An experiment in computational discrimination of English word senses

by Ezra Black

A number of researchers in text processing have independently observed that people can consistently determine in which of several given senses a word is being used in text, simply by examining the half dozen or so words just before and just after the word in focus. The question arises whether the same task can be accomplished by mechanical means. Experimental results are presented which suggest an affirmative answer to this query. Three separate methods of discriminating English word senses are compared information-theoretically. Findings include a strong indication of the power of domain-specific content analysis of text, as opposed to domain-general approaches.

1. Introduction

It is difficult to suggest a branch of natural-language processing which would fail to benefit from a procedure for identifying the senses of the words used in text. To take a single example, researchers in the field of speech recognition need information concerning the word sequence already recorded at a given point in time, so that they may accurately predict the word or words about to be received. For instance, consider the word *will*. While in its modal use it is one of the most frequent items in many sorts of text (cf., e.g., Francis and Kučera [1]), there are certainly a large number of text domains in which its appearance as, say, the *will* of *last will and testament* narrows the range of its likely right neighbors by impressive amounts, as against the uncategorized occurrence of the string *will*. Generalizing, it seems fair to assume that predictions of right context will differ appreciably depending upon the sense(s) of the predictor word(s).

This paper gives an account of experimentation, described more fully in Black [2], designed to compare and evaluate three possible methods of computational word-sense identification. As can be seen from the literature review in this section, the procedures chosen for contrast partially reflect current approaches to the problem. Section 2 presents the experiment itself, while Section 3 discusses it and suggests future research directions.

We now briefly survey computational sense research for which either a large-scale implementation or the materials for such development are in existence.

Débili comes to the task of word-sense discrimination armed with a listing, obtained computationally (see [3]), of word pairs observed to have entered into certain dependency relations in previously analyzed text. For instance, having processed the sentence, *Le bail a expiré* ('The lease has expired'), Débili's procedure "knows" that *bail* ('lease') and *expiré* ('expired') can occur in the subject/main-verb dependency relation. Other such relations include noun/adjectival modifier and verb/direct object. Simplifying

somewhat, a "validity" score of 1 is given to word pairs which are on this list, and a score of 0 is given to pairs which are not. Given this database, the disambiguation routine adopted is, briefly, as follows: Assume that a given word has three senses. Suppose the word is the French *expire*. Then an enumeration is made of all the synonyms of *expire* when the word has sense 1, and the same is done for senses 2 and 3. Call these enumerations "word families." Now suppose it is to be determined which sense of *expire* occurs in the new sentence, *Le bail expire à la fin du mois* ('The lease expires at the end of the month'), where *expire* can be either *expire*(FINIR) ('ends'), *expire*(MOURIR) ('dies'), or *expire*(RESPIRER) ('breathes out'). A preprocessor identifies the word pair *bail/expire* as standing in the subject/main-verb relation. The maximum is now calculated of the Cartesian products of the validity scores of the word families of the uncertain or polysemous word *expire* and of the *ex hypothesi* unambiguous word *bail*, and the sense is chosen whose word family yields this maximum. Hence, a sense is most likely to be selected as correct if it or its near semantic relatives have occurred in previously examined text in the exact dependency relation under scrutiny.

If for Débili, "sense disambiguation is morphology," for Gross and his associates, "sense disambiguation is syntax." Gross [4] presents an interesting and provocative orientation toward sense discrimination in English (and, by implication, in French and other Romance languages as well). Each word of Gross's lexicon corresponds to a line of a lexicon-grammar, which is a two-dimensional matrix, apparently on the order of 32 000 × 400 [4, 5] for French at present, and of uncertain size for English. Columns of the lexicon-grammar are labeled with possible syntactic properties of an entry word. Most of these consist of intraclausal syntactic environments in which the candidate word can or cannot appear in a given slot; these environments can be sequences of syntactic categories to the right and/or left of the word, or they can be sequences of lexical items to the right and/or left of the word, or, finally, they can be combinations of the latter two possibilities. In addition, certain properties of one or more nouns involved in the environment—properties which some linguists have called syntactic, others semantic—such as "concrete," "animate," and "human," are labels of still other columns. Each cell is marked with a + or a − depending on whether the team of linguists conducting the research finds the clause represented by the cell to be acceptable or unacceptable. (For a description of Gross's methods of linguistic team investigation, see [5].) By inspecting the syntactic context of a given word token, candidate senses are eliminated whose conditions for usage, as catalogued in the lexicon-grammar, are not met.

The problem confronted by Kelly and Stone [6] arose within the context of the operation of a suite of programs previously developed by Stone and his associates, whose purpose was to pass through a text, assigning to each appropriate word a label representing a particular content-analytic category. (For presentations of content analysis, see Budd et al. [7] and Rosengren [8].) They found early on in the application of the system that its utility was sharply reduced by its lack of ability to assign parts of speech and word senses [6, p. 1]. What they therefore did was to construct a preprocessing program to perform such assignments. (Part-of-speech labeling by computer is a problem that is completely distinct from sense identification—one that is well in hand at present. It has been performed, for instance, by the Continuous Speech Recognition Group of IBM's Thomas J. Watson Research Center, using the Viterbi algorithm (on which see, e.g., Forney [9]), and achieving a >98% accuracy of prediction for 29 syntactic categories. A different approach to part-of-speech labeling, also quite successful and implemented over a much larger set of categories, on the order of 200, is being taken by G. Leech and his associates at the University of Lancaster, England [10–13]. There are quite a number of other programs in existence which perform this task with a fairly high level of accuracy.)

