Natural Language Generation for Decision Support[*]

Ehud Reiter[†]

Abstract

One potential use of Natural Language Generation (NLG) systems is to generate texts that help users make decisions. I summarise current and previous work in this area, and explain why I think decision support is potentially a very promising application of NLG. I also briefly describe some of the research challenges in generating decision support texts, and introduce the new BabyTalk project at Aberdeen, which will generate texts for medical decision support.

1 Introduction

Natural Language Generation (NLG) is a technology which is still looking for its 'killer app', that is an application which (a) needs NLG and (b) is very valuable and useful in the real world. I believe that decision support, that is generating texts which summarise data and options, and thus help users make decisions, could be such as 'killer app'. In this note I describe why I think decision support is an exciting application of NLG, and summarise past and current research on this topic. I also briefly describe some of the research challenges for the NLG community in generating decision support texts. I conclude with a brief description of Aberdeen's new BabyTalk project, which will generate texts that help medical professionals make decisions about babies in a neonatal intensive care unit. My summary of past and current research is largely based on responses to a query I posted to the SIGGEN mailing list.

[*] Please cite as Aberdeen University Computer Science technical report AUCS/TR0602

[†] Dept of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK. Email: `ereiter@csd.abdn.ac.uk`

2 Applications of NLG

Perhaps the earliest application of NLG was as a component of machine translation systems. Transfer and interlingua MT systems needed a component which could generate texts in the target language from a semantic or syntactic representation, and NLG technology was used to perform this task. In recent years, however, there has been less interest in using NLG in MT (or indeed in other text-to-text NLP systems, such as summarisation), because the statistical techniques which dominate much current research in text-to-text systems work directly on word strings, and do not need or produce abstract semantic or syntactic representations of texts. Instead, recent interest in applied NLG has focused more on systems which generate texts from non-linguistic input data.

John Bateman and Michael Zock have created a very comprehensive list of NLG systems[1]. From an applications perspective, they classify 340 NLG systems into no less than 113 categories. The largest category, *medical*, includes 18 systems; but these systems are themselves very diverse, ranging from a deployed system that helps doctors write routine documents to a theoretical study of style in patient information leaflets.

The next largest category in Bateman and Zock's list is weather forecast generators (13 systems); this category includes several systems which have been operationally used. I think this category highlights two promising ideas for applied NLG:

- Helping humans produce routine documents more quickly, by generating drafts which humans post-edit into their final shape. All fielded weather-report generators which I am aware of work this way, and this idea has been picked up in many other areas of applied NLG as well.

- Producing texts which help humans make decisions, such as (in the weather forecast case) whether a ship should sail or stay in port.

The first idea has already been explored by a number of people; for example see (Paris et al., 1995) for a discussion of general issues, and (Sripada, Reiter, and Hawizy, 2005) for a detailed study of post-editing of generated texts. In this note I will focus on the second idea, using NLG to generate texts which help people make decisions.

[1] `http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/overview-domain.html`

3 NLG for Decision Support

'Decision support' means different things to different people. In this document, I will use it in the AI sense; IT tools that help users make specific decisions by analysing, summarising, and presenting relevant information. I do not include in this survey systems whose goal is to help a user explore and 'understand' a data set in general terms.

Perhaps the most obvious way to provide decision support (from an AI perspective) is to build a reasoning system which recommends a specific course of action; this of course was the goal of much research in expert systems in the 1970s and 1980s. One general finding of this research was that people wanted explanations as well as recommendations (Davis and Lenat, 1982). But even when explanations were provided, expert systems were not as successful as their early proponents had hoped for. Certainly in medicine usage of expert systems has been low, despite the fact that many expert systems seem to perform as well as or even better than most doctors.

This reluctance to use expert systems was perhaps partially because such systems were not always robust; in many cases existing medical knowledge was insufficient to allow a computer to make reliable recommendations in all cases. Perhaps another factor was that many doctors did not like computers telling them what to do. Hence there was a change of emphasis from building systems which performed doctor-like tasks by themselves, to systems which helped doctors perform such tasks; such systems could be used in cases where lack of medical knowledge made it difficult for a computer to make a decision on its own, and also were more acceptable to many doctors. These systems were called decision support systems.

