DEVELOPMENT, IMPLEMENTATION AND TESTING OF A DISCOURSE MODEL FOR NEWSPAPER TEXTS

Elizabeth D. Liddy [1], *Kenneth A. McVearry* [2], *Woojin Paik* [1], *Edmund Yu* [3], *Mary McKenna* [1]

[1] Syracuse University
School of Information Studies
Syracuse, NY 13244

[2] Coherent Research, Inc.
1 Adler Drive
East Syracuse, NY 13057

[3] Syracuse University
College of Engineering and Computer Science
Syracuse, NY 13244

ABSTRACT

Texts of a particular type evidence a discernible, predictable schema. These schemata can be delineated, and as such provide models of their respective text-types which are of use in automatically structuring texts. We have developed a Text Structurer module which recognizes text-level structure for use within a larger information retrieval system to delineate the discourse-level organization of each document's contents. This allows those document components which are more likely to contain the type of information suggested by the user's query to be selected for higher weighting. We chose newspaper text as the first text type to implement. Several iterations of manually coding a randomly chosen sample of newspaper articles enabled us to develop a newspaper text model. This process suggested that our intellectual decomposing of texts relied on six types of linguistic information, which were incorporated into the Text Structurer module. Evaluation of the results of the module led to a revision of the underlying text model and of the Text Structurer itself.

1. DISCOURSE-LEVEL TEXT MODELS

A discourse-level model of a text type can be likened to an interpretation model [Breuker & Wielinga, 1986] in that it specifies the necessary classes of knowledge to be identified in order to develop the skeletal conceptual structure for a class of entities. The establishment of text-type models derives from research in discourse linguistics which has shown that writers who repeatedly produce texts of a particular type are influenced by the schema of that text-type and, when writing, consider not only the specific content they wish to convey but also what the usual structure is for that type of text based on the purpose it is intended to serve [Jones, 1983]. As a result, one basic tenet of discourse linguistics is that texts of a particular type evidence the schema that exists in the minds of those who produce the texts. These schemata can be delineated, and as such provide models of their respective text-types which we suggest would be of use in automatically structuring texts.

The existence of and need for such predictable structures in texts is consistent with findings in cognitive psychology suggesting that human cognitive processes are facilitated by the ability to 'chunk' the vast amounts of information encountered in daily life into larger units of organized data [Rumelhart, 1977]. Schema theories posit that during chunking we recode individual units of perception into increasingly larger units, until we reach the level of a schema. Humans are thought to possess schema for a wide range of concepts, events, and situations [Rumelhart, 1980]. Discourse linguists have extended this notion to suggest that schema exist for text-types that participate regularly in the shared communication of a particular community of users.

What is delineated when a text schema is explicated is its discernible, predictable structure, referred to as the text's Superstructure. Superstructure is the text-level syntactic organization of semantic content; the global schematic structure; the recognizable template that is filled with different meaning in each particular example of that text-type [van Dijk, 1980]. Among the text-types for which schemas or models have been developed with varying degrees of detail are: folk-tales [Propp, 1958], newspaper articles [van Dijk, 1980], arguments [Cohen, 1987], historical journal articles [Tibbo, 1989], and editorials [Alvarado, 1990], empirical abstracts [Liddy, 1991], and theoretical abstracts [Francis & Liddy, 1991].

The goal of our current effort is to develop a component that can recognize text-level structure within a larger document detection system (DR-LINK) to enable the system to produce better retrieval results. For this system, we have focused our first efforts on newspaper texts, since the corpus we must process includes both the Wall Street Journal and the Associated Press Newswire.

2. DR-LINK

DR-LINK is a multi-stage document detection system being developed under the auspices of DARPA's TIPSTER Project. The purpose of TIPSTER is to significantly advance the state of the art in document detection and data extraction from large, real-world data collections. The document detection part of the project focuses on retrieving relevant documents from gigabyte-sized document collections, based on descriptions of users' information needs called topic statements. The data extraction part processes a much smaller set of documents, presumed to be relevant to a topic, in order to extract information which is used to fill a database.

The overall goal of DR-LINK is to simultaneously 1) focus the flow of texts through the system by selecting a subset of texts on the basis of subject content and then highlighting those sub-parts of a document which are likely spots of relevant text while 2) enriching the semantic representation of text content by: a) delineating each text's discourse-level structure; b) detecting relations among concepts; c) expanding lexical representation with semantically-related terms; and d) representing concepts and relations in Conceptual Graphs.