Drawing from a corpus of about six million tokens within the domain of interest to the content analysts they wished to serve—namely behavioral science—they selected a sample of 510 976 tokens [6, p. 5]. They then restricted their efforts to the part-of-speech labeling and/or sense disambiguation of 1815 types.

Approximately eight individuals—undergraduates, graduate students, Kelly, and Stone—together worked from a KWIC concordance of the items selected, over the 510 976-token corpus. [A KWIC concordance (Key Word In Context) for some word *w* is a file each record of which features *w* at roughly the same field, often set off on both sides by one or two padded blank characters. To the left and right of *w* in a given record is the sequence of words or other character strings which, respectively, precede and follow *w* in a particular line of the text over which the concordance has been constructed.] The first step in the data analysis carried out consisted of establishing senses for each word. General orientation to this task was obtained by consulting an unabridged dictionary's definition; at this point, attention was turned exclusively to the concordance itself as the source of information for the words under scrutiny. In the case of senses, as opposed to parts of speech, it was found that the dictionary served in only the most general way to guide the formulation of definitional criteria. Even where the dictionary senses could be adopted as such, "the set of senses is itself relative to our pragmatic aims—i.e., we ask which of 'the senses' of this entry seem useful, or worth discriminating. To this end we were aided by our accumulated knowledge of what kinds of distinctions are important to content analysis work" (p. 10). Presumably, by "content analysis work" is meant content analysis work in the behavioral sciences. It turned out that the influence of

the domain of application was quite radical in terms of the partitioning of senses, as compared either to a standard dictionary definition or to partitions from other domains.

The goal of the analysis of a given word was to come up with an ordered set of disambiguation rules which bear on the word itself—i.e., look for its morphological characteristics—and on any number of words within a window of word-plus-and-minus-four-words. Any word in this range can be tested by a rule for part of speech, for membership in one or more of a set of semantic categories listed below, and for Boolean combinations of these conditions. Passage or failure of a given test was allowed to determine either assignment of a specific word sense number, or a jump to a subsequent rule in the set, or even to a rule in the set for some other word.

As suggested above, senses were simply stipulated, as a function of the meaning differences which, it was believed, a behavioral scientist would find revealing or important, and after having read one dictionary's listing of meanings for general orientation. For the purpose of word-sense discrimination, Kelly and Stone created sixteen semantic categories: Animate, Human (Male, Female, Kinship), Collective, Abstract Noun (Abstract, Time, Distance), Social Place, Body Part, Political, Economic, Color, Communication, Emotions, Frequency, Evaluative Adjective, Dimensionality Adjective, Position Adjective, and Degree Adverb.

Sinclair [14] considers the lemma *yield*—the collection of word forms such as *yielding*, *yields*, *yield*, etc., each in its several functions. He claims that about 70% of the occurrences of the lemma *yield* in the 7.3-million-word Birmingham database of English display what he calls an "alignment" of "sense and structure." Specifically, if one knows which representative of the lemma occurs in a given citation, e.g., if *yields* as a plural noun occurs—then the assertion is that one can predict with 70% accuracy whether the 'give way' meaning, the 'produce' meaning, the 'lead to' meaning, or one of several "minor meanings" of *yield* is being used.

In Amsler [15], structurally determined keywords of definitions of the *New Merriam–Webster Pocket Dictionary* (G. and C. Merriam, Eds., 1971) are disambiguated essentially by hand. That is, concordances are labeled, as was done by Kelly and Stone. The resulting information is used to automatically construct lexical hierarchies based on the text of the Merriam–Webster definitions. Further work is suggested in which each definition would be disambiguated in its entirety, in the sense that each of the words it contained, and not simply its "keywords" or "head words," would be assigned a sense number. This work would proceed automatically, in a bootstrapping process by which each correctly disambiguated definition would constitute additional data of the store from which word-sense discrimination decisions were made. (For work which continues in the direction of dictionary-processing research taken by Amsler, see Chodorow et al. [16]. Here semantic hierarchies are extracted from *Webster's Seventh New Collegiate Dictionary* by automatic and semiautomatic methods.) In addition, the conjecture is made that such a fully disambiguated dictionary could serve to distinguish word senses in free text [15, p. 123].

