

12-1-95

SANDIA REPORT

SAND95-2532 • UC-705
Unlimited Release
Printed November 1995

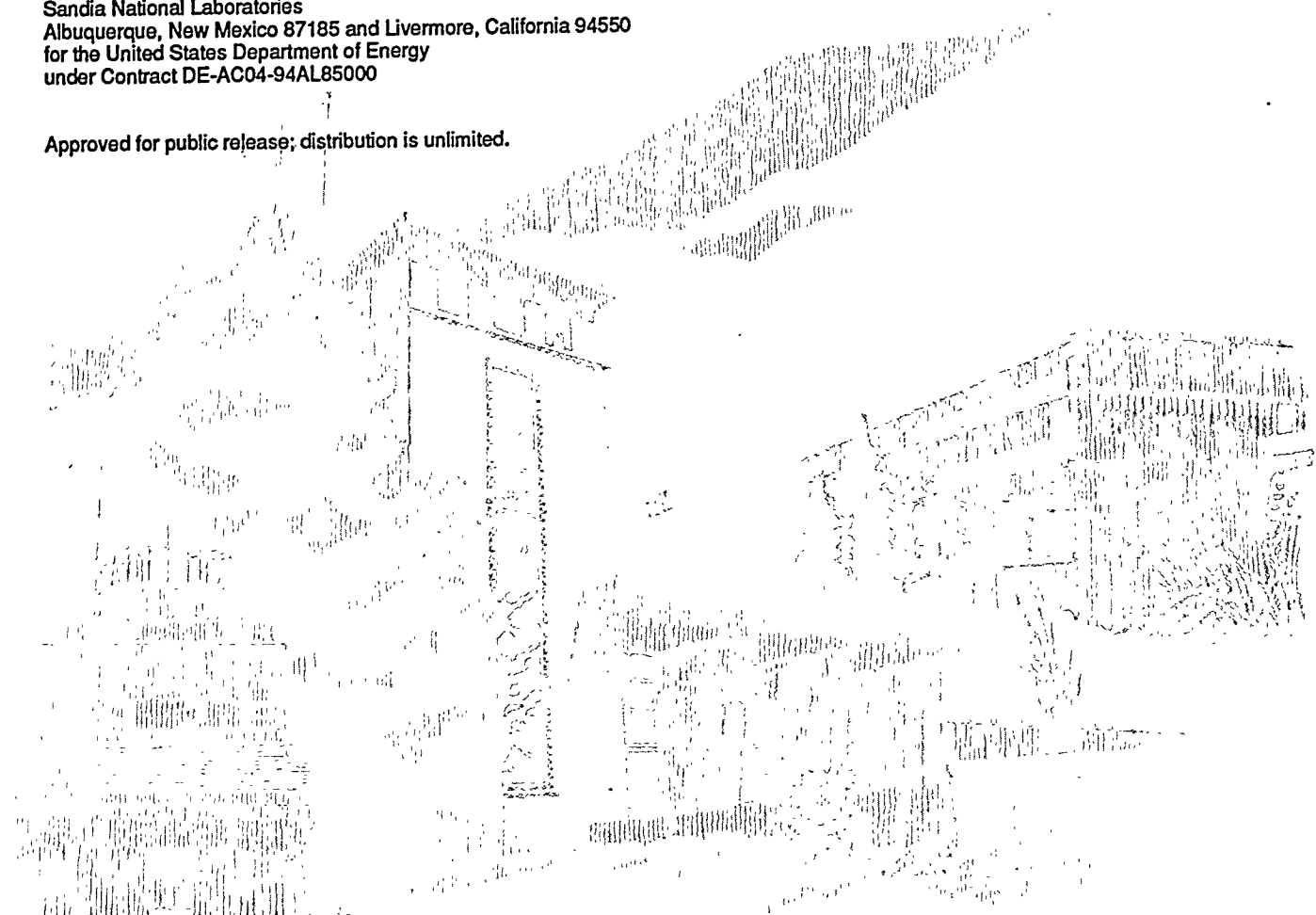
RECEIVED
DEC 12 1995
OSTI

Extraction of Information from Unstructured Text

Nancy H. Irwin, Sharon M. DeLand, Stephen V. Crowder

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550
for the United States Department of Energy
under Contract DE-AC04-94AL85000

Approved for public release; distribution is unlimited.



