From: AAAI Technical Report SS-98-06. Compilation copyright © 1998, AAAI (www.aaai.org). All rights reserved.

A Grounded Theory Approach to Expert Summarizing

Brigitte Endres-Niggemeyer
Polytechnic of Hanover
Department of Information and Communication
Hanomagstr. 8
D-30449 Hannover, Germany
phone +49 511 92 96 606
fax +49 511 92 96 610
ben@iks.ik.fh-hannover.de

Abstract

I report on an empirical approach to expert summarizing, emphasizing the qualitative modeling methods that shaped the system. In the framework of model-based knowledge engineering, the empirical grounded theory of summarizing is subsequently transferred to a small simulation system called SimSum (Simulation of Summarizing). SimSum shows that an empirical design of a summarizing system is feasible, and it explains the human performance by means of a multimedia presentation.

1. Introduction

As the title promises, this paper is about empirical methods and uses findings (in detail in Endres-Niggemeyer 1998) for illustrative purposes only. I argue that it is feasible to ground a summarization system on observables of human expert summarization. Wherever possible, this should be done.

To my knowledge, many designers implement their conjecture of summarizing, without probing the question of how summarizing works in human minds. They are satisfied as soon as their system can construct a defendable short version of a given text or information source. Often, their system organization reflects facts which are a bit beside the point, such as the subdivision of linguistics. In this case, the system typically begins processing with a morphological analysis of its input, applies the findings of syntax, afterwards calls upon semantic theories, and so on. Researchers concerned with poor computational resources or with processing speed in real-time applications may design their summarizing systems according to these constraints. It goes without saying that summarizing systems of this type are not intended to incorporate a theory of human summarizing, and it is at least debatable whether any fashion to achieve a justifiable short version of a text should be dubbed summarizing.

It is, however, clear that human expert summarizers possess and apply valid summarization know-how. The problem is to reconstruct their knowledge in the form of a theory and to implement a system that incorporates this theory. Precisely this is the main endeavor of SimSum, the simulation of human expert summarizing which is the background of this article.

In the following, I discuss the empirical methods that enabled system design by means of observation. Towards the end of the paper, the subsequent SimSum development story is told in short, simply to explain that the research really reached its goal. For lack of space in the printed paper, the discussion of the empirical methods must concentrate on principles. More detail is presented in Endres-Niggemeyer 1998.

2. Grounded Theory Development

The overall research procedure leading to the SimSum system is inspired by the KADS methodology of model-driven knowledge engineering (Schreiber, Wielinga, and Breuker 1993). KADS tells us to set up an empirical domain model first, and to think of implementation only later.

2.1 Empirical Methods for Modeling Complex Cognitive Processes

The empirical model of professional summarizing is a *grounded theory* (Glaser and Strauss, 1980, Strauss 1987), the outcome of qualitative field research (Diesing 1971, Mayring 1990) or *naturalistic enquiry* (Lincoln and Guba 1985). The stance of grounded theories is that meaningful theories about social phenomena, which are often complex by nature and poorly understood, can be obtained inductively by systematic data analysis. As humans react to their environment, the researcher has to study their behavior in its natural setting instead of laboratory tests. The initial knowledge about the domain is expanded incrementally by new concepts developed from empirical observation. Whether the model is valid is shown predominantely according to the coherence principle of truth. The more comprehensive and complex a network of statements is, the stronger are the constraints it imposes on new hypotheses. Since every hypothesis has to stand the pressure of many other observed facts, the survival chances of wrong hypotheses are indeed scarce.

The investigation of cognitive processes such as summarizing needs methods which are called *introspective* because they presuppose a person who reports what happens in her or his mind. A detailed discussion of introspective data and methods is given in (Ericsson and Simon 1984). In particular, thinking-aloud protocols are applicable for learning about summarization processes. Obeying the basic thinking-aloud instruction, the subject talks about all operations during task execution and thus records all states of the process. We obtain rich introspective data, but the record is also prone to gaps and distortions. Interpretation therefore draws upon additional evidence from input, output and other sources in order to reconstruct a more complete image of the cognitive activities from of their traces in the thinking-aloud protocol.

For interpreting introspective data about such complicated processes as summarizing, we are lost without a sophisticated interpretation framework based on preexisting theories or models. Fortunately, summarizing is a part of discourse processing. The cognitive model of discourse processing proposed by (Kintsch and van Dijk 1983) can therefore guide the interpretation of summarization process records. It provides an idea sketch of overall summarization processes and of specific strategies, based on experimental results and theoretical considerations. Besides the discourse processing model, the writing model for expository texts developed by (Hayes and Flower 1980) has been helpful in setting up a summarization model. It applies since summary writing is a variation of expository writing. Whereas the (Kintsch and van Dijk 1983) model is not very explicit about process organization, the writing model states stages of the process and broad hypotheses about their internal structure.

The summarization model reconstructs from observation a human knowledge processing system which generates the observational data, according to the line proposed by (Bereiter and Scardamaglia 1987). In the operational empirical phase of the project, 6 experts took thinking-aloud protocols of 9 summarizing processes per person: three examples of abstracting a short document (conference paper or journal article), abstracting a long document (report or monograph), and indexing and classifying a paper, respectively. All experts remained in their natural working context. They used 52 different documents of their personal choice, mostly taken from routine input. The group was composed of persons who combined summarization expertise as abstractors, indexers and technical editors with at least some experience in teaching. Teaching experience was required from experts to avoid a problem well-known in the knowledge engineering literature: experts who can perform a task without the ability to explain it. An expert who has to transmit his or her intellectual skills to others must reflect on them and verbalize them. (S)he will be able to describe them during thinking-aloud, too.

