 SUMTIME

The goal of the SUMTIME project is to develop better techniques for automatically generating textual summaries of time-series data, in part by integrating leading-edge NLG and time-series analysis technology. We are currently focusing on two domains:

Meteorology – producing weather forecasts from numerical weather simulations. This work is done in collaboration with Weather News Inc (WNI)/Oceanroutes, a leading meteorological company.
Gas Turbines – summarising sensor readings from a gas turbine. This work is done in collaboration with Intelligent Applications, a leading developer of monitoring software for gas turbines.

These domains are quite different in time-series terms, not least in the size of the data set. A typical weather forecast is based on tens of values for tens of parameters, while a summary of gas-turbine sensor readings may be based on tens of thousands of values for hundreds of parameters. We hope that looking at such different domains will help ensure that our results are generalisable and not domain-specific. We will start working on a third domain in 2002; this is likely to be a medical one, perhaps (although this is not definite) summarising sensor readings in neonatal intensive care units.

The first year of SUMTIME (which started in April 2000) has mostly been devoted to knowledge acquisition, that is to trying to understand how human experts summarise time-series data. This was done using various techniques, including corpus analysis, observation of experts writing texts, analysis of content rules suggested by experts, discussion with experts, and think-aloud sessions, where experts 'think aloud' while writing texts (Sripada, 2001).

3.3 Example

The following table shows an example segment of meteorological time series data, specifically predicted wind speed and wind direction at an offshore oil exploration site. The time field is shown in 'day/hour' format.

Time (day/hour)	**Wind Direction**	**Wind Speed** Knots
05/06	SE	22
05/09	SE	24
05/12	SE	30
05/15	SE	25
05/18	SSE	28
05/21	SSE	22
06/00	SE	16

This data was summarised by WNI's human forecasters as follows:

```
FORECAST 06-24 GMT,FRIDAY,05-Jan
2001
WIND(KTS)      CONF   HIGH
  10M: SE 20-25 OCCASIONALLY
       25-30, EASING 15-20 LATER
```

The above example is just a sample showing the data and its corresponding forecast text for the wind subsystem. Real weather forecast reports are much longer and are produced from data involving many more weather parameters than just wind speed and wind direction.

4 Human Summarisation

4.1 Meteorology

In the domain of weather forecasting, we observed how human experts carry out the task of summarising weather data by video recording a meteorologist thinking aloud while writing weather forecasts. Details of the KA have been described in Sripada (2001). Our observations included the following:

1. In the case of weather forecasts, time-series data represent the values of important weather parameters (wind speed, direction, temperature, rainfall), which collectively describe a single system, the weather. It seemed as though the expert was constructing a mental picture of their source using the significant patterns in time series. Thus the first activity is that of data interpretation to obtain a mental model of weather.

2. The mental model of the weather is mostly in terms of the elements/objects related to atmosphere, like cold fronts and warm fronts; it also seems to be qualitative instead of numerical. In other words, it qualitatively describes the meteorological state of the atmosphere. The expert calls this an 'overview of the weather'.

3. Building the overview involves the task of interpretation of the time series weather data. While interpreting this data the expert used his meteorological knowledge (which includes his personal experience in interpreting weather data) to arrive at an overview of the weather. During this phase, he appeared to be unconcerned about the end user of the overview (see 4.1.1 below). We call this process Domain Problem Solving (DPS) where information is processed using exclusively the domain knowledge.

4. Forecasts are written after the forecaster gets a clear mental picture (overview) of the weather. Building the overview from the data is an objective process which does not

depend on the forecast client (user), whereas writing the forecast is subjective and varies with client.

4.1.1 Examples

Two examples of the influence of the overview on wind texts (Section 3.3) are:

1. When very cold air flows over a warm sea, surface winds may be underestimated by the numerical weather model. In such cases the forecaster uses his 'overview of the weather' to increase wind speeds and also perhaps add other instability features to the forecast such as squalls.