The purpose of the Text Structurer component in DR-LINK is to delineate the discourse-level organization of documents' contents so that processing at later stages can focus on those components where the type of information suggested in a query is most likely to be found. For example, in newspaper texts, opinions are likely to be found in EVALUATION components, basic facts of the news story are likely to be found in MAIN EVENT, and predictions are likely to be found in EXPECTATION. The Text Structurer produces an enriched representation of each document by decomposing it into smaller, conceptually labelled components. Operationally, DR-LINK evaluates each sentence in the input text, comparing it to the known characteristics of the prototypical sentence of each component of the text-type model, and then assigns a component label to the sentence.

In a form of processing parallel to the Text Structurer, the Topic Statement Processor evaluates each topic statement to determine if there is an indication that a particular text model component in the documents should be more highly weighted when matched with the topic statement representation. For example, topic statement indicator terms such as *predict* or *anticipate* or *proposed* reveal that the time frame of the event in question must be in the future in order for the document to be relevant. Therefore, documents in which this event is reported in a piece of text which has been marked by the Text Structurer as being EXPECTATION would be ranked more highly than those in which this event is reported in a different text model component

3. DEVELOPMENT OF THE NEWS SCHEMA MODEL

The need for a text model specifically for newspaper text is necessitated by the fact that the journalistic style forsakes the linear logic of storytelling and presents the various categories of information in a recurrent cyclical manner whereby categories and the topics contained within them are brought up, dropped, and then picked up again for further elaboration later in the news article. This internal topical disorganization makes a story grammar, as well as the expository text models [Britton & Black, 1985] not appropriate as text models.

Therefore, we took as a starting point, the uniquely journalistic, hierarchical newspaper text model proposed by van Dijk [1988]. With this as a preliminary model, several iterations of coding of a sample of 149 randomly chosen Wall Street Journal articles from 1987-1988 resulted in a revised News Schema which took from van Dijk's model the terminal node categories and organized them according to a more temporally oriented perspective, to support the computational task for which our model was to be used. We retained the segmentation of the overall structure into van Dijk's higher level categories, namely: Summary, Story and Comment, but added several terminal components as warranted by the data.

The News Schema Components which comprise the model are the categories of information which account for all the text in the sample of articles. The components are:

CIRCUMSTANCE - context in which main event occurs

CONSEQUENCE - definite causal result of main event

CREDENTIAL - credentials of author

DEFINITION - definition of special terminology

ERROR - mention of error that was made (in a correction)

EVALUATION - author's comments on events

EXPECTATION - likely or possible result of main event

HISTORY - non-recent past history of main event

LEAD - first sentence or paragraph which introduces or summarizes article

MAIN EVENT - text which advances the plot or main thread of the story

NO COMMENT - refusal or unavailability of source to comment

PREVIOUS EVENT - immediate past context for main event

REFERENCE - reference to related article (title and date)

VERBAL REACTION - quoted reaction from source to main event

While coding the sample, we developed both defining features and properties for each component. The defining features convey the role and purpose of that component within the News Schema while the properties provide suggestive clues for the recognition of that component in a news article. The manual coding suggested to us that we were in fact relying on six different types of linguistic information during our coding. The data which would provide these evidence sources was then analyzed statistically and translated into computationally recognizable text characteristics. Briefly defined, the six sources of evidence used in the original Text Structurer are:

Likelihood of Component Occurring - The unit of analysis for the first source of evidence is the sentence and is based on the observed frequency of each component in our coded sample set.

Order of Components - This source of evidence relies on the tendency of components to occur in a particular, relative order. For this source of evidence, we calculated across the coded files we had of each of the sample documents, looking not at the content of the individual documents, but the component label. We used this data to compute the frequency with which each component followed every other component and the frequency with which each component preceded every other component. The results are contained in two 19 by 19 matrices (one for probability of which component follows a given component and one for probability of which component precedes a given component). These two can be used in conjunction when there is a sentence lying between two other sentences which have already been coded for component or even when only the component of the preceding or following sentence is known. For example, if a series of sentences, *a-b-c*, the component label for sentence *b* is unknown, but the labels for sentence *a* and *c* are known, these matrices provide evidence of the likelihood that *b* might be any of the components in the model.

Lexical Clues - The third source of evidence is a set of one, two and three word phrases for each component. The set of lexical clues for each component was chosen based on observed frequencies and distributions. We were looking for words with sufficient occurrences, statistically skewed observed frequency of occurrence in a particular component, and semantic indication of the role or purpose of each component. For example, all the clues for VERBAL REACTION reveal the distinctly informal nature of quoted comments and the much more personal nature of this component when compared to the other components in a newspaper text.

Syntactic Sources - We make use of two types of syntactic evidence: 1) typical sentence length as measured in average number of words per sentence for each component; 2) individual part-of-speech distribution based on the output of the part-of-speech tagging of each document, using POST. This evidence helps to recognize those components which, because of their nature, tend to have a disproportionate number of their words be of a particular part of speech. For example, EVALUATION component sentences tend to have more adjectives than sentences in other components.