A recent survey of medical decision support systems (Garg et al., 2005) suggests that well-engineered decision support systems can change doctors' behaviour to conform more closely to 'best practice' guidelines or protocols, especially with regard to details that doctors do not care much about (e.g., for drug prescriptions, it is easier to convince a doctor to change the amount prescribed than which drug is prescribed (Bates et al., 2003)). However, there is less evidence that such systems either improve patient outcomes (health) or decrease medical spending. But perhaps this reflects problems in medical knowledge, not decision support technology. From a computer science perspective, we could argue that a system should be regarded as successful if it encourages decisions that domain experts believe are optimal; if these decisions are not in fact optimal, that is an issue for domain experts, not for computer scientists.

Most existing decision-support systems have concentrated on making specific recommendations about decisions. However, another way of providing decision support is to analyse and summarise the information and data that a decision is based on, perhaps in the context of the specific choices that are available to the decision maker. A doctor may have available to him megabytes (or even gigabytes) of data about a patient, and only a few minutes in which to make a decision. Clearly it is not possible for a human to digest and analyse this amount of data in this short of a time. Hence I believe there is a real need for systems which try to summarise data for doctors (and other medical professionals) in a way which makes it easy for them to pick out key information for making these decisions.

I believe this is true in many other areas, in addition to medicine. Enormous amounts of data are produced in the modern world, far more than any one person can reasonably expect to digest. Society desperately needs technology to help doctors, engineers, logistics planners, and indeed ordinary people cope with this flood of data. The data mining community has developed many techniques for analysing data, but they have not thought much about communicating analyses to human users. Visualisation researchers have looked at communicating data, but they have tended to focus on communicating raw data for data exploration purposes, not communicating summaries of data for decision-making purposes.

I believe that natural language can be a good mechanism for communicating data for decision support, because

- A text can naturally communicate many kinds of information in addition to raw data, including abstractions, interpretations, uncertainty, background information, and causal links.

- A text can be very precise in what information it communicates and what information it does *not* communicate (Bernsen, 1995).

- Many people absorb information more easily from texts than from numerical/graphical presentations; this is especially important for systems which will be used by ordinary people

I am encouraged in my beliefs by an experiment carried out at Aberdeen and Edinburgh which showed that doctors made better decisions when shown text summaries of data instead of graphic visualisations (Law et al., 2005).

Of course I do not mean to imply that data summaries should be presented solely by text, no doubt the optimal presentation will involve a combination of media. But I believe text will play an important role, for the above reasons.

I also believe that decision support is a good venue for testing hypotheses about NLG, because there is a clear way of evaluating ideas and systems; show the output of the system to a user and see how likely he/she is to make a good decision (as in (Reiter, Robertson, and Osman, 2003; Law et al., 2005)). Also decision support systems take as input real data instead of abstract knowledge or linguistic representations; such data is usually easier to acquire than abstract AI representations, and also using real data as input protects us from the accusation that we are using unrealistic inputs.

Working on NLG for decision support could also encourage the NLG community to interact more with the data mining (for data analysis) and HCI (for information presentation) communities, which I think could lead to interesting insights.

4 Current and Previous Work on NLG for Decision Support

4.1 Explaining Reasoning

As mentioned above, an early application of NLG was generating explanations of expert system reasoning; this can be considered as decision support if the expert system in question is helping the user make a decision, and providing an explanation increases the user's willingness to use the expert system. A recent summary of expert system explanation is (Lacave and Diez, 2004), who in fact conclude that explanation systems would benefit from more sophisticated NLG techniques. Lacave and Diez also believe that expert system explanation systems would be more effective if they were better able to interact with users, and if they generated multimodal explanations which included both text and graphics.

Of course, expert systems are just one type of AI reasoning system, and explanations can be generated for other types of reasoning systems as well. In another paper, (Lacave and Diez, 2002) survey explanations of Bayesian networks. Such explanation systems are in many ways similar to explanation systems for symbolic expert systems, although they put more emphasis on communicating numerical information, and on multimedia explanations which include textual and graphical elements.

Research has also been done on generating explanations of mathematical proofs produced by AI theorem provers (another type of AI reasoning system). I will not discuss this here as it is not relevant to decision support; but for those interested, a recent paper with good pointers to previous work is (Fiedler, 2005).