2. If the data contains an outlier, such as a wind direction which is always N except for one time period in which it is NE, then the expert uses the overview to decide if the outlier is meteorologically plausible and hence should be reported or if it is likely to be an artefact of the simulation and hence should not be reported.

The above examples involve reasoning about the weather system. Forecasters also consider user goals and tasks, but this may be less affected by the overview. For example, in one think-aloud session, the forecaster decided to use the phrase *20-24* to describe wind speed when the data file predicted a wind speed of 19kt. He explained to us that he did this because he knew that oil-rig staff used different operational procedures (for example for docking supply boats) when the wind exceeded 20kt, and he also knew that even if the average wind speed in the period was 19kt, the actual speed was going to vary minute by minute and often be above 20kt. Hence he decided to send a clear signal to the rig staff that they should expect to use '20kt or higher' procedures, by predicting a wind speed of *20-24*. This reasoning about the user took place after the overview had been created, and did not seem to involve the overview.

4.2 Gas Turbine Sensors

Unlike the previous domain, in the domain of gas turbine (GT), currently there are no textual summaries of turbine data written by humans. Thus we have asked the domain experts to comment orally on the data. However, the experts have attempted to summarise their comments at the end of each session if they found something worth summarising. Our observations included:

1. The main task is to identify the abnormal data and summarise it. However, an abnormal trend in a specific channel might have been caused due to a change in another channel (for instance, an increase in the output voltage can be explained with a corresponding increase in the fuel input). Thus individual channel data needs to be interpreted in the context of the other channels.

2. The expert agrees during the KA session that he first analyses the data numerically to obtain qualitative trends relating to the GT before generating comments. Therefore the state of the GT that produced the data is constructed through data interpretation and the knowledge of the state is then used to check if the turbine is in a healthy state or not. Since GT is an artefact created by humans it is possible to have a fairly accurate model of states of a GT (unlike weather!).

3. The phrases used by the expert often express the trends in the data as if they were physical happenings on the turbine, like "running down" for a decreasing trend in shaft speed data. This indicates that the expert is merely expressing the state of the GT. This in turn indicates that at the time the summarisation is done, the mental model of the state of the GT is available.

5 Evidence from Other Projects

After making the above observations, we examined think-aloud transcripts from an earlier project at the University of Aberdeen, STOP (Reiter *et al* 2000), which involved building an NLG system that produced smoking-cessation letters from smoking questionnaires. These transcripts (from think-aloud sessions of doctors and other health professionals manually writing smoking-cessation letters) showed that in this domain as well experts would usually first build

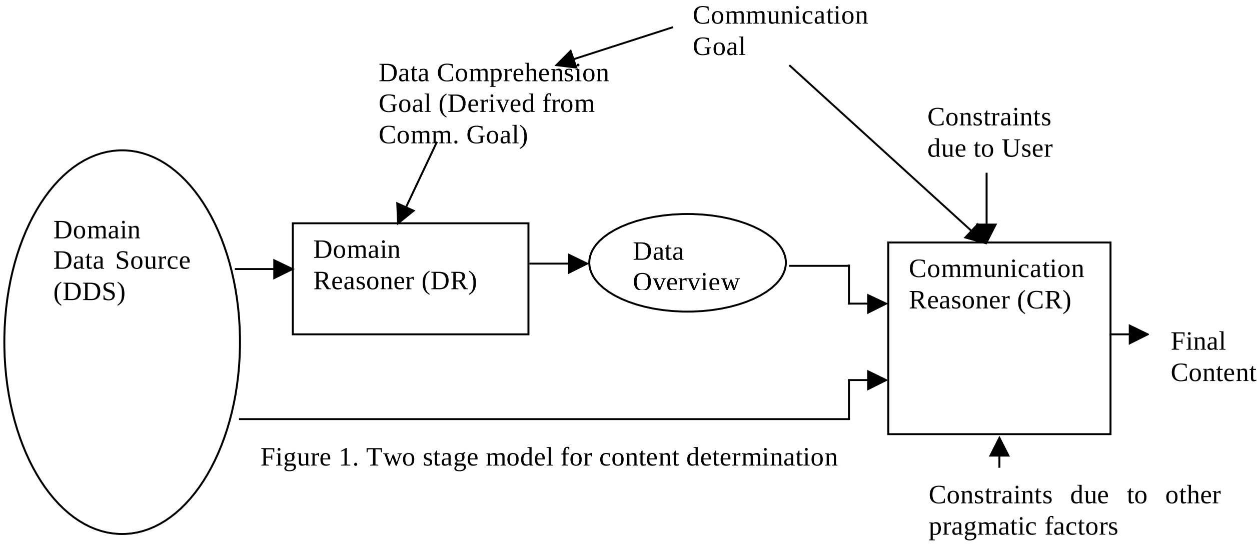
Communication Goal
Data Comprehension Goal (Derived from Comm. Goal)
Constraints due to User
Domain Data Source (DDS)
Domain Reasoner (DR)
Data Overview
Communication Reasoner (CR)
Final Content
Constraints due to other pragmatic factors