SF2900Q(8-81)

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831

Prices available from (615) 576-8401, FTS 626-8401

Available to the public from
National Technical Information Service
US Department of Commerce
5285 Port Royal Rd
Springfield, VA 22161

NTIS price codes
Printed copy: A03
Microfiche copy: A01

Extraction of Information from Unstructured Text

Nancy H. Irwin
Software Surety Department

Sharon M. DeLand
Advanced Concepts and Architectures Department

Stephen V. Crowder
Statistics and Human Factors Department

Sandia National Laboratories
Albuquerque, NM 87185

Abstract

Extracting information from unstructured text has become an emphasis in recent years due to the large amount of text now electronically available. This status report describes the findings and work done by the end of the first year of a two-year LDRD. Requirements of the approach included that it model the information in a domain independent way. This means that it would differ from current systems by not relying on previously built domain knowledge and that it would do more than keyword identification. Three areas that are discussed and expected to contribute to a solution include (1) identifying key entities through document level profiling and preprocessing, (2) identifying relationships between entities through sentence level syntax, and (3) combining the first two with semantic knowledge about the terms.

Acknowledgement

This work was funded by the Laboratory Directed Research and Development (LDRD) Program. The authors wish to thank John Taylor, chairman of the LDRD Office's National Security Technology Program Area, for his enthusiastic support.

Contents

Introduction.....	1
PreviousWork	2
System Description	2
Profiling.....	4
Keyword Identification	5
Cluster Analysis.....	6
Nominal Compounds.....	7
Unknown Words.....	9
Intra-sentence Relationships.....	11
Inter-sentence Relationships.....	11
Entity-relationships Tables	12
Identifying Equalities	12
Identifying Events.....	13
Summary and Future Work	15
References.....	17
APPENDIX - Cluster Analysis	18

Figures

1 Block Diagram of the Information Extraction System	3
2 Dendrogram from Cluster Analysis Example	22

Tables

1 Breadth and Frequency Data from Example Article	19
2 Initial Distance Matrix for Cluster Analysis Example	19
3 Distance Matrix after Step 1 of Clustering Algorithm	21
4 Distance Matrix after Step 2 of Clustering Algorithm	21
5 Distance Matrix after Step 10 of Clustering Algorithm	21

Extraction of Information from Unstructured Text

Introduction

Large amounts of unstructured information (e.g., text) is electronically available to humans. The problem is that the data is not well organized or structured in such a way as to allow for easy computer/human access. Work in this area has taken several forms. Keyword and keyword expansion searches in text retrieval applications find documents most relevant to a user request. Automatic indexing programs generate 'back of the book' indexes helping users locate pages where particular references are made. Information extraction systems are largely experimental and attempt to extract and "understand" the information in the text such that a user might query for exact information. Current research systems of the latter type typically choose a domain, build a priori knowledge about that domain and then define, also a priori, the script or kind of information to be extracted. This ambitious LDRD attempts to do the latter through an entity-relationship approach without the a priori domain knowledge or script.

Although the long term goal of this work is to support analysis in the arms control and verification arena, the original proposal was clear in stating that a solution in analyzing unstructured data should be domain independent. To this end documents in three areas have been chosen for testing and experimentation: treaties, baseball articles, and medical records. Domain-knowledge dependent systems, which today's research systems are, require knowledge engineering to switch domains and miss important information that was not considered or preselected during the human knowledge building phase.

Building a model of document information depends on having syntactic and semantic information about the words and phrases used in the document and on the sentences being grammatically parsed. Some public-domain software was used where it provided (often partial) solutions to the above-mentioned prerequisites to the extracting and modelling problem. In other cases, intermediate problems have been solved in-house but only in so far as to facilitate working on the latter problem. If promising results ensue, it would be reasonable to buy a parser with a rich lexicon or even more comprehensive pieces with the then-known desired characteristics.

Because of the domain-independent requirement of this problem, learning knowledge automatically from the document has been a focus and has driven a two-pass approach. During the first pass, the document is profiled and preprocessed. The sections on keyword identification, cluster analysis, nominal compounds, and unknown words are all products of this pass where the goal is to learn as much as possible that will aid in the correct identification of important entities in pass two. In pass two, once the entities are identified, relating the entities to each other and to the events is then the focus of the processing. The sections on intra-sentence unification, inter-sentence unification and entity-relationship tables reflect this emphasis. This approach distinguishes itself from the current work in this area in that a template of desired information is not predefined. This suggests that the outcome, if to be in the form of a template, can only be in a less specific template than these other systems.