Amsler and Walker presented a related idea for a research program in [17]. Word senses can be distinguished, they reported, in the following manner: Let each word within a paragraph be assigned all its possible "subject categories" as listed in the computer-readable version of the *Longman Dictionary of Contemporary English* (Longman Group Ltd., Eds., Longman, Burnt Mill, England, 1978). The print version of this unfortunately does not list these subject categories. The category most frequently represented among the words of the paragraph is the theme or subject area it covers. Accordingly, wherever one of the possible senses of a word within the paragraph is assigned the "theme category," it may be selected as correct. Actually, this account of their approach is probably oversimplified, since they report using some 1600 categories compounded in some way from the roughly 125 "major" and 250 "minor" subject categories of the *Longman Dictionary*.

A sampling of these subject categories, where in the interest of clarity we let the major categories begin with capital letters and the minor ones with lowercase letters, is as follows: Baseball, Building, car building, bricklaying, carpentry, plastering, plumbing, Beauty Culture, cosmetics, hairdressing, perfumery, Basketball, Bible, . . . , Numismatics, currencies, Occult, alchemy, palmistry, astrology, spiritualism, Occupation, medical profession, royal rank, . . . , Transport, Tobacco, Nonautomotive Vehicles, Water Sport, swim clothing, swimming, Winter Sport, curling, ice skating. As should be clear, the minor categories cited each belong with the closest preceding major category.

Weiss [18] confronts sense differentiation from the point of view of the discipline of information retrieval. He designs and tests a program embodying the following procedure: He wishes to learn the contextual concomitants of sense-label assignments associated with a set of input sentences all featuring some word w, and originating in any sort of corpus. To this end, he induces an ordered set of sense-determination rules on the basis of an initial, "training" corpus C1, and applies these to a "test" corpus C2 drawn (randomly?) from the initial set of citations, in such a way that C1 and C2 are disjoint.

Starting with the empty set of rules, his procedure examines each sentence/label pair of C1, and performs one or more of the following operations: It adds a rule or rules; it deletes a rule or rules; and/or it adds a deleted rule to a running list of "prohibited new rules." What is a rule? There are two sorts—template and contextual rules. Template rules are of the form, "word x occurs in the current sentence

Table 1 Results of Weiss's sense-label assignment [18, p. 40].

Word	*T*	*C*	*I*	*U*	*RR**	*RP*†
DEGREE	180	160	12	8	0.89	0.93
TYPE	180	164	4	12	0.91	0.98
VOLUME	180	152	15	13	0.84	0.91
Total	540	476	31	33	0.88	0.94

* RR: Number right over number seen.
† RP: Number right over number both seen and categorized.

within two words to the left or right of word *w* whose senses are being disambiguated." Contextual rules say: "word *y* occurs in the current sentence within five words to the left or right of word *w*." There are two additional differences between template and contextual rules:

- Template rules are ordered, *en bloc*, before contextual rules.
- Contextual rules do not count "function words"—in some sense of this term which need not occupy us—but template rules count all words.

When meeting a new sentence/label pair, the procedure first attempts to apply, in order, every rule it knows, until a rule is satisfied. It then matches the label predicted by rule with the actual label of the sentence. If there is a match, it simply proceeds to the next sample. Otherwise, it traces back to the offending rule, and both deletes it and adds it to the list of rules which cannot be coined in the future. There is a tendency for the useful rules within each set—the templates and the contextuals—to "rise to the top of the stack" as incorrect rules are deleted.

Once the final sentence of C1 is examined, attention shifts to C2. Now, in test mode, the only function remaining of the three utilized on C1 is rule application. That is, each sentence of C2 is labeled according to the rule set derived from C1. An additional function is added, specific to the present phase: tallying of right and wrong answers. Accuracy scores are determined on the basis of the following figures: the total T of samples in the data set ($=N$); the number C of correctly resolved ambiguities; the sum I of incorrectly resolved ambiguities; and the number U of unresolved ambiguities. The two evaluative statistics employed are "resolution recall" $RR = C/T$, and "resolution precision" $RP = C/(C + I)$.

Weiss obtains the results shown in **Table 1**, where *DEGREE*, *TYPE*, and *VOLUME* are the actual words used to test his procedures.

The work of Dahlgren [19] formally resembles both that of Weiss and that of Kelly and Stone in the sense that the mechanism of sense discrimination used is an ordered set of categorial, i.e., "yes-or-no" as opposed to probabilistic, rules. Like Weiss, she uses "frequent collocates" as one sort of criterion for establishing differences. Like Kelly and Stone, she also uses a second kind of rule, one based on the local syntax of the token under analysis, and relying on a previously-carried-out parse. Sample syntactic questions, used when a noun is being disambiguated, concern the presence or absence of an associated definite article, personal pronoun, or noun complement. Other such questions include whether the noun serves as object of a preposition, or as subject or object of particular verbs. A third sort of question, and the most novel of the three types, is the one using "common-sense knowledge" as defined on the basis of the results of a number of psycholinguistic studies (e.g., Ashcraft [20], Rosch et al. [21]) within what amounts to prototype theory. Each word to be disambiguated is represented via a "tangled hierarchy" of ontological predicates derived from the sort of study just referred to, and various questions are defined which turn on the similarity or lack of similarity between the representation of the word in focus and some other word or words in its environment. For instance, one question type looks for such resemblance in a second noun standing in conjunction with the noun under examination; if such commonality is indeed found, the proper sense is deemed to have been identified. The rules are applied in three tiers: first the "frequent collocate" questions, then the syntactic questions, and finally the common-sense-knowledge questions—with later tiers being reached only when earlier ones have failed to yield a unique sense selection.