Figure 1. Two stage model for content determination

an overview (in this case, of the smoker) before starting to determine the detailed content of a letter. Below is an excerpt from one of the transcripts of a KA session :

« …. The first thing I have got to do is to read through the questionaire just to **get some idea** of where he is at with his smoking. …… »

We did not investigate overview formation in any detail in STOP, but the issue did come up once in a general discussion with a doctor about the think-aloud process. This particular doctor said that he built in his mind a mental image of the smoker (including a guess at what he or she looked like), and that he found this image very useful in deciding how best to communicate with the smoker.

In another work, RajuGuide, once again there is evidence of an overview influencing content determination (Sripada 1997). RajuGuide is a system that generates route descriptions. At a higer level of abstraction, RajuGuide has two parts. The first part is responsible for planning the route the user wanted. The second module is responsible for generating the text describing the route. The route computed by the first part, which is in the form of a series of coordinates, is not directly communicated to the user. Instead the second part attempts to enrich the route depending upon what the user already knows and what additional information the knowledge base knows for that particular route. We believe that the route computed by the route planner is the overview in this case and it drives the content determination process in the second part.

6 Two-stage Model for content determination

These observations have led us to make the following hypotheses:

1. Humans form a qualitative overview of the input data set.
2. Not all the information in the overview is used in the text.
3. The overview is not dependent on pragmatic factors such as the user's taste, these are considered at a later stage of the content determination process.

Based on the above hypotheses, we propose a two-stage model for content determination as depicted in Figure 1. It is assumed that Domain Data Source (DDS) is external to the text generator. It has been assumed that a Domain Problem Solver or Domain Reasoner (DR) is available for data processing. This reasoning module is essentially useful to draw inferences while interpreting the input data set and ultimately is responsible for generating the overview. Communication Goal (CG) is the input to the data summarisation system in response to which it accesses DDS to produce an overview of the data using the DR. In the context of the overview produced by DR, the

Communication Reasoner (CR) system generates the final content specification taking into account the influence of the User Constraints (UC) and other pragmatic factors. This content is then sent to subsequent NLG modules (not shown), such as microplanning and surface realisation.

Our model has some similarities to the one proposed by Barzilay *et al* (1998), in that the Domain Reasoner uses general domain knowledge similar to their RDK, while the Communication Reasoner uses communication knowledge similar to their CDK and DCK.

The central feature of the above model is the idea of data overview and its effect on content selection. One possible use of overviews is to trigger context-dependent content rules. The time-series analysis part of SUMTIME is largely based on Shahar's model (1997), which makes heavy use of such rules. In Shahar's model contexts are inferred by separate mechanisms; we believe that these should be incorporated into the overview, but this needs further investigation.