Previous Work

In 1992 the DARPA-sponsored Fourth Message Understanding Conference (MUC-4) focussed on a competition of participants' text analysis systems. The task was to extract information in the form of templates from newswire articles on terrorism in Latin America. Information in the defined template included items such as perpetrator, victim, instrument, location, date, etc. A training set had been given to participants ahead of time; the competition included a new set. Scoring was based on recall and precision. Recall is the percentage of possible answers which are correct. Precision is the percentage of answers given which are correct. SHOGUN, a joint effort of the GE Research Center[RAU89] and Carnegie-Mellon University, and CIRCUS from the University of Massachusetts[RIL94] are two systems of this type.

Since that conference, groups have used the 1500 terrorist articles (including answers) as a testbed corpus for their systems and have used the scoring results of MUC-4 as a yardstick against which to measure their own systems. Participants of the experiment found that the domain-specific knowledge building in preparation for the conference to be extremely time consuming. Words such as 'bomb', 'murder' and 'attack' were used as keys to trigger frames and patterns for extracting the desired information. Interest has been sparked in the area of building tools to speed up the domain-specific knowledge building activity. AutoSlog[RIL93] from the University of Massachusetts and PALKA[KIM93] from the University of Southern California are two examples.

Information extraction systems are also tapping into large databases of information as they become available. WordNet[WOR93] is a semantic word database organized in "is-a" and "part-of" hierarchies and sets. Longman's learner dictionary has 81 syntactic codes for large numbers of words that provide information essential for clean parsing. With access to this kind of data, activity is also stirring in the area of resolving which meaning or sense applies in a particular context. These large databases, however, are still not providing the role-mapping that some of the domain-specific knowledge bases provide. For example, the actor of a 'buy' is the recipient, while the indirect object of a 'sell' is the recipient. Just knowing that they are both transfer events may not be enough.

System Description

The implemented approach uses a two-pass, multi-step process as shown in Figure 1. During pass 1, words are profiled and nominal compounds, important keywords, and certain unknown words are identified and preprocessed and provide input into the lexicon processing of pass 2. During pass 2, identification of the entities and relationships is targeted.

In both passes, text documents are processed on a sentence by sentence basis where a tokenizer divides sentences into word and punctuation tokens and assigns parts of speech to all words and semantic categories to nouns and verbs.

The parts of speech assignment is accomplished using several sources.

- A "special word" word list handles frequently used words and words needing particular attributes not supplied by other sources. 'Said', for example, is not only a past and past participle verb but also often introduces clauses where the 'that' is understood.
- PC KIMMO, a public domain lexicon, provides basic parts of speech such as

- adjective, singular noun, adverb, etc. for a larger set of words.
- Suffix analysis provides basic parts of speech to unknown words. E.g., -ly implies adverb; -ing implies present participle. Suffix analysis is also used to get the root form of a word which is then used in coordination with other approaches.
- Existence as noun in the semantic net.

The semantic category assignment of nouns and verbs depends heavily on the data supplied with WordNet, a public domain linguistic database. Among other things, WordNet groups nouns into the following categories: act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, and time. Words with multiple senses often have multiple categories. To this original set of categories was added amount, date, reference, and acronym. Each word is also associated with a definition in the WordNet, but this facility is mostly not being exploited at this time. About 7500 first names were also added to the data.

WordNet also supplies categories for verbs: body, change, cognition, communication, compete, consume, contact, creation, emotion, motion, perception, possession, social, stative, weather and process. Since half the nouns in the English language are also verbs and several of the categories overlapped already (cognition, communication), processing was simplified by folding all the verb categories into the noun categories. Emotion, for example, was mapped to feeling and stative mapped to state. Also from the point of view of understanding language content, it is probably not important as to whether the sentence used the noun or verb form ("made a purchase" vs. "purchased").

