Automatic generation of texts without using cognitive models: television news

Schmitz, Ulrich (1994)

In: Susan Hockey/ Nancy Ide (eds.): Research in Humanities Computing 2. Oxford: Clarendon Press 1994, pp. 186-192

0. Summary

Many types of texts - though not all of them - are so monotonous on the language level that they can be produced automatically without any very complicated cognitive AI models. A meticulous corpus analysis can reveal how similar elements and rules are used over and over again: one then only needs to reproduce them with a computer. This procedure will be demonstrated and discussed using television news.

1. Text generation by persons and computers: TV news

Automatic generation of texts in natural language is considered to be one of the most difficult areas within the language-orientated AI research (cf. Danlos 1987, Kempen 1987, MacKeown 1985 or Zock/Sabah 1987). One can dispense with theoretically and technically complicated models of AI, however, in many more areas of automatic text generation than commonly assumed. Television news here will be used to show how even highly complex texts, where one does not usually notice the "continuity of repetition" (Schleiermacher 1977, 82), can be generated by computer in a basically quite simple way, more simply at any rate than the research in newspaper articles and computers leads one to imagine (Cullingford 1977, Rösner 1986, Weber 1986).

A detailed study of a fairly large corpus of German television news (ARD channel, 2 months of the 8 p.m. main bulletin, i.e. some 110.000 words) was conducted using in part the classical tools of linguistic data analysis, stati- stics and text-analysis (Schmitz 1990). It revealed that, owing to particular pragmatic conditions, such news texts consist of an effective technique of moderately random text production using ready-made and semi-ready- made parts. This

rational procedure, which can be precisely described, en- ables the news writers to balance the tension between bringing the news fast and remaining objective. If texts are written or edited with a routine programme - as Luhmann (1971, 118) calls it - one saves time and wipes out any subjective traces left by individual authors.

Particularly routine programmes are not difficult to computerize. In such cases, text production does not have to be modelled as a cognitive process. There is no tension between a linear text and multi-dimensional knowledge (as often is in text production, cf. Rothkegel 1989). In contrast to the most frequently analysed types of text (e.g. stories, essays, dialogue systems) knowledge really only appears here as a "building block" without any inner logical links. In order to produce such a text one does not have to draw from implicit knowledge which is not formulated in the text. Furthermore, excepting stylistic variations and indexical elements, all the expressions and meanings are extremely stereotyped. And finally, one only needs to simulate the monological production of these not particularly coherent texts, i.e. without having to consider the essentially more complex conditions for dialogues which would require having to simulate the processes of text understanding.

2. Material and construction of the natural language text

Discarding the weather forecast, the 8 p.m. main bulletin of the "Tagesschau" consists of approximately 9 to 13 news items coming from a manageable number of subject areas. (The more loosely one defines the boundaries of the subject areas, the less they change over a certain period of time.) Each item contains one or more of seven text sorts (on- and off-speaker, on- and off-correspondent, interview question and answer, excerpt of a speech/statement) in varying combinations. The transition from one text sort to the other can be determined statistically according to the subject area; apart from that they do not depend on the content of the item.

The items vary in length but, as a rule, they total about 101 to 120 sentences. Sentences represent the key unit in the production of texts. The first sentence of an item contains the core of the news, and the rest of the item refers to it in some way or another. Thematic (not, however, linguistic or narrative) coherence is the only means of guaranteeing the unity of one whole item. Unlike almost any other text genre, no other structure exists beyond the sentence level if we forget the few anaphoric pronouns which could very well be produced by a random number generator. In other words: apart from the first sentence, all the other sentences in an item could, in principle, be strung to one another in any order.

The fact that all the elements can be mixed with one another is characteristic of the text production of the "Tagesschau" not only on the sentence level but also on two further levels. The first one is the general way facts are presented (e.g. stating sources, events leading up to a given event, concomitant circumstances). The empirical analysis shows that there are 38 moves which are typical of news bulletins and which can be combined in almost any order, i.e. not in a logical or conventionally ruled order.

The second level is the fund of meanings which ought to serve to describe reality. The "Tagesschau" uses semantic stereotypes interlocked with the ongoing text according to a multi-level hierarchy. There is a standing repertoire of semes typical of this text genre realized either as single words (e.g. "apparently") or as stylistically variable syntagms (e.g. "is

imminent"). The "Tagesschau" uses a set of 14 imaginary pictures (e.g. "Travel and Encounters", "Death, Accidents, and Bad Weather") along with 255 small semantic fields, which could be grouped in 23 larger ones such as "Toil, Obstacles, and Progress", "Confidence and Danger".