As a very general comment, there seems to be less interest in the NLG community in this area than there was 10-20 years ago, despite the fact that at least some people working in the area of explanation of AI reasoning believe that better NLG could help (Lacave and Diez, 2004).

4.2 Describing Options

There is a growing amount of research in using NLG to describe and explain choices, often based on detailed user models that describe users interests, preferences, constraints, etc. For example, GEA (Carenini and Moore, 2001) generated user-tailored descriptions of houses which helped people make real estate decisions; FLIGHTS (Moore et al., 2004) generated user-tailored descriptions of flights for people who needed to plan a travel itinerary; MATCH (Walker et al., 2004) generated user-tailored descriptions of restaurants for people who wanted to decide where to eat; and MADSUM (Harvey, Carberry, and Decker, 2005) generated user-tailored descriptions of stocks, for users who were deciding how to change their stock market portfolio.

Describing choices could be especially important in medicine, as there is growing legal and ethical pressure for the medical community to involve patients in decision making. But patients cannot make sensible decisions about complex medical options unless they understand the risks and benefits of the medical treatments and procedures being proposed. Unfortunately it is difficult for time-pressured doctors to fully explain medical options; hence there is a need for educational material and information aids for patients (Doyal, 2001). Some NLG researchers are exploring whether NLG technology can be used to generate such material. For example (Di Marco et al., 2005) propose an NLG system which explains the implications of different reconstructive surgical procedures.

A key issue in explaining medical options is communicating risk. There is an extensive psychological and medical literature on risk communication and perception (e.g., (O'Connor, Légaré, and Stacey, 2003)), but this does not seem to have had much impact on NLG. One exception is RAGs (Coulson et al., 2001), which communicated risk using a cognitively motivated argumentation strategy.

4.3 Persuading Users

There is also interest in the medical community in trying to persuade patients to change their lifestyle in medically useful ways, such as stopping smoking or improving their diet. In other words, in trying to persuade people to make the 'right' decision about their lifestyle. Several NLG systems have been developed which generated this kind of text, including the Aberdeen STOP system (Reiter, Robertson, and Osman, 2003), which attempted to persuade smokers to make an attempt to stop smoking. Several other systems which generate persuasive medical texts are described in a recent symposium (Bickmore and Green, 2006).

A clinical trial of STOP showed that it was not in fact effective at helping people stop smoking. I personally believe that this is because of STOP's limited user models and lack of any kind of user interaction; this meant that STOP could not focus on the specific interests, concerns, and background of individual smokers.

Systems that generate persuasive texts have been developed in many other domains as well, including education, culture, and marketing. A good recent overview of this area is Guerini's thesis (Guerini, 2006). Guerini argues that persuasive NLG systems should be based on the psychological models of persuasion as well as logical models of argumentation, and should consider affective (emotional) impact of choices (de Rosis and Grasso, 2000).

4.4 Summarising Data

Last but not least, an NLG system can also help people make decisions by summarising data sets that influence the decision. Perhaps the classic example is NLG systems which generate weather forecasts, as mentioned above. Most such systems generate texts for specific audiences who need to make decisions based on this information. For example, Aberdeen's SumTime-Mousam system (Reiter et al., 2005) generates forecasts for offshore oil rigs; these forecasts are intended to help rig staff make decisions about unloading supply boats, scheduling diving operations, and so forth.

While most work in this area has probably been done with weather forecasts, there is growing interest in generating data summary texts in other areas such as medicine and engineering. For example, MAGIC (McKeown et al., 2000) generates summaries of key relevant information for patients who are entering an intensive care unit (ICU); this helps ICU staff treat the patient appropriately. SumTime-Turbine (Yu et al., 2003) generates summaries of sensor data from gas turbines, which are intended to help engineers decide what maintenance actions (if any) should be performed.

One of the challenges in building systems which generate textual summaries of data is integrating data interpretation and abstraction techniques with NLG. Little seems to have been published on this; one exception is

(Sripada et al., 2003a), who argue that data interpretation techniques used in a system which produces textual data summaries need to be modified to better conform to the Gricean maxims.