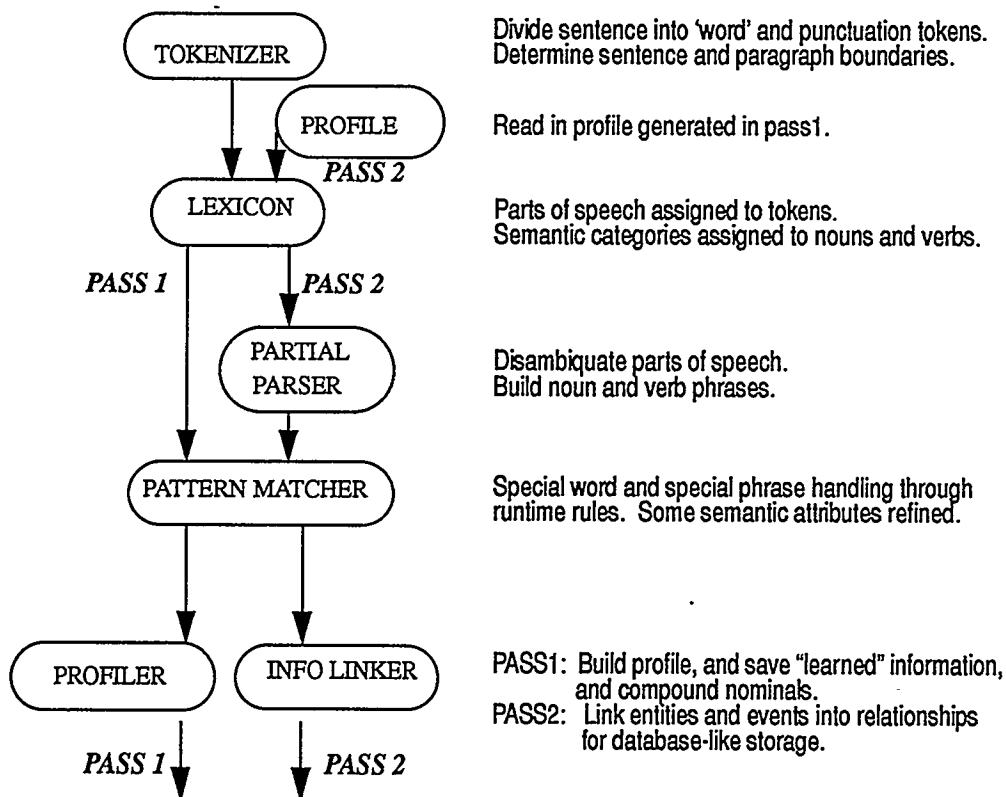


Figure 1: Block Diagram of the Information Extraction System

Some of the categories seem to group as events: act, cognition, communication, event, process, etc. Others intuitively fall into "entity" groupings: animal, body, plant, shape, substance, time, location. Given an event, entities may fill roles of an event: actor, recipient, object, time, location. Entities also may have roles that are filled by other entities: location, address, position name, date of birth, quantity, and frequency.

The output of pass 1 is a profile in which word occurrences and co-occurrences are counted. The profile is currently used to identify compound nominals and to save any "learned" information. Under investigation is whether the profile can provide clues as to the identification of main topics or selection of high content sections of the text where further processing might be focussed. (See section on cluster analysis.)

During pass 2, verb and noun phrases are identified and entities and relationships are established. The partial parsing step disambiguates the parts of speech and builds noun and verb phrases. The information linker is experimental at this point, but its goal is to provide the information in a database compatible format. If good progress can be made in moving entity relationships to a database-like storage, it may be reasonable to replace the current partial parser with a purchased parser with higher precision and clause analysis and an associated extensive lexicon.

Relationships can be established through pattern matcher rules, through subject, verb, object or complement attachment or simply through proximity in the cases of time, location or object of a prepositional phrase.

A small set of pattern matcher rules is activated during pass 1 to facilitate the handling of unknown words and the handling of special patterns. A larger set of pattern matcher rules is activated during pass 2 that help in linking events with entities or entities with entities.

Profiling

Profiling is done during the first pass of document processing. This pass counts the non-functional words. When preceded or followed by a non-functional word, the count on this co-occurrence is also recorded.