Concordances drawn from a corpus of legislative English were used to test the rule set. Seven nouns (*office*, *hand*, *company*, *idea*, *crop*, *people*, and *school*) were used, in an average of 313 different citations each. The correct sense was selected in 96% of the citations attempted.

2. Description and results of experimentation

The aim of the experiment to be reported was to compare three methods of computational determination of English word senses. The corpus selected for processing consisted of some 22 million tokens' worth of the Canadian House of Commons' official proceedings, for a period during the late 1970s. As such proceedings are called Hansards [22] in Canada, the corpus is referred to here as the Hansard database. A pool of about 1000 types occurring in the Hansards was chosen so as to ensure a varied frequency distribution of sample types. Of these 1000 or so, five types were selected to which each of the three methods would be applied, as a basis for comparison. Potential members of the set of five types were chosen randomly, and those candidates were retained which met preset requirements of usefulness for the experimental task (e.g., the type must have at least three senses within a single part of speech).

The experimental types were *interest*, *point*, *power*, *state*, and *terms*. A concordance over all 22 million Hansard

tokens was compiled for each. [The TUPLES text analysis system (Byrd [23]) was used for the generation of concordances and for additional purposes to be described subsequently.] After close inspection of a concordance, it was decided which part of speech and which selection of possible senses within that part of speech would be chosen for the type in question. About 2000 concordance lines were obtained for each test word, after elimination of duplicate entries, incorrect-part-of-speech entries, and entries in which the test word bore a sense other than one of those selected. All lines of each such concordance were then hand-labeled as to which of the selected senses characterized the node word (in the terminology of Sinclair [24])—the token in focus in each particular concordance line. All of the test words except one had four stipulated senses; the remaining word had three such senses. The part of speech chosen for all five words was "nominal." See [2] for details and commentary. Each concordance was then randomly partitioned into an approximately-1500-line "training corpus" and a 500-line "test corpus." The former would be submitted to each method under consideration, as the basis for the formulation of generalizations and predictions as to when the node word would most probably bear each of the given senses. The latter would permit the evaluation of the degree of correctness of each method's predictions concerning a test word.

As stated earlier, three methods of sense discrimination were under scrutiny. A "method" consisted of a set of 81 "contextual categories" or "contextual event types" which could be defined anew for each of the test words; or once and for all, independently of which test word they were applied to; or in a mixture of both these modes. The notion of utilizing "context" to discriminate word senses was thus made precise: The "context" of a token with reference to the concordance line in which it figures was taken as the pattern of presences and absences in that line of exponents of each of 81 "event types."

In order to quantify the predictions of each method with regard to the contextual conditions most likely to be associated with the occurrence of each stipulated sense of a test word, and thereby to permit the comparison of the methods, a decision tree (Meisel [25]) with maximum mutual information as the node-label selection criterion (Lucassen [26]) was constructed for each of the five 1500+-line training corpora, using each of the three methods being compared, for a total of 15 trees. (See the Appendix for an introductory sketch of the concepts of "decision tree" and "maximum mutual information.") Every tree so derived from a training corpus was then employed to predict the senses of the 500 lines of the corresponding test corpus. The probability was calculated that the sense associated with each of the 500 lines was correctly predicted. The average of these probabilities was taken, and the result expressed in logarithmic form, yielding the entropy of the test data as modeled by each of the fifteen training-data trees. The average entropy for each method over the five test words analyzed was the statistic used to compare the three methods. The method achieving the least uncertainty in its sense predictions of the words examined would display the lowest entropy.

A dictionary-based domain-general method (henceforth method DG) was developed, based loosely on the work of Amsler and Walker reviewed above. Each of the 500 most frequently appearing words of the entire 2000-line concordance was automatically looked up in the on-line version of the *Longman Dictionary of Contemporary English* (henceforth LDOCE). If any of the word's definitions in LDOCE features a given "subject code" (see the account of the work of Amsler and Walker described earlier), then the word is added to a file listing all the words in the concordance whose definitions include that subject code. When processing is completed, what results is a sort of profile of the words which have occurred most frequently in an entire 2000-line concordance, from the point of view of the subject codes employed in LDOCE. The 500 most frequently occurring words in one of these concordances correspond to all those words which occur there from ≥ 5 times to ≥ 10 times. The requisite 81 categories were obtained in this manner for each of the five test words.