In each of these pictures and fields there are specific and recurrent formulations which can be mixed beyond the boundaries of the given pictures or fields in almost any order to compose whole sentences. The complete text (including the opening sentences of the items) can be composed of semantic building blocks combined freely. If, sticking to our metaphor, we would want to put away the blocks after use, we would find that they all fitted in a two-dimensional meaning space (with the coordinates "Good/ Disastrous" and "Little/much Movement") which would enable all the non-deictic "pieces of language" composing the sentences to be determined non-ambiguously according to its coordinate points.

Thus in this model the semantic stereotypes (in several stylistic variations) can be formulated very closely to language because in the "Tagesschau" on the literal level the same elements and set phrases occur (e.g. "said today that", "the reason mentioned was", "ended his/her X-day visit to Y"). This means that the grammar to be included can be limited to few elementary rules (e.g. agreement rules for number).

The sentences consist, however, of two kinds of text pieces which, roughly speaking, can be related to the "Symbolfeld" or the "Zeigfeld" (Bühler 1934, 149 ff).

The first kind consists of the virtually constant number of semantic stereotypes which, as we have sketched above, can be mixed in an almost random way to create a sentence meaning. The variables for the "Zeigfeld" arising here can be replaced by the corresponding names taken from a data bank. By these we mean proper names or words similar to proper names; they make up one seventh of all the word tokens occurring in the text. This data bank must be variable, in principle. De facto, however, within a decade it hardly ever needs to be brought up to date in the main areas (e.g. names for countries and institutions) and only comparatively seldom in the other areas (e.g. surnames). The same applies to the probability of their occurring in the text.

Thus, in principle, the text of a "Tagesschau" bulletin can be produced according to simple rules using almost random choice and mixture of a limited supply of expressive and linguistic possibilities. This is due to the fact that each not too small a sample of actually broadcast "Tagesschau" texts (over one month, for instance) is part of the same discourse world and differs from other samples of the same size only (1) by another mixture of the pieces belonging to the same "Symbolfeld", (2) by other indexical elements (particularly other proper names) and (3) by stylistic variations in unimportant details. A computer with a specially designed program with access to a data bank for the relevant proper names, which would be kept up to date from time to time, could in ten or twenty years still produce texts "worthy" of being broadcast.

3. A machine for producing German news bulletins

Due to the comparatively simple structure of the object to be simulated, there is absolutely no need for constructing a knowledge-based system in the sense of artificial intelligence. Neither

semantic networks nor precisely devised frames or scripts are required (Brachman 1977, Minsky 1975, 1981, Metzing 1980, or Schank/Abelson 1977; cf. Wettler 1980) in order to represent contextual knowledge needed for text production and understanding. (Besides, owing to the thematic spectrum of the news, the total contextual knowledge would take up much more space than all natural-language oriented AI systems designed to date taken together; cf. McKeown 1985 for the relation between technical effort and size of discourse areas.) Rather, a variable form of text templates could be quite sufficient: it would be something like a book of patterns with partly ready text pieces which a formulating machine could use as raw material for producing the final texts.

The simulated text of the main German television news bulletin can be produced with a computer in ten successive steps:

1. Number and topic of items and number of sentences they consist of
2. Transitions from one sort of text to the other within each item
3. Shortened form of the first sentence of each item
4. Sentence by sentence listing of the moves typical of news bulletins
5. Combining of the semantic stereotypes for each sentence
6. Producing provisional text patterns for the "Symbolfeld" of each sentence (including variables for the indexical elements)
7. Further formulating to partly finished sentences (possibly using the dictionary of syntagmatic constructions and the frequency word book)
8. Instantiation with the indexical constants which are relevant for that day
9. Printing out of the complete news bulletin
10. If desired, editing by the user.

The user can intervene in any one of the steps number 1, 2, 3, 8 or 10.

Schmitz (1990) presents a PROLOG program for the first five steps. The other five steps are either trivial or a question of much not particularly difficult work. A complete system would require extensive yet hardly structured elements of a data bank. To date there exist only the following three fragments:

a) A dictionary of proper names and other indexical terms (including their links with particular topics and the probability of their appearing there)

b) A dictionary of syntagmatic constructions (including the small fragments of grammar required for getting correct sentences and stylistic variations)

c) A list of the short versions of the schemata for the sentences used to begin a news item.

4. Simulation and criticism of reality

The program is based on an empirical analysis of news bulletins which have actually been broadcast; it could, however, keep on generating texts which could be broadcast right into the next century. This is due to the fact that the news texts acquire their topicality solely from the random mixture of constant elements belonging to the "Symbolfeld" and the variables belonging to the "Zeigfeld" valid at the time.