In the output the original word is followed by a number indicating the number of times this word occurred in the document and then the semantic net category names of this word. Indented on the next lines are occurrences of the word in pairs and the number of times that pair occurred. The small section of output that follows was the result of a run on 180K document containing a series on small articles on the baseball strike. A sample article is shown in Appendix A in the cluster analysis discussion. "Antitrust exemption", as will be explained later, will be added as a compound because of the high frequency of the pair. 'April' is marked 'active' because it is a trigger word for a pattern in the pattern matcher. 'Antitrust', 'anaheim' and 'appropriate' have no categories listed as they are not found in the noun/verb modified semantic net.

american	21	PERSON	COMMUNICATION
		american league	8
		american people	3
amount	7	COGNITION	ATTRIBUTE POSSESSION
anaheim	16	ACTIVE	
		anaheim stadium	8
analysis	2	COGNITION	COMMUNICATION ACT

angel	9	PERSON	OBJECT
announcement	4	COMMUNICATION	
annual	6	COMMUNICATION	
anticipation	2	COGNITION	FEELING STATE
antitrust	20		
		baseball's antitrust	4
		antitrust exemption	12
COMPOUND ADDED: antitrust_exemption. <ACT STATE exemption>			
apartment	3	ARTIFACT	
approach	3	ACT	ARTIFACT
appropriate	2		
april	5	ACTIVE	<DATE PERSON TIME April>
arbitration	29	COMMUNICATION	ACT
		salary arbitration	6
		binding arbitration	8
		arbitration award	3
arbitrator	3	PERSON	

As will be discussed in following sections, experiments suggests that the profile information could be used (1) with a second measure to identify key or important words, and (2) to identify compound nominals such as 'cruise missile' which should be treated as a parsed unit.

Keyword identification

"Important" words in a document may exhibit two measurable characteristics: (1) the word occurs *infrequently* in the language as a whole and, conversely, (2) the word occurs *frequently* in this particular document. The Brown Corpus and others lists common English words along with their frequency of occurrence over a large document set. With WordNet already integrated into the system, it was hypothesized that a word's frequency in the language would likely be proportional its number of semantic senses, or *breadth b*, in WordNet. 'Block' in WordNet, for example, has 12 semantic senses while 'ballpark' has only one. The frequency of occurrence, *f*, is simply a count of the number of times a particular word appears in the document and is returned in the profile.

A few experiments were run to see if the higher content words would have higher *f/b* (frequency in document divided by breadth in WordNet) scores. In a composite of articles on the baseball strike the following words were the highest *f/b* scorers in a possible field of 918 nouns.

baseball	freq = 149	breadth = 2	f/b = 74.5	*****
bet	freq = 22	breadth = 1	f/b = 22.0	**
clinton	freq = 45	breadth = 1	f/b = 45.0	****
game	freq = 191	breadth = 3	f/b = 63.7	*****
income	freq = 28	breadth = 1	f/b = 28.0	**
major_league	freq = 81	breadth = 1	f/b = 81.0	*****
national	freq = 22	breadth = 1	f/b = 22.0	**
negotiation	freq = 23	breadth = 1	f/b = 23.0	**
negotiator	freq = 28	breadth = 1	f/b = 28.0	**
owner	freq = 222	breadth = 2	f/b = 111.0	*****
people	freq = 30	breadth = 1	f/b = 30.0	***
photo	freq = 20	breadth = 1	f/b = 20.0	**
player	freq = 240	breadth = 3	f/b = 80.0	*****
revenue	freq = 46	breadth = 2	f/b = 23.0	**
salary	freq = 70	breadth = 1	f/b = 70.0	*****
season	freq = 80	breadth = 3	f/b = 26.7	**

selig	freq = 32	breadth = 1	f/b = 32.0	***
stadium	freq = 35	breadth = 1	f/b = 35.0	***
strike	freq = 183	breadth = 6	f/b = 30.5	***
team	freq = 58	breadth = 2	f/b = 29.0	**
total	freq = 62	breadth = 2	f/b = 31.0	***
week	freq = 32	breadth = 1	f/b = 32.0	***
year	freq = 68	breadth = 3	f/b = 22.7	**

These results look encouraging as the top six scorers (baseball, game, major_league, owner, player and salary) and many of the others are intuitively words that one would expect to see amongst its keywords.