In theory, one of the main advantages of presenting a data summary as a text instead of as a list of bullet points is that the structure of the text (eg, the order information is presented in, cue phrases and discourse structure, sentence and paragraph breaks) should help readers digest and understand the content of the text. Unfortunately, there does not seem to be much recent research in the NLG community on improving text structuring.

4.5 General Comment: Why isn't this in the NLP Literature

A striking aspect of the above literature survey is that *none* of the papers cited in this section appear in the computational linguistics literature. I have cited many papers in the general AI literature, many papers in domain literature (e.g., medicine), and a few papers in the user modelling literature. But none of the papers cited above appear in INLG, ACL, *Computational Linguistics*, or any other venue which focuses on natural language processing.

I suspect that this is because papers on generating texts for decision support often tend to focus more on data analysis, user modelling, and domain requirements than on linguistics, and hence authors feel their papers would not be welcome at NLP events. In other words, such papers might not fit the definition of NLG suggested by (Evans, Piwek, and Cahill, 2002). However, I must admit that I personally think that it is a shame that such work is not presented at NLG events. I believe that the NLG community needs to look at content-related issues if it is going to develop useful real-world technology; I also believe that from a theoretical perspective, we cannot ignore content issues if we want to understand how language relates to the world (Roy and Reiter, 2005).

5 Research Challenges

There are of course many research challenges in developing NLG systems and technology for decision support. I list below a *few* challenges which are of particular interest to me and also to many of my colleagues at Aberdeen (these emerged from a discussion in the Aberdeen NLG group):

- *Combining textual and graphical decision support:* This seems especially important for explanations (Section 4.1) and data summarisation (Section 4.4). A related question is how we enable users to effectively interact with multimodal (or indeed text-only) decision support systems (they can already interact with graphics-based decision support).

- *Effectively communicating individual bits of information:* For example, how do we effectively communicate risks and probabilities, especially to non-technical users? I believe this is one of the main challenges in effectively describing options (Sections 4.2).
- *Understanding emotional/affective impact:* For example, how should we communicate bad news about the user's health; is it best to do this directly or indirectly, and how does this choice depend on the user's personality? I think this is a fundamental question for persuasive NLG (Section 4.3).
- *Integrating NLG with reasoning/interpretation systems:* How do we integrate an NLG system with a data interpretation (or reasoning) system? What constraints do we need to impose on the interpretation and NLG systems in order to allow them to work together? This is very important for data summarisation (Section 4.4) in particular.

6 BabyTalk

This note was written as a preliminary step in a new project at Aberdeen, BabyTalk (`http://www.csd.abdn.ac.uk/research/babytalk/`). BabyTalk will generate summaries of medical data about premature babies in a neonatal intensive care unit (some background and a very initial demonstrator is described in (Sripada et al., 2003b)). In fact we intend to generate three kinds of summaries, for three different purposes

- *Decision support*: summaries which help doctors and other medical professionals decide what actions (if any) should be done. This is partially inspired by an experiment which showed that manually written text summaries were useful decision aids in this domain (Law et al., 2005).
- *Report drafting*: drafts of shift summaries, which nurses post-edit (nurses are required to write these the end of each shift). The primary goal is to speed up the production of these documents, although we also hope to improve their usefulness as decision aids by making them more accurate and complete.

- *Parent reports*: summaries of a baby's status for the baby's parents. These are primarily intended to reassure parents and reduce their stress, although it is possible they could also play a role in informed decision making.

We hope to investigate all of the issues mentioned in Section 5 in the course of developing the above systems.

Acknowledgements

The literature review in Section 4 was largely based on suggestions from members of the SIGGEN mailing list. My thanks to Giuseppe Carenini, Michael Elhadad, Nancy Green, Marco Guerini, Terry Harvey, Ed Hovy, Guy Lapalme, Johanna Moore, Yaji Sripada, Sandra Williams, and Ingrid Zukerman for their help and suggestions. The challenges mentioned in Section 5 are based on a discussion with other members of the Aberdeen NLG group. BabyTalk is supported by the UK Engineering and Physical Sciences Research Council (grant EP/D049520/1).

References

Bates, David, Gilad Kuperman, Samuel Wang, et al. 2003. Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, 10:523–539.

Bernsen, Niels. 1995. Why are analogue graphics and natural language both needed in HCI? In F. Paternó, editor, *Design, Specification and Verification of Interactive System*, pages 235–251, Heidelberg. Springer-Verlag.

