Project EPISTLE: A system for the Automatic Analysis of Business Correspondence

Lance A. Miller

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT: The developing system described here is planned to provide the business executive with useful applications for the computer processing of correspondence in the office environment. Applications will include the synopsis and abstraction of incoming mail and a variety of critiques of newly-generated letters, all based upon the capability of understanding the natural language text at least to a level corresponding to customary business communication. Successive sections of the paper describe the Background and Prior Work, the planned System Output, and Implementation.

I. BACKGROUND AND PRIOR WORK

Based on previous work in natural language processing [1,2], the long-term objective of the EPISTLE project is to develop general-purpose system algorithms for implementing highly practical applications requiring some level of understanding of natural language texts (the acronym derives from "Executive/Principal's Intelligent System for Text and Linguistic Endeavors"). Ultimately, we intend that these algorithms could apply to any texts as long as the pragmatics controlling their interpretation can be objectively represented. Initially, however, we are focusing on business correspondence, and much of our preliminary work has been to establish that business letters, rather than being highly idiosyncratic, do indeed serve a set of knowable purposes and that people pretty much agree on what these are. To this end we collected a database of 400 letters, from which we selected a sample of 20 believed to represent the range of purposes, styles, contents, and tones of the database. We then asked experienced writers to rate letters on these features. The primary findings of the experiment [3] were: (1) letter features were identified by subjects with high inter-personal agreement; (2) the majority of purposes detected in the letters were accounted for by a small number of categories; (3) the ordering of text segments fulfilling ordered purposes conformed to a few frequent patterns; and (4) the subjective judgements could often be predicted quite well from objective word frequency statistics.

We conclude from these behavioral findings that there are indeed extensive regularities in the characteristics of business letters, determined primarily by the purpose objectives. It is these constraints that most strongly indicate to us the feasibility of developing automatic means for recognizing content-themes and purposes from the letter text (as well as the converse, generating letter text from information about purposes).

Other analyses have been undertaken to estimate the linguistic complexity and regularities of the texts. The average letter appears to contain 8 sentences, with an average of 18 words each; in the 400 letter-bodies there are roughly 57,000 words and 4500 unique words total. An ongoing hand analysis of the syntactic structure of sentences in a 50-letter sample reveals a relatively high frequency of subject-verb inversions (about 1 per letter) and complex lengthy complementizers (1-4 per letter). These features, along with very frequent noun phrase and sentence coordination, accompanied by a wide variety of grammatical but unsystematic structure deletions, indicate an exceptionally high level of grammatical complexity of our texts. With respect to overall text syntax we have analyzed 10 letters for text cohesion, using a modification of Halliday and Hasan's coding scheme [4]; 82 percent of the instances of cohesion detected were accounted for by 4 categories: lexical repetitions (29%), pronouns (28%), nominal substitutions (9%, e.g., "one", "same"), and lexical collocations (words related via their semantics, 16%). In an extension of this discourse structure analysis we are analyzing 50 letters, coding all occurrences of functional nouns in terms of (1) the grammatical case function served and (2) the cohesive relation to prior nouns. Preliminary results indicate consistent patterns of case-shift and type of cohesion as a function of the pragmatic and content themes. The results of these linguistic analyses will help determine the strategy ultimately adopted for selecting surface parses and meaning interpretations.

II. SYSTEM OUTPUT

The planned system will provide the following for each letter in our database: (1) surface syntactic parses for each sentence; (2) meaning interpretations for each sentence, adjusted to the context of prior sentences; (3) a condensed synopsis of the overall meaning content of the letter; (4) a critique of each letter's spelling, punctuation, and grammaticality; (5) a mapping of the meaning content onto common business communication themes; and (6) some characterization of the author's style and tone. In addition to the above, we plan to develop a limited facility to generate short letters of a certain type (e.g., information requests) conforming to a particular author's normal "style" and "tone".

III. IMPLEMENTATION COMPONENTS

A. Semantic Representations: Many of the lexical items in our texts appear to have two or more literal word-sense usages (as well as occasional non-literal ones); it also appears that the discriminating semantic features among highly-related lexical items cannot be ignored if the intended letter nuances are to be preserved. We therefore do not expect much reduction in cardinality when mapping from the lexical to the concept space; we also anticipate that our representations will have to be unusually rich, in terms of both a large number of features distinguishing the concepts underlying lexical items and the capability to relate different concepts together. Among the most important anticipated semantic features are those describing the preconditions and conse-

quences of ACTIONS and those characterizing the internal states of ACTORS (e.g., their intentions, expectations, and reactions).