In order to see the real contribution of the breadth component b , however, one needs to see the top performers using only the frequency f . Using the same document set as above the following list show the top f scorers.

baseball	freq = 149	breadth = 2	f/b = 74.5	*****
business	freq = 32	breadth = 9	f/b = 3.6	
clinton	freq = 45	breadth = 1	f/b = 45.0	****
day	freq = 51	breadth = 4	f/b = 12.8	*
development	freq = 36	breadth = 7	f/b = 5.1	
game	freq = 191	breadth = 3	f/b = 63.7	*****
key	freq = 35	breadth = 7	f/b = 5.0	
major_league	freq = 81	breadth = 1	f/b = 81.0	*****
owner	freq = 222	breadth = 2	f/b = 111.0	*****
president	freq = 38	breadth = 4	f/b = 9.5	
quote	freq = 32	breadth = 2	f/b = 16.0	*
revenue	freq = 46	breadth = 2	f/b = 23.0	**
salary	freq = 70	breadth = 1	f/b = 70.0	*****
season	freq = 80	breadth = 3	f/b = 26.7	**
selig	freq = 32	breadth = 1	f/b = 32.0	***
side	freq = 49	breadth = 9	f/b = 5.4	
stadium	freq = 35	breadth = 1	f/b = 35.0	***
strike	freq = 183	breadth = 6	f/b = 30.5	***
team	freq = 58	breadth = 2	f/b = 29.0	**
time	freq = 57	breadth = 9	f/b = 6.3	
total	freq = 62	breadth = 2	f/b = 31.0	***
union	freq = 64	breadth = 11	f/b = 5.8	
week	freq = 32	breadth = 1	f/b = 32.0	***
year	freq = 68	breadth = 3	f/b = 22.7	**

Comparing the two lists, one can see that very top scorers are in both lists. Included in the f only list and eliminated in the f/b list were the words: *business, day, development, key, president, quote, side, time, and union*. All of the words which were included in the f/b list which did not make the f only list (*bet, income, national, negotiation, negotiator, people and photo*) were words with a single semantic sense which moved up in the list by virtue of others being moved down.

One might hope that the f/b scores, or possibly the f only list, could serve as a guide either toward focussing further processing toward high content areas or toward establishing priorities on what information should be extracted in the templates.

Cluster Analysis

While f/b as a ratio looks promising as a measure, cluster analysis provides an

alternative method of combining the measures of f and b . For example, the f/b ratio would treat the (f,b) combinations (4,2) and (20,10) the same, while cluster analysis would treat them differently.

Cluster analysis is an exploratory statistical procedure that is often helpful in understanding the complex nature of multivariate relationships. It consists of graphical techniques (dendograms) and step-by-step rules (clustering algorithms) for grouping "similar" objects which can be either variables or items. Searching the data for a structure of "natural" groupings is an important exploratory technique. In the context of extracting information from text, the clustering technique is being considered as a tool for grouping words appearing in text that are the most important (as well as grouping together unimportant words). Having identified the most important words (and unimportant words) in a given text, along with other known attributes of the words (such as part of speech or semantic category), one can attempt to piece together information content in the form of entity relationships.

Cluster analysis makes no assumption concerning the number of groups or the group structure. Grouping is done on the basis of similarities or "distances". The inputs required to the clustering algorithm are the distances between items, or data from which distances can be computed. The usefulness of the grouping that results depends on the appropriateness of the distance or similarity measure. In our application, a "distance" between words appearing in an article is defined in terms of the difference in frequency of occurrence of the words in the article, and difference in the breadth of the words in the WordNet. The cluster analysis technique is being explored as a way to group words according to these two measures, to identify those words with high frequency of occurrence and low breadth. An example of the clustering technique is given in the appendix.

Nominal Compounds

In addition to the straight word count, the profiler counts the number of times a word occurs adjacent to another word. This co-occurent frequency is the basis for building the first set of compound nominals discussed in this section. Identifying these compounds improves sentence parses and consistency in entities by always treating these word pairs as units.

Using a 750-line section of a chemical weapons treaty, forty-two compounds were identified based on profile co-occurrence in the document. Extraction depends on either of the words being used with the adjoining (preceding or succeeding) word at least a specific number of times in its total usage in the document. The specified number of times was based on the formula $y = 1/4 x + 2$ where x is the total usage in the document and y is the number of times it occurred with this word. This formula allows for larger documents requiring lower ratios of total occurrences to still meet the cutoff criteria. This formula was selected based on a *limited* sample of experiments and hand comparing actual results against desired, subjective results. Compound nominals were not built if either of the two words were function words such as determiners, pronouns, auxiliaries verbs or modals, prepositions, or conjunctions. The semantic categories of the last word is assigned to the unit.