Tense Distribution - Some components, as might be expected by their name alone, tend to contain verbs of a particular tense more than verbs of other tenses. For example, DEFINITION sentences seldom contain past tense, whereas the predominate tense in HISTORY and PREVIOUS EVENT sentences is the past tense, based on POST tags.

Continuation Clues - The sixth and final source of evidence is based on the conjunctive relations suggested in Halliday and Hasan's Cohesion in English. The continuation clues are lexical clues which occur in a sentence-initial position and were observed in our coded sample data to predictably indicate either that the current sentence continues the same component as the prior sentence (e.g. *And* or *In addition*) or that there is a change in the component (e.g. *However* or *Thus*).

4. EMPIRICAL TESTING OF THE MODEL

The above computational method of instantiating a discourse-level model of the newspaper text-model has been incorporated in an operational system (DR-LINK). The original Text-Structurer evaluated each sentence of an input newspaper article against these six evidence sources for the purpose of assigning a text-level label to each sentence. This implementation uses the Dempster-Shafer Theory of Evidence Combination [Shafer, 1976] to coordinate information from some very complex matrices of statistical values for the various evidence sources which were

generated from the intellectual analysis of the sample of 149 Wall Street Journal articles.

Operationally, the text is processed a sentence at a time, and each source of evidence assigns a number between 0 and 1 to indicate the degree of support that evidence source provides to the belief that a sentence is of a particular news-text component. Then, a simple supporting function for each component is computed and the component with the greatest support is selected.

The Text Structurer was tested using five of the six evidence sources. (The algorithms for incorporating evidence from the continuation clues were not complete at the time of testing, so that evidence source was not added to the system.) We tested the Text Structurer on a set of 116 Wall Street Journal articles, consisting of over two thousand sentences.

The first run and evaluation of the original Text Structurer resulted in 72% of the sentences being correctly identified. A manual simulation of one small, heuristic adjustment was tested and improved the system's performance to 74% of the sentences correctly identified. A second manual adjustment for a smaller sample of sentences resulted in 80% correct identification of components for sentences.

5. ATTRIBUTE MODEL

After evaluating the preliminary results from the Text Structurer, we became dissatisified with some aspects of the model we developed and the processing based on that model. We needed a more precise way to define the components in the model, and we saw that frequently a sentence contained information for more than one component.

As a result, we developed a set of attributes of newspaper text in order to first better distinguish between similar components, and then to assign the attributes to text independent of component labels. These attributes are usually binary in nature. We identified eight attributes: Time, Tense, Importance, Attribution, Objectivity, Definiteness, Completion, and Causality.

For example, the Importance attribute has two possible values: "foreground" and "background". Components which are in the foreground include LEAD and MAIN EVENT; background components include CIRCUMSTANCE, DEFINITION, PREVIOUS EVENT, HISTORY, VERBAL REACTION, and NO COMMENT. The Objectivity attribute is also binary: its possible values are "objective" and "subjective". Objective components include CIRCUMSTANCE, MAIN EVENT, PREVIOUS EVENT, and HISTORY; subjective components include VERBAL REACTION, EVALUATION, and EXPECTAION. The Time attribute is multi-valued: its possible values are "past", "present", "past or present", and "future".

6. CURRENT MODEL

As a result of our analysis of text based on its attributes, we revised both the text-type model and the algorithms used by the Text Structurer. Revisions to the model focused primarily on subdividing components and adding new components to fill in gaps in the model and make it more precise. Revisions to the processing algorithms include: 1) restricting the sources of evidence used to lexical clues only; 2) establishing an order of precedence for components; 3) moving from a single lexicon to a lexicon for each component; 4) discontinuing the use of the Dempster-Shafer method of evidence combination; 5) moving the level of analysis from the sentence to the clause level.

The new components:

CIRCUMSTANCE-STOCK - closing price of stock mentioned in the article

CONSEQUENCE-PAST/PRESENT - past or present causal result of main event

CONSEQUENCE-FUTURE - future causal result of main event

EVALUATION - opinion attributed to a source

EVALUATION-JOURNALIST - opinion not attributed to a source

EXPECTATION-JOURNALIST - likely or possible result of main event not attributed to a source

FIGURE DESCRIPTION - text which describes a nearby figure, table, etc.