In contrast with domain-general method DG, two domain-specific methods were developed, henceforth DS1 and DS2. DS1, inspired in some measure by Weiss [18], described earlier, might better be called text- or concordance-specific than domain-specific, however, as it is based completely on the frequencies of different lexical items in the 1500+-line training corpus. Two classes of categories, on the analogy of Weiss's template and contextual rules, were used here. The first consisted of the 41 types occurring most frequently in the window $n \pm 2$ of the training corpus being processed, where $n \pm x$ means the sequence of words beginning x words to the left of the node word, and ending x words to its right. The aim here was to capture those words in close grammatical construction with the node. The second class of categories excluded "function words" from its purview (see Black [2] for details) and ranged over an entire concordance line. The 40 most frequent words fulfulling these conditions for a given training corpus made up this class. The idea of this second sort of category was to recover collocates (see, for example, Sinclair [24], Jones and Sinclair [27]) of the node. Thus, in the case of DS1 each category consisted of a single word, whereas in DG a category was more often comprised of a list of words—all those with one or more LDOCE definitions that included the subject code which named the category.

The second of the two domain-specific methods, DS2, partially resembles DS1 in that 20 of its 81 categories were concordance-specific, consisting of the 20 words or two-word sequences which occur most frequently in a given training

Table 2 Entropy of test data for test words selected.

Test word	*Entropy(chance)*
interest	2.00
point	1.20
power	1.97
state	1.99
terms	1.93
Average	1.82

corpus in the windows $n \pm 1$ and $n \pm 2$, respectively. However, the bulk of the categories of DS2—the remaining 61—were derived not from the test-word concordance itself at all, but rather, strictly from the concordances of 100 other types occurring in the Hansards. Crucially, none of the five test words was included in this set of 100 concordances. These 61 files were generated in a manner completely unlike any of the procedures described so far. Recall that about 1000 types occurring in the Hansards were chosen to guarantee a broad frequency distribution of sample types. A chance selection was made from these 1000, of 100 types, for which concordances were then produced over the full approximately 22-million-word corpus. Then each such concordance was analyzed quite closely from what might be called a content-analytic point of view, and many of the words and expressions which were uncovered in the course of this analysis were loaded into one or another of the 61 "content-analytic" DS2 files.

Actually, the loading of the 61 files took place through a bootstrapping process which started with no "content-analytic" files at all and, of course, no words or expressions entered in such files; and which ended with 61 fully stocked files. The procedure by which this result was obtained was the following: The 100 concordances were examined seriatim. In the case of any given concordance, this examination began with a first reading whose aim was the stipulation of a set number of node-word meanings relative to the corpus. Each sense was assigned a label consisting of a number. There followed a second reading of the concordance, in which all words and expressions which occurred were listed and partitioned according to the node-sense number of the line in which they were found. Next, an attempt was made to partition along thematic lines each list of sense-particular words and expressions. In the course of processing the first ten or fifteen concordances, any theme was entertained. Some of these early themes which did not survive were Advertising, Capital Goods, Industry, Obligative, and Try. However, beyond this point it started to become clear what the useful thematic categories were for this domain, or, better, "world." (For the notion "world," see Black [2, Ch. 2].) A theme was considered useful and adopted if the presence of one of its exponents in a concordance line—either alone or in conjunction with exponents of a small number of other themes—sufficed in a large percentage of cases to determine the sense number of the line's node. The proliferation of themes was, in an informal sense, asymptotic, and fell off dramatically after about 40 concordances had been considered. In a final reading, conducted after the processing of all 100 concordances was complete, those concordances taken up at the beginning of the process were reanalyzed in terms of the final set of thematic categories, so that uniformity of classification was ensured. A sample of the 61 categories arrived at in this manner is as follows: CONTROVERSY, DOCUMENT, ENERGY, GOVERNMENT_BODY, NEGATIVE_CONNOTATION, MILITARY/FORCE, PARLIAMENTARY_MOVES, POWERFUL_PEOPLE, RESPONSIBILITY, TRANSPORT, VOTERS. It is essential to note that these content-analytic categories were hypothesized to be valid for all or most of the types occurring in the Hansards, not simply for the five test words. Part of the motivation for selecting the five experimental words via random methods was to test the validity of the DS2 categories across the range of Hansard types.

Fifteen analyses of test data were performed on the basis of the decision trees obtained from the fifteen corresponding analyses of training data. These results are presented below.

As a preliminary, it will be useful to give a brief explanation of the statistical measure in terms of which the results are expressed. This is the entropy measure. Consider that if we had four possible senses for a node, and if all senses were equally likely, the entropy of the data under examination would be 2 bits. That is, it would require 2 bits to convey the information that four choices are present. (Hereafter, entropy will be understood to be expressed in bits.) If it happened that all senses were not equally likely—i.e., that we knew there were more instances of some senses than of others—the entropy of the data would be a little less than 2. Now if by using method DG or DS1 or DS2 we were able on average to eliminate from consideration two of the possible sense choices, and if the remaining two choices were equally likely, then the entropy of the data would be 1. If there were a perfect method, one which always correctly narrowed the four possibilities down to a single choice, then the entropy of the data would be 0.

In the case of the data we are considering, all words except one had four senses, while the remaining one (*point*) had three. If the only facts known about the test data were how many instances were present of Sense 1, Sense 2, and so forth, the results shown in **Table 2** would occur on average. The results obtained via methods DG, DS1, and DS2 (see **Table 3**) should be compared with the results of chance selection in Table 2.