B. Parsing:

We will employ a system called NLP as the basic "operating system" for our application development. This system has been used for other understanding projects and provides all of the general components required, including a word-stem dictionary, a parser, knowledge representation, and natural language generation [2,5]. In particular, the parser proceeds left to right, character by character, in processing a sentence, generating all possible descriptions of text segments in a bottom-up fashion by application of rules from an augmented phrase structure grammar -- essentially a set of context-free rules augmented with arbitrary conditions and structure-building actions. In writing the grammar we are attempting to keep to an absolute minimum the use of semantic information, to increase the applicability of the parser over a variety of semantic domains. Our general strategy to minimize the number of surface parses (and increase parsing efficiency) is to attach lower syntactic constituents (e.g., post-nominal prepositional phrases) to the highest-possible unit (e.g., to the verb-phrase rather than the noun-phrase), with the final decision as to the most appropriate attachment to be resolved by the interpretation components.

C. Meaning-Assignment:

Our planned strategy for choosing the meaning to assign to a sentence is basically to find that action-concept whose case-relations are most completely "satisfied" by all of the concepts implied by the sentence. In the expected frequent case of only partial "fits" to several action-concepts, preference among these will be based on several factors, including: (1) the number of case relations of a concept "filled" or "unfilled" by elements of the present (or prior) text, and the relative importance in the intensional definition of each of these; (2) the "directness" of the mappings of text segments to underlying concepts; and (3) the syntactic structure of the sentence (e.g., syntactically "higher" components usually will be preferred to "lower" ones).

D. Text-Interpretation:

We propose to keep separate lists of each action-concept and entity-concept encountered in the text. Following the meaning-assignment to a sentence, the sentence will be re-examined to determine if it supplies qualification information for any prior-mentioned action or entity; if so, these separate representations will be so augmented, a process called "updating". Statistics of each such updating of information will be kept for each sentence for subsequent characterization of style. Next, these separate entity/action representations will be examined directly to determine whether they can be combined as elements of some broader concept. By this process we will therefore be able to update and condense our representations as we go along, facilitating eventual synopsis and abstraction of content-themes.

In addition to the above semantic interpretations for the complete text, we will also build up a composite representation of the syntactic structure of the text. We, first, are hopeful of being able to discover a relatively small number of schema for characterizing syntactic structures within sentences; we then believe that the majority of letter text can be accounted for in terms of frequently occurring patterns of these schemas.

E. Adaptation:

As a unique feature, we plan to implement the capability to dynamically modify or adapt our system so as to change the manner in which word-senses are selected or meanings assigned as a function of the system's experience with various kinds of texts. This would be accomplished by assigning "weight" attributes to each lexical item and to each underlying concept (and its attribute-values); the weight-values along the path finally selected for mapping text to concepts would then be incremented by some amount. Given that preference ordering of text to concept paths is determined by such overall path-weights, the system could thus achieve self-adaptation to word-usages of particular application domains. This facility could also be employed to characterize individual authors' styles.

F. Abstraction and Critique of Letters:

Concerning Content-Themes and Purposes, we plan to map the system's meaning interpretations onto a set of common business content-themes and communication purposes, and we are presently conducting behavioral and analytical studies to determine these. With respect to Grammaticality, we anticipate being able to detect incomplete sentences, subject-verb disagreements, and inappropriate shifts in verb tenses; in addition, we will be able to identify ambiguities and some instances of clearly "awkward" syntax. Spelling errors of the "non-word" type are easily caught, and certain spelling errors in which the misspelled word is in the dictionary may also be caught if they contain sufficient syntactic information. In addition, some fraction of "spelling" errors involving semantically inappropriate words should be detectable. Finally, we may be able to discover a number of Punctuation errors.

The last aspect of critiques is that of style and tone. We are aware of the several "indices" for measuring various aspects of these but consider them to be at best very crude indicators [6]. As a starting point we have identified five dimensions for each concept, and we will implement the capability to assess texts on these dimensions until we are better informed. For Style, defined as "the organizational strategy for conveying content", the dimensions are: sentence precision, sentence readability, reference clarity, information-value, and cohesion. Tone, defined as "the connotations of interpersonal attitudes", is to be rated on the dimensions of: personal-ness, positive-ness, informal-ness, concrete-ness, and strength. We plan to output and highlight those text segments which fall below a certain level of acceptability on these measures.

REFERENCES

[1] Miller, L. A. "Behavioral studies of the programming process." IBM Research Report RC 7367, 1978.

[2] Heidorn, G. E. "Augmented phrase structure grammars". In Nash-Webber, B.L. and Schank, R. C. (Eds.), Theoretical Issues in Natural Language Processing. Association for Computational Linguistics, June, 1975.

[3] Miller, L. A. and Daiute, C. "A taxonomic analysis of business letters". IBM Research Report, in preparation, 1980.

[4] Halliday, M. A. K. and Hasan, R. Cohesion in English. London: Longman Group Ltd., 1976.

[5] Heidorn, G. E. "Automatic programming through natural language dialogue: A survey". IBM Journal of Research and Development, 1976, 20 302-313.

[6] Dyer, F. C. Executive's Guide to Effective Speaking and Writing. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1962.