Looking at the following results, notice that most of the generated compounds do make semantic sense ('verification activities', 'on-site inspection', and 'Geneva Protocol'). There are a few that are questionable such as 'due regard', and 'set forth'. It is likely that these could be refined if, in fact, the building of them as compounds causes any

difficulties in meaning extraction.

verification_activities	11	<ACT STATE ATTRIBUTE activity>
scientific_advisory_board	4	<GROUP SUBSTANCE FOOD ARTIFACT board>
verification_annex	52	<ARTIFACT annex>
annual_destruction	4	<ACT EVENT destruction>
cd/1173_appendix	29	<COMMUNICAT BODY appendix>
national_authority	3	<GROUP PERSON STATE ATTRIBUTE RELATION authority>
available_information	3	<COGNITION COMMUNICAT information>
chemical_weapons	102	<ARTIFACT weapon>
national_chemical_industry	5	<GROUP ACT ATTRIBUTE industry>
key_component	3	<COGNITION EVENT ARTIFACT RELATION component>
confidentiality_annex	3	<ARTIFACT annex>
due_regard	3	<COGNITION ACT FEELING STATE gaze>
production_facilities	20	<COGNITION ARTIFACT ATTRIBUTE facility>
production_facility	11	<COGNITION ARTIFACT ATTRIBUTE facility>
regional_factors	4	<QUANTITY EVENT RELATION factor>
first_session	3	<GROUP TIME COMMUNICAT session>
voluntary_fund	3	<POSSESSION fund>
geneva_protocol	3	<COGNITION protocol>
regional_group	5	<COGNITION SUBSTANCE group>
highest_priority	3	<STATE RELATION priority>
on-site_inspection	8	<ACT inspection>
on-site_instruments	6	<GROUP PERSON COMMUNICAT ACT ARTIFACT instrument>
part_iv	16	
limit_verification	3	<COGNITION confirmation>
multilateral_agreement	6	<COGNIT COMMUNIC STATE POSSESS RELATION agreement>
nameplate_capacity	3	<COGNIT ACT QUANTITY STATE ATTRIB PHENOM capacity>
old_chemical	6	<SUBSTANCE chemical>
part_v	14	
states_parties	62	<GROUP PERSON party>
state_party	76	<GROUP PERSON party>
peaceful_purposes	3	<COGNITION MOTIVE ATTRIBUTE function>
weapons_production	31	<COMMUNICAT ACT ARTIFACT production>
toxic_properties	3	<LOCAT COGNIT ARTIFACT ATTRIB POSSESS property>
united_nations_scale	3	<COMMUNICAT ANIMAL ARTIFACT PLANT RELAT scale>
technical_secretariat	22	<GROUP secretariat>
set_forth	8	
significant_national	5	<PERSON national>
special_session	4	<GROUP TIME COMMUNICAT session>
systematic_verification	7	<COGNITION confirmation>
technological_development	4	<LOCATION ACT EVENT PROCESS RELAT exploitation>
article_viii	4	
article_x	5	

Nine additional compounds were extracted from the same treaty section based on occurrence in the WordNet. One benefit of finding the compound in WordNet is that the number of associated semantic meanings has now been substantially reduced. 'Board', for example, has 7 senses attached to it in WordNet. 'Advisory board', on the other hand, has only one attached meaning and it is under the major category 'group'.

advisory_board	4	GROUP
chemical_industry	6	GROUP
chemical_reaction	2	PROCESS
executive_council	57	GROUP
general_assembly	2	GROUP
international_law	3	COMMUNICAT
latin_america	2	LOCATION

riot_control	5	ACT
united_nations	9	GROUP

Although not implemented at this time, processing could favor this semantic meaning of 'board' in nearby sections of the text. Using co-occurrence and a semantic network in this way could also contribute to automatically solving the semantic keyword expansion problem found in text document retrieval applications [DES95] by refining the relevant semantic sense before expansion

Unknown Words

Finding a compound nominal in WordNet allows the refinement of the semantic sense relevant to this usage where the individual words may have each had several senses. The opposite problem occurs when a word is not in the semantic net at all. This section discusses some algorithms for "guessing" its semantic category. Most of these techniques depend on the words immediate preceding or following the unknown word. In implementation many of these depend on the unit of code labelled the 'pattern matcher' in Figure 1.

Probably the largest category of unknown words in text are found as part of a people's names. A list of approximately 7500 first names (public domain) were merged to the data obtained from WordNet. If a capitalized, unknown word (or capitalized initial plus a capitalized, unknown word) follows a capitalized "first name", then it too is marked category "PERSON" for consecutive usage.