At the current stage of our project we have only a gross idea of what makes up the proposed data overview. Our suspicion is that it is hard to make a generic definition of the data overview for all domains. Instead, we would like to imagine the data overview as the result of inferences made from the input data so as to help in triggering the right content determination rules. For example, in out meteorology domain, the input time-series data comes from a numerical weather prediction (NWP) model, but even the most sophisticated NWP models do not fully represent the real atmosphere – all models work with approximations. Thus the NWP data displayed to the meteorologist is interpreted by him to arrive at a conceptual model in his or her head, which is the overview.

7 Issues with the two stage model

There are a number of issues that need to be resolved with respect to the two-stage model described above.

7.1 Is overview creation a human artefact?

The main basis for including the overview in two stage model has been the observation made during the think aloud sessions that experts form overviews before writing texts. Now it can be argued that even if humans need an overview, computer programs may not. Evidently, it is hard to ever prove the contrary. But what can be done is to show the advantages gained by a computer program by using an overview for content selection.

7.2 Does the overview have any other utility than just providing context for content determination rules?

We believe that the overview can play multiple roles in the overall process of writing textual forecasts. First, the overview can bring in additional information into the text that is not directly present in the underlying raw data. In Reiter and Dale's (2000) terminology, overviews are a technique for generating Computable Data content, that is content which is not directly present in the input data but can be computed or inferred from it. Such content provides much of the value of summary texts. Indeed, one could argue that simple textual descriptions of a set of data values without extra computed or inferred content, such as those produced by TREND (Boyd, 1997), might not be that much more useful than a graph of the data.

The overview may also help in deciding how reliable the input data is, which is especially important in the meteorology domain, since the data comes from an NWP simulation. This could, for example, help the generation system decide whether to use precise temporal terms such as *Midnight* or vague temporal terms such as *tonight*. Again one could argue that the ability to convey such uncertainty and reliability information to a non-specialist is a key advantage of textual summaries over graphs.

In general, the overview allows reasoning to be carried out on the raw data and this will probably be useful in many ways.

7.3 How is the overview related to the domain ontology?

The basic concepts present in an overview may be quite different from the basic concepts present in a written text. For example, the overview built by our expert meteorologist was based on concepts such as lapse rate (the rate at which temperature varies with height), movement of air masses, and atmospheric stability. However, the texts he wrote mentioned none of these, instead it talked about wind speed, wind direction, and showers. In the STOP domain, overviews created by doctors seemed to often contain many qualitative psychological attributes (depression, self-confidence, motivation to quit, etc) which were not explicitly mentioned in the actual texts written by the doctors.

This suggests that the conceptual ontology, that is the specification of underlying concepts, underlying the overview may be quite different from the ontology underlying the actual texts. The overview ontology includes concepts used by experts when reasoning about a domain (such as air masses or motivation), while the text ontology includes concepts useful for communicating information to the end user (such as wind speed, or longer life expectancy).

7.4 What do experts think about the two-stage model?

When the two stage model was reported back to a WNI expert who participated in a think-aloud session, the expert agreed that he does build an overview (as he did during the KA session) while writing forecasts, but felt that it's use may not be necessary for writing all forecasts. In his opinion, the interpretation of most data sets doesn't require the use of the overview. However, he was quick to add that the quality of the forecasts can be improved by using overviews which faciliate reasoning with the weather data.

8 Evaluation

We are currently building a testbed system called SUMTIME-MOUSAM which will enable us to test the hypotheses we have presented in this paper and other hypotheses suggested by our KA activities. SUMTIME-MOUSAM is a framework system that consists of

- "Infrastructure" software for accessing data files, regression testing of new software versions, etc.
- An ontology, which defines a conceptual level of representation of texts.
- A corpus of human-written texts with their corresponding conceptual representations defined using the above ontology.
- Scoring software which compares the output of a module (either at a conceptual or text level) against the human corpus.

Because we are primarily interested in content issues, it is important to evaluate our system at a content level as well as at a text level. To support this, we are developing conceptual representations of the texts we will be generating, which can also be extracted from human texts by manual analysis.

