Increasing Cohesion in Automatically Generated Natural Language Texts

Elisabeth A. Maier and Elena Not

Istituto per la Ricerca Scientifica e Tecnologica (IRST), 38050 Trento-Povo, Italy

Abstract. In this paper we show how the quality of machine-generated texts can be improved by augmenting a text planner with sophisticated components for the generation of textual phenomena. We discuss this point by introducing an example from our corpus of texts as so far generated by the schema-based planner RSE. We then describe the weaknesses of schematic approaches and show how a hybrid text planning architecture, which includes techniques to flexibly generate links between discourse units eliminates those disadvantages. Finally, we will show how the hybrid architecture adopted from [7] has been changed and extended in order to treat thematic progression and to generate referring expressions.

1 Introduction

The growing need for the quick availability of large amounts of documentation - ranging from machine manuals to weather reports - has led to an increasing importance of text generation. About 20 years ago research started in the area of sentence generation and over the years systems have been developed which cover the full repertoire of linguistic phenomena for various languages. Recently, interest moved towards issues related to the production of multiparagraph texts.

A whole family of text generators was based on a schema-oriented architecture ([15], [19], [14]). Schemata are abstract skeletons of the text to be generated that consist of sequences of predicates which are instantiated with knowledge from a data base.

Schematic text planners were soon found to be too inflexible for the generation of variant texts. Therefore, a number of researchers tried to formalize Rhetorical Structure Theory (RST, [13]) in order to use it as a tool for the planning of texts. RST builds on the assumption that the coherence of texts can be explained by means of relations which hold between segments of text. In RST about 20 such relations have been identified: among them are relations like SEQUENCE, which connects chronologically related events, ELABORATION, which occurs when one text segment provides more detail for another text unit, MOTIVATION, which links an action with the motivation for its execution, etc. In various implementations of the theory these relations have been operationalized

* We would like to thank Lee Fedder, Oliviero Stock and Carlo Strapparava for comments on a previous version of this paper.
as so-called plan operators (see, e.g., [16], [8]). For every plan operator, a set of application conditions and a specification of the effects of the corresponding relation exist. Both the condition and the effect parts of the plan operators address various types of knowledge: the use of a plan operator can, for instance, depend on the type of text to be generated, on the knowledge available in the knowledge base, on the preferences and the expertise of the potential reader, etc. This dependency on many different types of information may, however, introduce excessive complexity and intransparency in the plan operators, when they are extended for application in different domains. To solve this problem, a new architecture was developed ([7]) which separates out the various types of knowledge addressed in plan operators. The textual phenomena generated by this text planner were, however, restricted to the production of text structures.

We adopted this architecture as a starting point for the development of a system also capable of generating textual phenomena like thematic progression, a subclass of conjunctions and referring expressions.

2 The Problem

In this section we show how texts have been generated so far in the framework of the IRST-projects MAIA ([22]) and ALFRESCO ([21]), where the Rhetorical Schema Environment RSE has been used for text planning ([3]). We will point out why this approach is insufficient for the production of highly connected text.

Both ALFRESCO and the natural language component of MAIA, CONCIERGE, include intelligent multimedia interfaces that allow the user to access textual and visual information. In case the system chooses text as output modality the text planning component selects and organizes the content to be included in the response by “filling” one or more schemata with the relevant knowledge. Each schema is specified in a language designed for that specific purpose. It consists of two types of constructs: queries addressed to the knowledge base and operators for the construction of the text plan. The resulting text plan is specified in a format, which is an adaption of SPL, the Sentence Plan Language ([10]) which served as input specification for the PENMAN generation system ([9]).

In RSE new schemas can easily be written and executed at different levels of generality. Even though it can generate a broad variety of texts, RSE suffers from the deficiencies typical for schema-based text planners: they do not allow reasoning about the text plan produced, since they do not provide enough information about the discourse structure. In fact, schemas neither can give any motivation for their specific order nor do they have an explicit representation of the links between the different text units. All this hampers the introduction of textual factors that improve the quality of text. Although being perfectly coherent the texts lack linguistic cohesion. As an example, consider the following text produced by RSE, which is a description of the Natural Language Processing group of IRST.

---

2 English gloss: 1. The Natural Language group is part of the Interfaces department.