Development and Implementation of a Discourse Model for Newspaper Texts

Elizabeth D. Liddy, Woojin Paik, Mary McKenna
School of Information Studies, Syracuse University
Syracuse, New York, 13244-4100
{liddy, wjpaik, memckenn}@mailbox.syr.edu

Introduction

In this paper, we will focus on the development, implementation, and evolution of a discourse model which is used to computationally instantiate a discourse structure in individual texts. This discourse model was developed for use in a Text Structuring module that recognizes discourse-level structure within a large-scale information retrieval system, DR-LINK (Liddy & Myaeng, 1993). The Text Structurer produces an enriched representation of each document by computationally decomposing it into smaller, conceptually labelled components. This delineation of the discourse-level organization of each document's contents facilitates retrieval of those documents which convey the appropriate discourse semantics that are responsive to the user's query.

The recognition of the existence 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, texts of a particular type evidence the schema that exists in the minds of those who produce the texts. These schema can be delineated, and as such provide models of their respective text-types which are of use in automatically structuring texts. A text schema explicates a discernible, predictable structure, the global schematic structure 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 are: folk-tales (Propp, 1958), newspaper articles (van Dijk, 1980), arguments (Cohen, 1987), historical journal articles (Tibbo, 1989), editorials (Alvarado, 1990), empirical abstracts (Liddy, 1991), and theoretical abstracts (Francis & Liddy, 1991).

Development of the News Text Schema

Our first effort towards including discourse-level semantics in the DR-LINK System were focused on newspaper texts, taking as a starting point, the hierarchical newspaper text model proposed by van Dijk (1988). Several iterations of human analysis and coding of several hundred randomly selected Wall Street Journal articles using the components from van Dijk's model motivated us to develop a revised News Schema which re-organized van Dijk's categories according to a more temporally-oriented perspective and added several new components. The resulting News Schema Components were: CIRCUMSTANCE, CONSEQUENCE, CREDENTIALS, DEFINITION, ERROR, EVALUATION, EXPECTATION, HISTORY, LEAD, MAIN EVENT, NO COMMENT, PREVIOUS EVENT, REFERENCES, and VERBAL REACTION.

The process of manually coding the training sample also served to suggest to us the different types of linguistic information which we implicitly relied on during our intellectual decomposing of texts. These intuitions were further explored by means of statistical analyses of the linguistic differences exhibited by text in the various components. These results were translated into computationally recognizable text characteristics for use by the Text Structurer to assign a single component label to each sentence. Briefly defined, the sources of evidence used in the Text Structurer were:

Lexical Clues - A set of one, two and three word phrases for each component, based on observed frequencies and distributions. Clues are words with sufficient occurrences, and a statistically skewed observed frequency of occurrence in a particular component. Not surprisingly, many of clues strongly

suggest the semantic role or purpose of each component.

Order of Components - The tendency of components to occur in a particular, relative order determined by calculating across the coded training files.

Likelihood of Component Occurring - The observed frequency of each component in our coded sample set.

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.

Syntactic Sources - 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, a part-of-speech tagger (Meteer et al, 1991).

Continuation Clues - Based on the conjunctive relations suggested in Halliday and Hasan's Cohesion Theory (1976), lexical clues which occur in a sentence-initial, or near sentence-initial position, and which were observed in our coded sample data to predictably indicate either that the current sentence continues the same component as the prior sentence or that there is a change in the component.

These sources of evidence for instantiating a discourse-level model of the newspaper text-type were incorporated in the computational Text Structurer in our system, which evaluates each sentence of an input newspaper article against these evidence sources, comparing it to the known characteristics of each component of the text-type model, for the purpose of assigning a text-level label to each sentence.

The computational implementation of the Text Structurer used the Dempster-Shafer Theory of Evidence Combination (Shafer, 1976) to coordinate information from the various evidence sources. In the implementation, each document is processed a sentence at a time, and each source of evidence assigns a value between 0 and 1 to indicate the degree of support that each evidence source provides to the belief that a sentence is functioning as a particular news-text component. The probability of each observed value for each piece of evidence for each component is calculated and is used as a belief in the Dempster-Shafer algorithm for evidence combination. Then, a simple supporting function for each component is computed and the component with the highest assigned belief value is selected as the correct component tag for that sentence.

The Text Structurer was tested on one hundred sixteen WSJ articles comprising several thousand sentences. This first testing resulted in 72% of the sentences being correctly identified. A second run of a smaller sample resulted in 80% correct identification of components for sentences. Ongoing efforts have improved the quality of the evidence sources used by the Text Structurer and promise to enhance these results significantly.

Attribute Model of News Text

After completing the first implementation of this model, we moved to a new, attribute model of news-text structure. One factor which precipitated this move was the difficulty we encountered in manually coding some new training data. These difficulties appeared to be caused by our increased awareness of the multiple attributes or dimensions embedded in each of the component labels. For example, we realized that PREVIOUS EVENT was defined by a combination of particular values on the dimensions of Importance, Time, Completion, and Definiteness. Although each of the individual dimension values was shared by other components, PREVIOUS EVENT was a unique combination of dimension values. That is, although several components shared the same values on some dimensions, each component was distinguished from all other components by its value on at least one dimension. We felt that the more holistic tagging of sentences with component labels such as PREVIOUS EVENT, CIRCUMSTANCE, and LEAD had not adequately reflected these micro-level similarities and distinctions.