SUMTIME-MOUSAM is currently best developed in the area of producing wind texts. In this area, we have developed a conceptual representation and manual annotation guide (with good inter-annotator agreement, generally kappa values of .9 or higher); built an initial software system to automatically produce such texts based on a threshold model without an overview; and begun the process of analysing differences. We are currently working on extending SUMTIME-MOUSAM to other parts of weather forecasts, such as statements describing clouds and precipitation, and plan in the future to extend it to the gas-turbine domain.

With regard to testing hypotheses specifically about two-stage content determination (the subject of this paper), our plan is as follows

1. Compare the output of the non-overview based software to human summary texts, and identify cases where an overview seems to be used.

2. Ask human experts to build an overview (using a GUI), modify our software to use this overview when generating texts, and see if this results in texts more similar to the human texts.

3. Attempt to automatically generate the overview from the data, and again compare the resultant texts to human texts.

At some point towards the end of SUMTIME, we also hope to conduct user task evaluations. For example, we may show gas-turbine engineers our summary texts and see if this helps them detect problems in the gas turbine.

9 Conclusion

Our experience in three domains shows that human experts build qualitative overviews when writing texts, and that these overviews are used by the experts for inference and to provide a context for specific content rules. We believe that overviews could also be very useful in computer NLG systems, and are currently working on testing this hypothesis, as part of the SUMTIME project.

Acknowledgements

Many thanks to our collaborators at WNI/Oceanroutes and Intelligent Applications, especially Ian Davy, Dave Selway, Rob Milne, and Jon Aylett; this work would not be possible without them! Thanks also to Sandra Williams and the anonymous reviewers for their comments on a draft of this paper. This project is supported by the UK Engineering and Physical Sciences Research Council (EPSRC), under grant GR/M76881.

References

Allen J. and Perrault C. R. (1980). Analyzing Intention in Utterances. *Artificial Intelligence*, **26**:1-33.

Barzilay R, McCullough D, Rambow O, DeChristofaro J, Korelsky T, and Lavoie B (1998) A New Approach to Expert System Explanations, In *Proceedings of INLG-1998*, pages 78-87.

Boyd S (1997). Detecting and Describing Patterns in Time-varying Data Using Wavelets. In *Advances in Intelligent Data Analysis: Reasoning About Data,* X Lui and P Cohen (Eds.), Lecture Notes in Computer Science 1280, Springer Verlag.

Fiedler A (1998). Macroplanning with a Cognitive Architecture for the Adaptive Explanation of Proofs. In *Proceedings of INLG-1998*, pp 88-97.

Goldberg E, N Driedger and RL Kittredge (1994), Using Natural-Language Processing to Produce Weather Forecasts, *IEEE Expert*, **9**, 2, pp 45-53.

McKeown K, Kukich K, Shaw J (1994). Practical Issues in Automatic Document Generation. In *Proceedings of ANLP-1994,* pp 7-14.

Mittal V, Moore J, Carenini G, and Roth S (1998). Describing Complex Charts in Natural Language: A Caption Generation System. *Computational Linguistics* **24**: 431-467.

Reiter E. and Dale R. (2000) *Building Natural Language Generation Systems.* Cambridge University Press.

Reiter E., Robertson R. and Osman L. (2000) *Knowledge Acquisition for Natural Language Generation.* In Proceedings of the First International Conference on Natural Language Generation (INLG-2000), 217-224 pp.

Shahar Y (1997), “Framework for Knowledge-Based Temporal Abstraction”, *Artificial Intelligence* **90**:79-133..

Sripada S. G. (1997) *Communicating Plans in Natural Language: Planning and Realisation.* PhD Thesis, Indian Institute of Technology, Madras, India.

Sripada S. G. (2001) *SUMTIME: Observations from KA for Weather Domain.* Technical Report, Computing Science Dept. Univ of Aberdeen, Aberdeen AB24 3UE, UK. Awaiting approval from industrial collaborators.