A further result is that, for two of the words selected at random, when the 20 structural categories of DS2 were replaced by the 20 DS1 categories that appear highest in the

DS1 decision tree, the result was nearly identical in each case with the minimum of DS1 and DS2. Specifically, this combination of DS1 and DS2 for the word *point* resulted in an entropy of 0.64, which is exactly identical with min(DS1, DS2) for this word. The same combination applied to the word *state* yielded an entropy of 0.66, which is within 0.02 of min(DS1, DS2) for *state* (in fact, it is better by that amount). Projecting this result onto the entire set of five test words by averaging the minima of DS1 and DS2, we obtain an estimated entropy of 0.78 using the particular method of combining the two DS approaches which is described just above.

An additional statistic of interest is "percent correct," i.e., the number of correct sense choices divided by the total number of predictions made. **Table 4** shows the results of random sense selection. The percent-correct results obtained when methods DG, DS1, and DS2 are applied to the data are given in **Table 5**.

3. Discussion

In the experiment under discussion, chance selection of senses yielded 1.82 on average—close to the "worst case" figure of 2.00 referred to in the previous section. Now both methods DS1 and DS2 were able to reduce the entropy of the data to below 1.00, whereas method DG turned out considerably closer to chance in its results (at 1.49, on average) than to either of the remaining two methods. In fact, in the case of the test word *point*, random sense selection would have resulted in an entropy of 1.20, while method DG actually did worse than chance, at 1.48.

In terms of percent correct, again methods DS1 and DS2 do roughly twice as well as chance, at 72% and 75% respectively, *vis-à-vis* the chance percentage of 37% correct. But method DG achieves only a 27% improvement over chance, with a percent-correct rate of 47.

An analysis of the DG trees themselves, and of the contents of the DG categories, suggests why this method's performance was somewhat lower than that of DS1 or DS2. Those categories which rise to the top of a useful decision tree, in the present experimental environment, are typically connected with the thematic or structural functioning of the node word in a way that is intuitively obvious. Thus, the category "PLACE" appears at the top of the DS2 decision tree for *point*, since almost every sample in which the word is used to mean 'geographical location' has some exponent of this category, and not many other lines do. Hence, "PLACE" is quite "helpful" (informative) for the disambiguation of the nodes of this concordance. Similarly, in the DS1 tree for *terms*, the categories "IN" "THE," and "OF" rise to positions at or near the root, and the reason for this seems to have to do with the predictive power of such frequent expressions as *in terms of* and *the terms of*, each of which characterizes two of the four senses of *terms* (although not the same two) to the practically complete exclusion of the remaining meanings. But this phenomenon appears not to characterize the DG trees. More specifically, there are cases where the categories that exist in the list of 375 Longman subject classes—categories one might choose on an intuitive basis as "relevant"—do not seem to attract any of the 500 most frequently appearing words of a test item, and therefore do not get a chance to appear near the root. On the other hand, there are instances where the intuitively likely categories are in fact represented in the tree. But then they tend not to inhabit the upper reaches of the tree. For instance, in the DG tree for *interest*, classes such as "COMMERCE," "BANKING," and "FINANCIAL" do occur in the decision tree, but they are not near the top, and only serve to categorize small numbers of samples. What this situation suggests is that the DG method may fail to reflect the thematic organization of the concordances analyzed.

A look into the DG category files themselves provides further possible explanations of the lower DG scores. Even

Table 3 Entropy of leaves of generated decision trees for test data.

Test word	*Entropy(DG)*	*Entropy(DS1)*	*Entropy(DS2)*
interest	1.47	0.83	1.23
point	1.48	0.76	0.64
power	1.50	0.74	0.77
state	1.63	0.94	0.68
terms	1.38	1.21	0.99
Averages	1.49	0.90	0.86

Table 4 Results for random sense selection: percent correct.

Test word	*% Correct(chance)*
interest	28
point	66
power	33
state	27
terms	31
Average	37

Table 5 Results for sense selection by methods DG, DS1, and DS2: percent correct.

Test word	*% Correct(DG)*	*% Correct(DS1)*	*% Correct(DS2)*
interest	47	72	63
point	46	76	83
power	48	77	76
state	42	71	81
terms	50	62	70
Averages	47	72	75

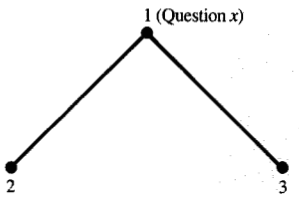

Figure 1

One-level decision tree.

the most common "function words" are given very complete descriptions in terms of the subject classes, so that, for example, every occurrence of *on* or *off* is categorized as indicating that there is some sense in which the topic "ELECTRICITY" is being discussed. Therefore, one might guess that DG performance would improve substantially if function words were removed from the input to the Longman categories, and only content words were retained. However, what typically happened when this was tried was that words which were "content" items but still high in frequency, such as *point* when it occurred in concordances other than its own, were placed in a rather large number of categories, including, e.g., "MATHEMATICS" and "LAW." And most of these categories did not seem to bear directly on the themes addressed in the Hansard concordances. Frequent content words accumulated in these less helpful classes, so that powerful categories emerged which perhaps served to diminish the effectiveness of the DG classification scheme.