In addition, questions from members of the potential community of users of the structured output - questions such as: "Do the component labels indicate the status (e.g. journalist vs. participant in the news) of the views in the text?" or "Do the component labels indicate whether an event is ongoing or completed?" made us realize that the granularity of the components in the original model did not explicitly indicate these facts, although they were implicit in the components' definitions which we had developed and relied on for manual coding. Therefore, we concluded that there was a dual need to: 1) capture and represent the basis of the commonality amongst some components, as well as;

2) make more distinct the uniquenesses which distinguished components.

In an attempt to accomplish these goals, we developed the Attribute Model of the news text in which pieces of text are evaluated for their specific value on each of eight dimensions or attributes: time of event, tense, importance, attribution, objectivity, definiteness, completion, and causality. Plus or minus values on these attributes were assigned to the text pieces without consideration of the component labels from the earlier model. At this point, we also began coding text at the clause rather than the sentence level, since we recognized that single sentences do contain multiple discourse-level components. These might be reflected in tense changes within a single sentence, or appositional statements of past events within a straight-forward reporting of a current news event.

After reviewing our recoding of the sample texts, we realized that the move to the Attribute Model had resulted in the loss of a very important function which had been performed by the earlier discourse-component labelling of sentences using the News Text Schema. That is, the recoded data seemed to prove the old adage that the whole is sometimes greater than the sum of its parts - that is, labelling a segment of text PREVIOUS EVENT had conveyed more than simply identifying that text segment's values on the eight dimensions. In other words, the discourse-component label conveys the **role** or **function** within the larger news-text model, information that is not conveyed by the Attribute Model coding. That is, discourse-level structured news articles based on the News-Text Schema convey a great deal of significant linguistic and pragmatic information that is not available without this discourse level analysis and processing.

Revised News Text Model

Although we recognized the superiority of the earlier News Text Schema over the newer Attribute Model, we did not want to lose the distinctions and similarities amongst text segments which we were able to recognize when using the eight dimensions of the Attribute Model. Therefore, we moved to a revision of our original News Text Schema, a refinement of the earlier components via addition of some of these distinguishing attributes to the earlier components. Operationally this was accomplished via the addition of sub-codes. For example, LEAD was sub-coded for its temporal aspect via the codes HISTORY, PREVIOUS, and FUTURE; CONSEQUENCE was sub-coded for PAST, PRESENT and FUTURE; EVALUATION was sub-coded for JOURNALIST to distinguish opinion which is not attributed to a source and therefore likely to be the journalist's view from plain EVALUATION, which is an opinion attributed to a named source. In addition, PREVIOUS EVENT and HISTORY had sub-codings added for CONTINUOUS, and MAIN was sub-coded for FUTURE, as well as SECOND EVENT, and EXAMPLE.

Given this more complex News Text Schema, the original Text Structurer implementation which made use of eight sources of linguistic evidence did not appear reasonable for processing gigabytes of text for our DR-LINK Project. Based on an analysis of the automatically-structured documents produced by the first implementation, we measured how much each evidence source contributed to the system's ability to assign correct components. From that analysis, we determined that the more important evidence sources were lexical clues, tense data, and continuation clues. Therefore we reduced the number of evidence sources to these three. These evidence sources were evaluated heuristically by a combination of rules and lexicon. Frequency of occurrence of each component, sentence length, and distribution of parts of speech were dropped as evidence sources. Ordering information, which was ineffectively implemented as an evidence source in the first implementation, is currently being re-incorporated, as is a return to the Dempster-Shafer approach to evaluating and combining evidence.

The development of a leaner implementation in which only those evidence sources which contributed most significantly to the system's ability to correctly recognize pieces of text as particular components was used. The new implementation of the revised News-Text Schema instantiated a more precise model both in terms of the specificity of the model's components and the unit of text assigned a discourse component.

Conclusion

The process of developing, implementing and iteratively revising a discourse model for one text-type for use in the computational recognition of discourse-level structure in text is not yet finished. We have empirical results which indicate the News Text Schema's positive contribution to a text retrieval application by enabling DR-LINK to recognize documents which are relevant to a query on the basis of their discourse-level semantics as captured by the News Text Schema. We are currently engaged in efforts to both improve the current

implementation as well as efforts to generalize the model.

Acknowledgements

The research was supported by ARPA's TIPSTER Initiative.

References

Alvarado, S. J. (1990). Understanding editorial text: A computer model of argument comprehension. Kluwer Publishers.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. Computational Linguistics, 13, pp. 11-24.

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. Greenwood Press.

Halliday, M. A. K. & Hasan, R. (1976). Cohesion in English. London, Longmans.

Jones, L. B. (1983). Pragmatic aspects of English text structure. Arlington, TX: Summer Institute of Linguistics.

Liddy, E. D. (1991). The discourse-level structure of empirical abstracts: An exploratory study. Information processing and management, 27:1, pp. 55-81.

Liddy. E.D. & Myaeng, S. H. (1993). DR-LINK: A system update for TREC-2. In Harman, D., (Ed.), Proceedings of the second Text Retrieval Conference.

Meteer, M., Schwartz, R. & Weischedel, R. (1991). POST: Using probabilities in language processing. Proceedings of the Twelfth International Conference on Artificial Intelligence. Sydney, Australia.

Propp, V. (1958). Morphology of the folk-tale. (L. Scott, trans.). Bloomington, Indian University Press. (Original work published 1919).

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

Tibbo, H. R. (1989). Abstracts, online searching, and the humanities: An analysis of the structure and content of abstracts of historical discourse. PhD Dissertation, College of Librar nd Information Science.

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

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