In example sentences one usually finds "what has been reluctantly thought up in order to fill previously constructed schemata" (Lipps 1938, 20, our own translation). We find no traces of such an operation in the (not particularly "intelligent") program, and this is due to the fact that even the "Tagesschau" texts written by persons - as opposed to machines - were constructed according to the same procedure. (As in the case of the construction of expert systems, the implicit knowledge always present in what the experts do was merely formulated when constructing the program.) The editing team of the "Tagesschau" can make do with a very economic procedure because the "Zeigfeld" actually relates the text to the day's reality. "What is said with it is raining is an insight which can be used only in a particular situation" (ibid. 22). Although one hardly ever notices it, the same set phrases always mean something else in one's everyday life because one is directly involved in the situation - in this fleeting practice. In contrast to that, mass media must first produce the relation to a situation; leaving pictures aside, television news does this by means of indexical expressions.

The concept presented here for a computer simulation of the German television news can be perceived both as a project for the entire setting up of a news machine and as criticism of the hollow simplicity of the usual news bulletins. In other (polemic) words: artificial intelligence - of a particularly simple kind - could replace natural stupidity or could show it up. The political conclusions of the "dequalification of knowledge leading to the fact that people are informed rather than being able to discuss issues" (Schmidt 1986, 22, our own translation) have not yet been drawn. First of all, one should consider what happens when "the power to regulate and therefore reproduce tends to be taken away from the administrators and turned over to machines" (Lyotard 1986, 52, our own translation). This should be given particular thought when considering the mass media.

References

- Brachman, R. (1977): 'On the epistemological status of semantic networks'. In: Findler, N. (ed.): Associative Networks. New York, pp. 3-50
- Bühler, Karl (1934): Sprachtheorie. Die Darstellungsfunktion der Sprache. Jena
- Cullingford, Richard Edward (1977): Script Application: Computer Understanding of Newspaper Stories. Yale University, Ph.D.
- Danlos, Laurence (1987): The Linguistic Basis of Text Generation. Cambridge
- Kempen, Gerard (ed. 1987): Natural Language Generation. New Results in Artificial Intelligence, Psychology, and Linguistics. Dordrecht
- Lipps, Hans (1938): Untersuchungen zu einer hermeneutischen Logik. Frankfurt/M.
- Luhmann, Niklas (1971): 'Lob der Routine' (1964). In: Luhmann, Niklas: Politische Planung. Aufsätze zur Soziologie von Politik und Verwaltung. Opladen, pp. 113-142
- Lyotard, Jean-François (1986): Das postmoderne Wissen. Ein Bericht (fr. 1979). Graz, Wien
- McKeown, Kathleen R. (1985): Text Generation. Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge
- Metzing, Dieter (ed. 1980): Frame Conceptions and Text Understanding. Berlin/W., New York
- Minsky, Marvin (1975): 'A framework for representing knowledge'. In: Winston, Patrick H. (ed.): The Psychology of Computer Vision. New York, pp. 211-280
- Minsky, Marvin (1981): 'A Framework for Representing Knowledge' (1975). In: Haugeland, John (ed.): Mind Design. Philosophy, Psychology, Artificial Intelligence. Cambridge/Mass., London, pp. 95-128

- Rösner, Dietmar (1986): Ein System zur Generierung von deutschen Texten aus semantischen Repräsentationen. (Diss.) Stuttgart
- Rothkegel, Annely (1989): 'Textualisierung von Wissen. Einige Forschungsfragen zum Umgang mit Wissen im Rahmen computerorientierter Textproduktion'. In: LDV-Forum 6, No. 1, pp. 3-13
- Schank, Roger C/ Abelson, Robert P (1977): Scripts, Plans, Goals and Understanding. An Inquiry into Human Knowledge Structures. Hillsdale, N.J.
- Schleiermacher, F[riedrich] D[aniel] E[rnst] (1977): Hermeneutik und Kritik (1838). (Ed. Manfred Frank). Frankfurt/M.
- Schmidt, Burghart (1986): Postmoderne - Strategien des Vergessens. Ein kritischer Bericht. Darmstadt, Neuwied
- Schmitz, Ulrich (1990): Postmoderne Concierge: Die "Tagesschau". Wortwelt und Weltbild der Fernsehnachrichten. Opladen
- Weber, Heinz J. (1986): 'Faktoren einer Textbezogenen Maschinellen Übersetzung: Satzstrukturen, Kohärenz- und Koreferenz-Relationen, Textorganisation'. In: Bátori, István/ Weber, Heinz J. (ed.): Neue Ansätze in Maschineller Sprachübersetzung: Wissensrepräsentation und Textbezug. Tübingen, pp. 229-261
- Wettler, Manfred (1980): Sprache, Gedächtnis, Verstehen. Berlin/W., New York
- Zock, Michael/ Sabah, Gérard (eds. 1987): Advances in Natural Language Generation: An Interdisciplinary Perspective. London, Norwood/NJ