2.2 Deriving Summarizing Strategies from Observation

Grounded theories are constructed inductively in a theoretical framework. Since the model of summarizing is a process model, it is first and foremost constructed from process concepts, called strategies according to (Kintsch and van Dijk 1983). The strategies are the empirical foundation of the implemented cognitive agents (see below). For every expert an individual process model of summarization was elaborated. The data interpretation procedure was the same for all processes, including the transcription of the audiotape, subdivision of the working process into steps, their interpretation in terms of the discourse processing model, discussion of the interpretation with the expert (with correction if needed), and organization and presentation of the model.

Inductive concept formation works by abstracting from data, in the present case by defining strategies from the evidence given in a working step, under the constraints which are given by the theoretical framework, the whole of the observational data and the method of grounded theory development. Combined constraints largely preclude unfounded hypotheses, without excluding alternative interpretations. The researcher assigns observed cognitive activities to strategies - recurring tasks of cognitive processing - until all activities are explained. If a suitable strategy is missing, it is created. If a hypothesis turns out to be wrong, the whole theory is reorganized.

A tiny example illustrates the discovery of intellectual strategies. A thinking-aloud protocol begins with the following utterance:

> So (noise) reading introductory material (noise) eh international operations of one kind another...

It is easy to set up hypotheses about the cognitive strategies which have left their traces in this short passage:

- There was a start strategy (recorded by the interjection "so" and the evidence that something happened after this moment).
- A reading activity was started (evidence is provided by the word "reading" and the actual reading from the input "eh international operations...").
- The summarizer applied document structure knowledge. It enabled him to recognize that he is looking at an introduction.

In the investigation reported here the three candidate strategies are supported by much more evidence. They gave rise to valid cognitive strategies and to agents of the SimSum system.

2.3 Global Features of Professional Summarizing

The process model of summarizing first and foremost describes the cognitive activities that take place during summarizing: assimilating information, selecting, transforming, reformulating and presenting. It explains the individual thinking procedures by reducing them to simpler, more transparent cognitive acts and by indicating their reasons and their effect on data. Main findings include

- 552 summarization strategies which have been observed in the empirical investigation. They have been ordered according to their function and put into the summarizers' intellectual toolbox.
- the organization of human summarizing processes presenting a sequence of working steps
- the cooperation structure inside working steps, where cognitive strategies share in data and tasks.

The following small illustrative sample shows a functional class in the intellectual toolbox. The experts who use the respective strategies are listed in parenthesis.

4 prof A32.222: Relevance judgements based on information value

288. *no-truism:* Leave out what is trivial. (Andreas, Edward, Hanne, Harold, Inge, Marliese)

289. *no-void:* Leave out what lacks content. (Edward, Hanne, Harold, Inge, Marliese)

290. *own-only:* No cited results, own results of the authors only. (Hanne, Marliese)

291. *relevant-contrast:* Important is what stands out from the rest. (Harold, Marliese)

292. *relevant-new:* Important is what is new and original. (Edward, Hanne, Harold, Inge, Marliese)

3. System Design and Implementation

After empirical modeling, the research went through the system design phase that reconstructs the empirical (or conceptual) model with computer-oriented concepts and theories, yielding an implementable model. It was transformed into a software product, the SimSum system included in (Endres-Niggemeyer 1998). The multimedia presentation is done with the Macromedia Director, the cognitive simulation is programmed in CLOS (Common Lisp Object-Oriented System). At the core of the simulation, cognitive agents realized as Lisp objects carry out summarization strategies observed in expert summarizers. On a CD-ROM that complements a text book, SimSum helps to explain better what it means to summarize.

Acknowledgements

I thank the German Science Foundation (DFG) for supporting the empirical research (grant En 186/1-3) and the German Federal Ministry for Education and Research (BMBF) for funding the implementation (grant F0916.00).

References

Bereiter, C., and Scardamaglia, M. 1987. *The psychology of written composition.* Hillsdale NJ: Erlbaum.

Diesing, P. 1971. *Patterns of discovery in the social sciences.* Chicago: Aldine Atherton.

Endres-Niggemeyer, B. 1998. *Summarizing information.* Heidelberg: Springer. Forthcoming.

Endres-Niggemeyer, B., and Neugebauer, E. 1998. Professional summarizing: No cognitive simulation without observation. *Journal of the American Society for Information Science.* Forthcoming.

Ericsson, K.A., and Simon, H.A. 1984. *Protocol analysis: Verbal reports as data.* Cambridge MA: MIT Press.

Glaser, B.G., and Strauss, A.L. 1980. *The discovery of grounded theory: strategies for qualitative research.* 11. ed. 1980. New York: Aldine Atherton.

Hayes, J.R., and Flower, L.S. 1980. Identifying the Organization of Writing Processes. 17-35 in Gregg, L.W.; Steinberg, E.R. eds. 1980. *Cognitive processes in writing.* Hillsdale NJ: Erlbaum.

Kintsch, W., and Dijk, T.A.van 1983. *Strategies of Discourse Comprehension.* Orlando FLA: Academic Press.

Lincoln, Y.S., and Guba, E.G. 1985. *Naturalistic inquiry.* Beverly Hills CA: Sage.

Mayring, P. 1990. *Einfuehrung in die qualitative Sozialforschung* [Introduction into qualitative social research]. Muenchen: Psychologie-Verlags-Union.

Schreiber, G., Wielinga, B., and Breuker, J. 1993. *KADS. A principled approach to knowledge based system development.* London: Academic Press.

Strauss, A. L. 1987. *Qualitative analysis for social scientists.* Cambridge: Cambridge University Press.