Bickmore, Timothy and Nancy Green, editors. 2006. *Papers from the AAAI-06 Spring Symposium on Argumentation for Consumers of Healthcare*. AAAI Press.

Carenini, Giuseppe and Johanna D. Moore. 2001. An empirical study of the influence of user tailoring on evaluative argument effectiveness. In Bernhard Nebel, editor, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, pages 1307–1314. Morgan Kaufmann.

Coulson, Andrew, David Glasspool, John Fox, and Jon Emery. 2001. RAGs: A novel approach to computerised genetic risk assessment and decision support from pedigrees. *Methods of Information in Medicine*, 40:315–322.

Davis, Randall and Douglas Lenat. 1982. *Knowledge-Based Systems in Artificial Intelligence*. McGraw Hill.

de Rosis, Fiorella and Floriana Grasso. 2000. Affective natural language generation. In Ana Paiva, editor, *Affective Interactions*. Springer, pages 204–218.

Di Marco, Chrysanne, Peter Bray, Dominic Covvey, et al. 2005. Authoring and generation of tailored preoperative patient education materials. In *Proceedings of the UM05 Workshop on Personalisation for e-Health*.

Doyal, Len. 2001. Informed consent: moral necessity or illusion? *Quality in Health Care*, 10:i29–i33.

Evans, Roger, Paul Piwek, and Lynne Cahill. 2002. What is NLG? In *Proceedings of the Second International Conference on Natural Language Generation*, pages 144–151.

Fiedler, Armin. 2005. Natural language proof explanation. In Dieter Hutter and Werner Stephan, editors, *Mechanizing Mathematical Reasoning*. Springer Verlag, pages 342–363.

Garg, Amit, Neill Adhikari, Heather McDonald, et al. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *Journal of the American Medical Association*, 293:1233–1238.

Guerini, Marco. 2006. *Persuasion Models for Multimodal Message Generation*. Ph.D. thesis, IRST, University of Trento, Italy.

Harvey, Terrence, Sandra Carberry, and Keith Decker. 2005. Tailored responses for decision support. In Liliana Ardissono, Paul Brna, and Antonija Mitrovic, editors, *Proceedings of User Modeling 2005*, pages 164–168. Springer.

Lacave, Carmen and Francisco Diez. 2002. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17:107–127.

Lacave, Carmen and Francisco Diez. 2004. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19:133–146.

Law, Anna, Yvonne Freer, Jim Hunter, Robert Logie, Neil McIntosh, and John Quinn. 2005. Generating textual summaries of graphical time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

McKeown, Kathlenn, Desmond Jordon, Steven Feiner, et al. 2000. A study of communication in the cardiac surgery intensive care unit and its implications for automated briefing. In *Proceedings of AMIA-2000.*

Moore, Johanna, Mary Ellen Foster, Oliver Lemon, and Michael White. 2004. Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of FLAIRS 2004.*

O'Connor, Annette, France Légaré, and Dawn Stacey. 2003. Risk communication in practice: the contribution of decision aids. *British Medical Journal*, 327:736–740.

Paris, Cecile, Keith Vander Linden, Marcus Fischer, Anthony Hartley, Lyn Pemberton, Richard Power, and Donia Scott. 1995. A support tool for writing multilingual instructions. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-1995)*, pages 1398–1404.

Reiter, Ehud, Roma Robertson, and Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter, and Jin Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Roy, Deb and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167:1–12.

Sripada, Somayajulu, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learned. In *Proceedings of ENLG-2005*, pages 133–139.

Sripada, Somayajulu, Ehud Reiter, Jim Hunter, and Jin Yu. 2003a. Generating English summaries of time series data using the Gricean maxims. In *Proceedings of KDD-2003*, pages 187–196.

Sripada, Somayajulu, Ehud Reiter, Jim Hunter, and Jin Yu. 2003b. Summarising neonatal time-series data. In *Proceedings of EACL-2003*, pages 167–170.

Walker, Marilyn, Stephen Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.

Yu, Jin, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. 2003. SumTime-Turbine: A knowledge-based system to communicate gas turbine time-series data. In *Proceedings of IEA/AIE-2003*, pages 379–384.