Summarising as a lever for studying large-scale discourse structure

Karen Sparck Jones
Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK
sparckjones@uk.ac.cam.cl

June 1993

We are using summarising as a way of studying large-scale discourse structure. Much computationally-oriented work on discourse structure has been concerned with dialogue, rather than with 'single-source' text. Some proposals have been made for single-source text e.g. Rhetorical Structure Theory (Mann and Thompson 1987), but are open to criticism (e.g. Moore and Pollack 1992); and single- source work has been primarily concerned with generation (e.g. McKeown 1985, Maybury 1991). We believe that large-scale discourse structure has a crucial part to play in summarising and therefore needs to be captured in the source text representation, for use in summarising, regardless of its contribution to source interpretation itself.

We have been engaged in a systematic examination of alternative types of large-scale text structure, designed to throw light on the kinds of information they make available for the text above the level of individual sentence representations, and how these can be used in summarising. Thus source text interpretation will provide a source representation capturing discourse structure over sentences, to be exploited in a condensing transformation through which the summary representation is formed, in turn leading to the output summary text.

This is a deliberately analytical investigation, taking a broad view without preconceptions. We distinguish three types of discourse information with structural implications: linguistic, domain, and communicative, and are seeing what large-scale text structures these respectively give. Thus we are investigating representation types categorised as dealing with information either about the linguistic properties of the source text (e.g. parallelism), or about its domain content (e.g. class membership), or about its communicative function (e.g. counterclaim). We are further, for any of these types, considering two alternative forms of structure that we have labelled 'bottom-up' and 'top-down' respectively. Bottom-up structures are individually created using general rules (e.g. by inference from domain facts); top-down structures are obtained by instantiating prior proformas (e.g. using domain frames). This is not a processing distinction, and the same formal structure (e.g. hierarchical) may result in either case; there may also be intermediate possibilities of the 'grammar' type. These distinctions of information type and representation form are broad ones that we are using as heuristics to explore discourse structure. Our aim is a comparative one, to see what each kind of approach leads to both for representation and for summarising. We can then consider how the structures relate to one another, whether as dependent, complementary, or reinforcing ones.

We are as far as possible using 'exemplar' approaches taken from previous research in the field, primarily in order to ground our work in what has been done so far: we are obliged in the current state of the art to work primarily through simulation, but we are trying to constrain the research by following approaches already proposed in the literature and preferably computationally investigated. Thus as an experimental strategy we are taking logical forms with resolved anaphors as a baseline representation for sentences, and then applying exemplar strategies of each type to these to obtain full representations of the source text. These full representations capture further relations across the sentences, embodying the large scale source text structure.

We have obtained alternative discourse structures and summaries for a set of short test texts. Some of the source structures are very simple, others more complex, importing significant additional information. So far, we have used the source representations in natural ways to obtain summaries: thus a linguistic-type source representation leads to a linguistically-motivated summary representation, in a way appropriate to the kind of the linguistic representation.

As linguistic structures we have so far provided analyses and derived summaries from the most simple approach, exploiting focus history to pick out key discourse entities, to more elaborate ones provided by RST (taking rhetorical relations as linguistic). These are bottom-up forms: rhetorical schemata might suggest a complementary top-down approach, but we could not readily analyse our texts as instantiations of these, and we therefore tried an intermediate 'story (or text) grammar' approach (cf Rumelhart 1975) To obtain domain-based structures we have used an extremely simple bottom-up approach using predication participation to identify discourse entities which figure largely in the source: we would like to try more sophisticated strategies where the baseline representation is enriched using general inference rules. We have applied scripts (and frames) as a top-down representation form (cf DeJong 1979; Tait 1983). Finally, for communicative structure we have used Grosz and Sidner (1986)'s approach to get intentional representations for our test texts. This constitutes a bottom-up approach: we have not yet identified an exemplar top-down one.

The results we have obtained have provided stimulating insights into the properties and roles of different types of text structure, and into the respective contributions they may make to summarising. For summarising, all the large-scale structures provide good leverage and help to identify source material which is intuitively important for use in the condensed summary, through selection or generalisation, though the alternative results for the same text may differ noticeably and individual results may be only semi-satisfactory. The results also illustrate the genuine role, but incomplete contribution, of each type of information.

Our deliberate separation of information types with their application strategies is thus allowing us to examine each type; to see how large-scale structure of any one kind is related to local structure, for instance through focus; and to formulate a view of a discourse model as a whole which subsumes distinct contributing models with their own necessary functions. Thus for example for one text, 'Biographies', there is a linguistic structure showing heavy presentational parallelism, a simple sequence of persuasive communicative intentions, and a separate domain object categorisation. There are complex relations between these, with reinforcing effects on the indication of key content. Our comparative analyses are thus providing the base (Grosz and Sparck Jones, in preparation), for the development of an account of discourse structure, or a discourse model, as a higher-level structure over subsidiary structures each with their own character and role.

P. Gladwin, S. Pulman and K. Sparck Jones 'Shallow processing and automatic summarising: a first study', TR 223, Computer Laboratory, University of Cambridge, 1991.

K. Sparck Jones 'Discourse modelling for automatic summarising', TR 290, Computer Laboratory, University of Cambridge, 1993, and in press.

K. Sparck Jones 'What might be in a summary?', Proceedings of the German Information Retrieval Conference, 1993, in press.