Integrating engineering and linguistic knowledge to build a domain-independent text generation system

Xinyu Deng * Sadao Kurohashi † Satoshi Sato ‡ Jun-ichi Nakamura ‡

* Corporate Research & Development Center, Toshiba
deng@eel.rdc.toshiba.co.jp

† Graduate School of Information Science and Technology
The University of Tokyo, kuro@kc.t.u-tokyo.ac.jp

‡ Graduate School of Informatics, Kyoto University
sato@i.kyoto-u.ac.jp

1 Introduction

Natural language generation (NLG) is the automatic generation of natural language by computer in order to meet communicative goals (Smith, 1995). Many natural language generation systems have been developed to generate coherent texts (Hovy, 1991; Wanner and Hovy, 1996). Until now, almost all of the existing generation systems are domain-dependent, i.e., they were built for some specific applications. For example, IDAS produces on-line hypertext help messages for users of complex machinery (Reiter et al, 1995); STOP is a natural language generation system which produces personalised smoking-cessation letters (Reiter et al, 1999). From the economic perspective, the need of reuse reveals the usefulness of a general-purpose, domain-independent text generation system. However, how to build a domain-independent generation system is still regarded to be a difficult problem.

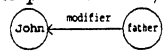
We think that the main difference between domain-dependent and domain-independent generation system is due to the problem of text structuring. Generally speaking, before building a domain-dependent text generation system, designers should make a domain model. In the model, the characteristics of the structure of text to be generated are described. For domain-dependent systems, text structure varies with application domains. In contrast with text structures of domain-dependent generation systems, text structures of domain-independent systems reflect the general features of English text. Our opinion is that in a generation system, instead of domain model, if a text structure model that represents the general features of English text is built, this system can be called domain-independent text generation system. In this paper, we outline our research in this field.

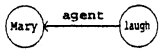
Using semantic network as the internal knowledge representation, we built a domain-independent English text generation system. Our research includes three parts: sentence-level generation, discourse-level generation and text-level generation. The rest of the paper is organized as follows. Section 2 introduces semantic network. Section 3 describes sentence-level generaion. Section 4 introduces discourse-level generation. Section 5 describes text-level generation, i.e., text structure model. At last, we draw conclusions and introduce future directions.

2 Semantic network

A semantic network of this study is a graph structure in which nodes represent concepts and links represent relations. In our English generation system, nodes are classified into two types

according to the word classes: entity node (e.g., noun) and event node (e.g., verb, adjective); links are classified into three types according to the grammatical relations:

1. Type I includes modifier link and predicative link. For example:
   (a) modifier link: the node which is modified by the other one is parent node; the node which modifies the other one is son node, e.g., John's father. 

2. Type II represents the relation between entity node and event node. Event node (e.g., verb) is parent node, entity node (e.g., noun) is son node. Type II consists of 9 kinds of links. For example:
   (a) agent link: e.g., Mary laughs. 

3. Type III is event-event link that represents the relation between two discourse segments. It is classified into 11 kinds. For example:
   (a) concession(-) link: e.g., I go to school though I have a headache.

Figure 1(a) shows how to represent "I go to school though I have a headache." by semantic network.

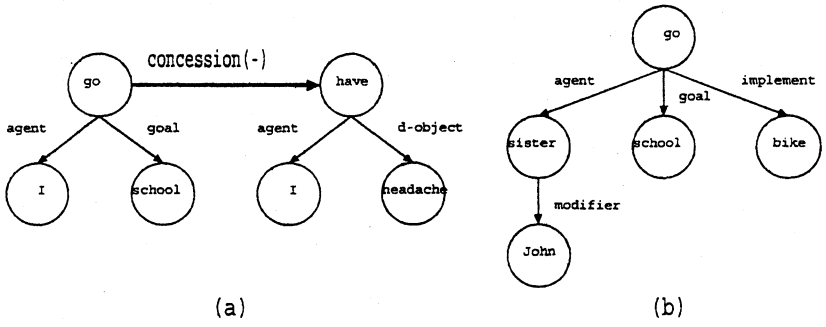


Figure 1: Representing sentences by semantic network

3 Sentence-level generation

On sentence level, we regard one single sentence as a tree whose leaf nodes are words and whose root is an event node. Figure 1(b) shows how to present a sentence ("John's sister goes to school by bike.") by semantic network. (Quirk et al., 1985) describes seven basic structures of English single sentence (p.721): SV, SVO, SVC, SVA, SVO_iO_d, SVOC, SVOA. According to these structures, we decided generation order of the links that construct single sentence: (agent link) > (i-object link) > (d-object link) > (complement link) > (source link) > (goal link) > (time link) > (implement link) > (place link). We defined a compare function to decide generation order of these links and used bottom-up algorithm to generate one single sentence.

4 Discourse-level generation

Generally speaking, cue phrases, such as "but", "because", "in contrast", are used to signal discourse relations between two discourse segments. Many papers discussed discourse structure and discourse generation, the most influential theory about discourse structure relations is RST

(Rhetorical Structure Theory) (Mann et al., 1988) which is widely applied to domain-dependent text generation systems. On the basis of RST, (Marcu, 1997) carried out an empirical study about rhetorical relations (discourse structure relations). He used cue phrases to determine rhetorical relations, and through corpus analysis of cue phrases, he calculated strength of each rhetorical relation. According to Marcu's results, we decided generation order of discourse structure relations. The order constraint is: (example link, detail link, summary link) > (condition link) > (concession(+) link, concession(-) link) > (contrast link) > (cause-result link) > (result-cause link) > (alternative link) > (sequence link). We use a compare function to realize it. Figure 2 shows a discourse strucutre whose generation result is "I take an examination today, so I go to school though I have a fever.".