Turning to DS1 and DS2, it seems that well-chosen (and automatically chosen) exclusively structural categories did nearly as well as a combination of about 3/4 thematic and 1/4 structural classes. The important point, it appears, is that the two methods each beat each other soundly once (with DS1 the winner for *interest* and DS2 the victor for *terms*), and that overall DS1 is superior two of five times and DS2 three of five times. This indicates that neither method can stand by itself, but rather that each is in need of the other to perform well. Clearly this observation suggests tasks for future research, such as, first, varying the number and character of DS1 categories entering into combination with the DS2 host; and second, attempting to discern the source or sources of DS1's power by teasing apart its elements and investigating their performance in different combinations, and in combination with the DS2 structural classes.

The experimentation presented suggests a number of additional avenues for future research. The underlying purpose of the study discussed here has been to test the efficacy of variant basic orientations in the discrimination of English word senses. None of the approaches employed could now be used exactly as is to actually process large volumes of text with the aim of automatically differentiating word senses. This is because of the necessity for hand-labeling of concordances, and further, because of the grand scale on which this labeling is required. In the background lies another difficulty, but one whose elimination is probably a more distant objective. This difficulty is that the content analysis itself had to be carried out by a person. While automation of this latter task does not seem a forlorn hope, obviating the need for hand-labeling does appear the greatly more tractable task of the two at present.

Appendix

An intuitive approach to the terms "decision tree" and "maximum mutual information" is as follows: Assume that we wish to construct a system of event classification which yields us progressively more certainty as more details are considered. In the case of word-sense disambiguation, we need to reduce our uncertainty concerning which of a pool of preselected senses best characterizes a particular instance of the use of a test word. The formal measure of uncertainty is called entropy, and is equivalent to the average number of bits of information it takes to transmit the identity of an event using an optimal coding scheme. One system of classification of the type we seek is called the binary decision tree. To understand what this amounts to, start with a set of yes-or-no questions which may be asked about events of the type under study; for our purposes, these might be questions of the form, "Was there a word or expression of contextual event type T in the concordance line under scrutiny?" From this set of questions choose that question with the following characteristic: Knowing the answer to this question reduces the average entropy of the outcomes of the process more than knowing the answer to any of the other questions would. The question that reduces entropy the most is said to display "maximum mutual information" with the outcome of the process. We can picture what we now have as a one-level decision tree, consisting of a root and two leaves (see **Figure 1**). Note that node 1, the root node of the tree, is associated with Question x, the one which best met the decision criterion, just outlined, for node-label selection. The average entropy of nodes 2 and 3 is lower than the entropy of node 1. We extend the decision tree beyond level 1 by processing nodes 2 and 3 in the same manner as node 1 was processed. To treat node 2, we find the best question to ask of all samples for which Question x was false; to treat node 3, we find the best question to ask of all samples for which Question x was true. Again, "best" means "leading to the greatest reduction in average entropy." So the average entropy of the leaves of the two-level decision tree (**Figure 2**) is less than that of the leaves of the one-level tree shown in Figure 1. A binary decision tree is a vehicle, then, for

discovering an answer to the question, "Which of this pool of questions are most useful to ask of my data, and in what order should they be asked?"

Acknowledgments

The author is grateful to L. Bahl, P. Brown, and R. L. Mercer of the Continuous Speech Recognition Group of the IBM Thomas J. Watson Research Center for discussions concerning the statistical basis of the research presented here.