LEAD-ATTENTION - attention-getting lead (does not summarize)

LEAD-FUTURE - lead which refers to the future

LEAD-HISTORY - lead which refers to the non-recent past

LEAD-PREVIOUS - lead which refers to the recent past

MAIN-EXAMPLE - specific instance or example of main event

MAIN-FUTURE - main event set in the future

MAIN2 - alternate main event (new story)

PAST - undated past context of main event

7. FUTURE WORK

There are several areas we would like to explore, both in improving the operation of the Text Structurer and in demonstrating its applicability. One obvious way to improve the accuracy and coverage of the Text Structurer is to expand the lexicons for each component, via corpus-guided acquisition of synonyms. Another possibility is that ordering and continuation evidence can in fact be used to augment lexical evidence, e.g. for sentences which should be labeled HISTORY and which follow a HISTORY lexical clue but which themselves do not contain any HISTORY clues. One area which needs improvement is distinguishing between foreground and background components, e.g. MAIN EVENT vs. CIRCUMSTANCE. It is clear that purely lexical information is not sufficient to make the distinction, and that patterns of verbs and other words, ordering, and other information are required, if not some internal understanding of the subject of the text.

There are several possible uses of the Text Structurer module in a document detection system. Within DR-LINK, it can be used as a focusing mechanism (filter or weighting) for other modules, e. g. the Relation-Concept Detector, which identifies concepts and relations between concepts in text. For example, the Relation-Concept Detector can be set to emphasize those sentences which are labeled with a foreground component (LEAD, MAIN EVENT, etc.) by the Text Structurer. Another application outside of DR-LINK is as an intermeditate process between document detection and data extraction. Once a document is determined to be relevant, the Text Structurer can focus the data extraction process on those sentences or sentence fragments which are most likely to contain the information required to fill the database.

8. CONCLUSIONS

Although we are clearly in the early stages of development of the Text Structurer, we find these results quite promising and are eager to share our empirical results and experiences in creating an operational system with other computational linguists. To our knowledge, no similar, operational discourse structuring system has yet been reported in the literature.

We have applied the newspaper text-type model to text from a different source, by coding a sample of AP Newswire articles. This effort verified that the model was general enough to handle news text from various sources; in fact, a subset of the model covered all cases seen in the AP text.

We are in the process of evaluating the latest version of the Text Structurer based on the current newspaper text model. We will next apply a similar methodology to the development of a model and processing component for automatically structuring full-length, technical journal articles.

REFERENCES

1. Alvarado, S. J. (1990). Understanding editorial text: A computer model of argument comprehension. Boston, MA: Kluwer Academic Publishers.
2. Breuker & Wielinga. (1986). Models of expertise. ECAI.
3. Britton, B. & Black, J. (1985). "Understanding expository text: From structure to process and world knowledge." In B. Britton & J. Black (Eds.), Understanding expository texts: A theoretical and practical handbook for analyzing explanatory text. (pp. 1-9). Hillsdale, NJ: Lawrence Erlbaum Associates.
4. Cohen, R. (1987). "Analyzing the structure of argumentative discourse." Computational Linguistics. 13, pp. 11-24.
5. Francis, H. & Liddy, E. D. (1991). "Structured representation of theoretical abstracts: Implications for user interface design." In Dillon, M. (Ed.), Interfaces for information retrieval and online systems. NY: Greenwood Press.
6. Halliday, M. A. K. & Hasan, R. (1976). Cohesion in English. London, Longmans.
7. Jones, L. B. (1983). Pragmatic aspects of English text structure. Arlington, TX: Summer Institute of Linguistics.
8. Liddy, E. D. (1991). "The discourse-level structure of empirical abstracts: An exploratory study." Information Processing & Management. (pp. 55-81).
9. Meteer, M., Schwartz, R. & Weischeidel, R. (1991). "POST: Using probabilities in language processing." Proceedings of the Twelfth International Conference on Artificial Intelligence. Sydney, Australia.
10. Propp, V. (1958). Morphology of the folk-tale. (L. Scott, Trans.). Bloomington: Indiana University Press. (Original work published 1919).
11. Rumelhart, D. (1977). "Understanding and summarizing brief stories." In D. LaBerge & S. J. Samuels (Eds.), Basic processes in reading: Perception and comprehension (pp. 265-303). Hillsdale, NJ: Lawrence Earlbaum Associates.
12. Rumelhart, D. (1980). "Schemata: the building blocks of cognition." In R. Spiro, B. Bruce, & W. Brewer (Eds.), Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education (pp. 33-58). Hillsdale, NJ: Lawrence Earlbaum Associates.

13. Shafer, G. (1976). A mathematical theory of evidence. Princeton, NJ: Princeton University Press.

14. Tibbo, H. R. (1989). Abstracts, online searching, and the humanities: An analysis of the structure and content of abstracts of historical discourse. Ph.D. Dissertation, College of Library and information Science.

15. van Dijk, T. A. (1980). Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition. Hillsdale, NJ: Lawrence Earlbaum Associates.

16. van Dijk, T. A. (1988). News analysis: Case studies of international and national news in the press. Hillsdale, NJ: Lawrence Earlbaum Associates.