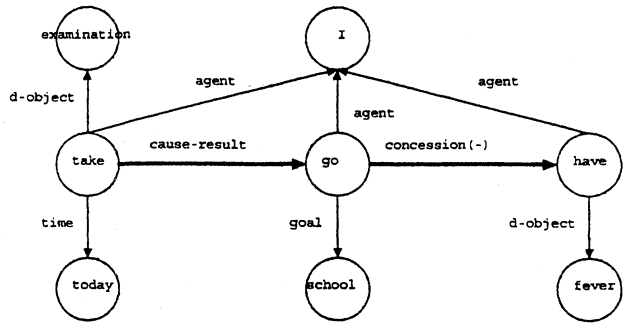


Figure 2: Example of discourse generation.

5 Text-level generation

In our study, we assume that a text consists of several units. One unit is a discourse or a single sentence. At text level, we built a text structure model to make the generation results coherent. In the model, we consider the following three aspects mainly:

1. The principle of END-FOCUS. (Quirk, et al, 1985)(p.1357) presents a basic characteristic of English text which is called the principle of END-FOCUS. That is, in a text, each unit ends with new information and begins with old information. The information value of "new information" is the highest, the information value of "old information" is the lowest. In English, it is common to process information in a unit "so as to achieve a linear presentation from low to high information value", i.e., the change of information value in English text has a regular pattern. We use three generation strategies to achieve this structure.

2. Attentional state of the readers. Text structure influences on attentional state of the readers. For example, the text "John went to his favorite music store to buy a piano. He had frequented the store .... He was excited that .... He ...." is coherenct and readable because it centers around a single individual "John" (Grosz et al., 1995). In our system, we use generation strategies to make generation results more readable.

3. How to generate pronouns. Pronoun generation is an important issue as well. We made two strategies to generate the text such as: "Iceland is very rich in natural heat. Earthquakes are frequent in it."

Figure 3 is an example of text generation. It represents "Tom lives in Canada. He manages three hotels. He likes fruits, so he eats apple everyday. Canada possesses numerous lakes. Many tourists enter it from USA. "

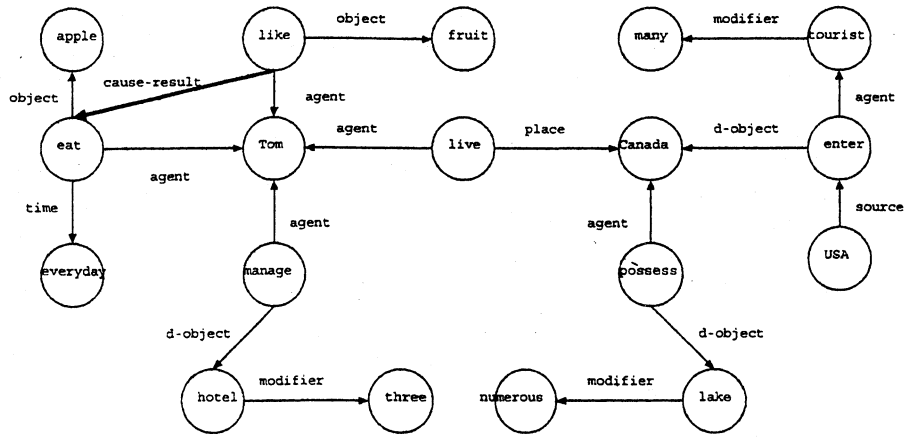


Figure 3: Example of text generation.

6 Conclusion and future directions

This paper puts forward a novel approach to building domain-independent text generation system. Our research includes three parts: sentence-level generation, discourse-level generation and text-level generation. We hope that this approach will be applied to other text generation systems. Future work will focus on enhancing the overall quality (coherence, flexibility, and readability) of the generation results.

Reference

Grosz, B., Joshi, A., and Weinstein, S. (1995). "Centering: A Framework for Modelling the Local Coherence of Discourse", *Computational Linguistics*, 21(2), 203-225.

Hovy, E. (1991). "Approaches to the planning of coherent text", *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pp.153-198, Kluwer.

Mann, C. W., and Thompson, S. A. (1988). "Rhetorical structure theory: toward a functional theory of text organization", *Text*, 3:243-281.

Marcu, D. (1997). *The rhetorical parsing summarization, and generation of natural language texts*, Ph.D thesis, University of Toronto.

Quirk, R., Greenbaum, S., Leech, G., and Scartvik, J. (1985). *A comprehensive Grammar of the English Language*, Longman Group Limited.

Reiter, E., Mellish, C., and Levine, J. (1995). "Automatic generation of technical documentation", *Applied Artificial Intelligence*, 9(3):259-287.

Reiter, E., Robertson, R., and Osman, L. (1999). "Types of knowledge required to personalise smoking cessation letters". *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making (AIMDM'99)*, pp.389-399, Springer-Verlag, Berlin.

Smith, M. H. (1995). *Natural Language Generation in the LOLITA System: A Engineering Approach*, Ph.D. thesis, Durham University.

Wanner, L., and Hovy, E. (1996). "The HealthDoc sentence planner", In *proc. of the International Workshop on Natural Language Generation*, pp.1-10.