References and note

1. W. N. Francis and H. Kučera, *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin, Boston, 1982.
2. E. Black, "Towards Computational Discrimination of English Word Senses," Ph.D. Dissertation, Graduate Center, City University of New York, 1987.
3. F. Débili, "Analyse syntaxico-semantique fondée sur une acquisition automatique de relations lexicales-semantiques," Thèse de Doctorat d'Etat (No. d'ordre 2541), Université Paris XI, 1982.
4. M. Gross, "Projecting the Lexicon-Grammar on Texts," presented at the First International Roman Jakobson Conference, New York University, October 12, 1985; to appear in *New Vistas in Grammar: Invariance and Variation*, L. R. Waugh and S. Rudy, Eds., John Benjamins, Amsterdam, the Netherlands.
5. M. Gross, "A Linguistic Environment for Comparative Romance Syntax," *Papers from the XIIth Linguistic Symposium on Romance Languages*, University Park, MD, April 1–3, 1982, P. Baldi, Ed., John Benjamins, Amsterdam, 1984, pp. 373–446.
6. E. F. Kelly and P. J. Stone, *Computer Recognition of English Word Senses*, North-Holland Publishing Co., Amsterdam, 1975.
7. R. W. Budd, R. K. Thorp, and L. Donohew, *Content Analysis of Communications*, Macmillan Publishing Co., New York, 1967.
8. *Advances in Content Analysis*, K. E. Rosengren, Ed., Volume 9, *Sage Annual Reviews of Communication Research*, Sage Publications, Beverly Hills, 1981.
9. G. D. Forney, Jr., "The Viterbi Algorithm," *Proc. IEEE* **LXI**, 268–278 (1973).
10. E. S. Atwell, "Constituent-Likelihood Grammar," *Newsletter of the International Computer Archive of Modern English (ICAME News)* **7**, 34–66 (1983).
11. A. D. Beale, "Grammatical Analysis by Computer of the Lancaster–Oslo/Bergen (LOB) Corpus of British English Texts" (ms. available from the Unit for Computer Research on the English Language, University of Lancaster, Bailrigg, Lancaster LA14YT, England), 1985.
12. A. D. Beale, "A Probabilistic Approach to Grammatical Analysis of Written English by Computer" (ms. available from the Unit for Computer Research on the English Language, University of Lancaster, Bailrigg, Lancaster LA14YT, England), 1985.
13. R. Garside and F. Leech, "A Probabilistic Parser" (ms. available from Unit for Computer Research on the English Language, University of Lancaster, Bailrigg, Lancaster LA14YT, England), 1985.
14. J. McH. Sinclair, "Sense and Structure in Lexis" (ms. available from English Language Research, 357 Bristol Road, Edgbaston, Birmingham B5 7SW, England), 1985.
15. R. A. Amsler, "The Structure of the *Merriam–Webster Pocket Dictionary*," Ph.D. Dissertation (*TR-164*), University of Texas at Austin, 1980.
16. M. S. Chodorow, R. J. Byrd, and G. E. Heidorn, "Extracting Semantic Hierarchies from a Large On-Line Dictionary," *Proceedings of the 23rd Meeting of the Association for Computational Linguistics*, University of Chicago, 1985, pp. 299–304.

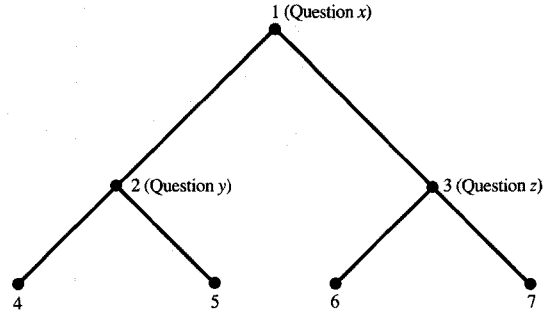

Figure 2

Two-level decision tree.

17. R. A. Amsler and D. Walker, "Knowledge Resource Tools," presented at the IBM Corporate Technical Institute, Thornwood, NY, December 6, 1985.
18. S. F. Weiss, "Learning to Disambiguate," *Information Storage and Retrieval* **9**, 33–41 (1973).
19. Kathleen Dahlgren, "Using Commonsense Knowledge to Disambiguate Word Senses," *Proceedings of the Natural Language Understanding and Logic Programming Conference*, Vancouver, B.C., Canada, 1987, pp. 215–229.
20. M. H. Ashcraft, "Property Norms for Typical and Atypical Items from 17 Categories: A Description and Discussion," *Memory & Cognition* **6**, 227–232 (1976).
21. E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic Objects in Natural Categories," *Cogn. Psychol.* **8**, 382–439 (1976).
22. "Hansard" is used in England to refer to the official published reports of the debates and proceedings of the British Parliament, and in Canada, similarly, for the Canadian House of Commons. One Luke Hansard (1752–1828) and his descendants compiled the British reports until 1889.
23. R. Byrd, "The TUPLES Text Analysis System: Description and Users' Guide," IBM Research Report, available from R. Byrd, Manager, Lexical Systems Project, IBM Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598.
24. J. McH. Sinclair, "Beginning the Study of Lexis," C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robins, Eds., *In Memory of J. R. Firth*, Longman, London, 1966, pp. 410–430.
25. W. S. Meisel, *Computer-Oriented Approaches to Pattern Recognition*, Academic Press, Inc., New York, 1972.
26. J. M. Lucassen, "Discovering Phonemic Base Forms Automatically: An Information Theoretic Approach," *Research Report RC-9833*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, 1983.
27. S. Jones and J. McH. Sinclair, "English Lexical Collocations: A Study in Computational Linguistics," *Cahiers de lexicologie* **24**, 15–61 (1974).

Received March 18, 1987; accepted for publication August 5, 1987

Ezra W. Black *IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Black is a Research Staff Member in the Computer Sciences Department at the Thomas J. Watson Research Center. He received his B.A. in French from the City College of New York in 1971, his M.A. in French from Columbia University Teachers College in 1973, and his Ph.D. in linguistics from the Graduate Center of the City University of New York in 1987. Dr. Black joined IBM at the Research Center in 1987, working on language modeling for continuous speech recognition and on French–English machine translation. He is interested in the construction of broad-coverage generative grammars of English and the Romance languages, in collocations, and in the philosophical foundations of language study.