Through the pattern matcher, unknown words or compounds can also be "learned". Mrs. X, Dr. Y, and 3 mg of Z are examples where unknown words X, Y, and Z can be assigned attributes <PERSON female>, <PERSON doctor> and <SUBSTANCE drug> through the pattern matcher.

Compounds relevant to specific domains can also be learned this way. If being used for baseball articles, for example, patterns for baseball teams could be added:

toronto blue_jays: >1-L@14#03935679; baseball team
 baltimore_orioles : >1-L@14#03935679; baseball team
 san_diego padres: >1-L@14#03935679; baseball team
 cleveland indians: >1-L@14#03935679; baseball team
 new_york yankees: >1-L@14#03935679; baseball team
 detroit tigers: >1-L@14#03935679; baseball team

Because they are added through the pattern matcher, a recompilation of code is not necessary to activate.

There are a number of rules currently in the pattern matcher that are non-domain dependent and key off of words such as 'which', 'that' and 'of the'. This makes for easy, fast development in trying out new attachments or ideas. The caveat, however, is that when multiple patterns match to the same sentence fragment unexpected results often occur. These grammar level, non-domain specific rules, once tested, should probably be moved to code.

Example paragraph:

Mrs. Tannager is currently on 50 mg of Tenormin per day and has been on that for about two weeks. She has noticed no significant improvement in her headaches. The Tenormin was used in conjunction with the recommendations

made by Dr. Kilder.

Profile:

conjunction	1	COMMUNIC EVENT STATE ARTIFACT ATTRIB RELATION
day	1	TIME
headaches	1	COGNITION STATE
improvement	1	ACT EVENT RELATION
kilder	1	ACTIVE <PERSON doctor>
made	1	ACT
recommendations	1	COMMUNICATION
tannager	1	ACTIVE <PERSON female>
tenormin	2	ACTIVE <SUBSTANCE>
weeks	1	TIME

Notice in the above profile that 'kilder', 'tannager' and 'tenormin' are marked 'ACTIVE <PERSON doctor>', 'ACTIVE <PERSON female>' and 'ACTIVE <SUBSTANCE>' respectively. These three words were not contained in the semantic database. The information assigned to them was the result of the execution of a matching pattern. The clue to these assignments came from the preceding words 'Dr.', 'Mrs.', and '# mg of'.

The full set of rules are read in during pass 1. Trigger words are marked and saved for fast matching as is the rule itself. The rule has 4 components: the match component, the delimiter ':>', the execute component, and an optional comment field beginning with a semicolon. The three simple rules executed in this case are described below.

Syntax of Rule: match_component ':>' execute_component [; comment_field]

Rule 1: %t17 mg of %t19: >4@27; 25(num) mg of morphine(singn) {substance}

Match component: %t17 mg of %t19

Explanation of match: %t17 = match must have type 'number'.

mg = exact string match.

of = exact string match.

%t19 = match must have type 'singular noun'

Text match: 50 mg of Tenormin

Execute component: 4@27

Explanation of execution: assign the 4th element ('Tenormin') the category type 'SUBSTANCE' (27 ==> substance).

Rule 2: dr. %t: >1-2@18#04552709; Dr. x {doctor}

Match component: dr. %t

Explanation of match: dr. = exact string match

%t = word of any type

Text match: Dr. Kilder

Execute component: '1-2@18#04552709'

Explanation of execution: assign first 2 elements the type 'PERSON' (18==>person), subtype 'doctor' (04552709=doctor) and build unit as a single noun phrase.

Rule 3: mrs. %t: >1-2@18#04413284; Mrs. Bolland {female}

Match component: mrs. %t

Explanation of match: mrs. = exact string match

%t = word of any type

Text match: Mrs. Tannager

Execute component: 1-2@18#04413284
Explanation of execution: assign first 2 elements the type 'PERSON'
(18==>person), subtype female (04413284=female) and build
unit as a single noun phrase.

An issue that has not yet been addressed is for what period of time the "learned" knowledge should remain active. Since knowledge is learned during the first pass and used during the second pass, scoping the knowledge for particular sections becomes complicated. Text could contain references to "Mrs. Tannager" and her son "Adam Tannager" causing "Tannager" to be both male and female.

Intra-sentence Relationships

Having identified entities and events, the next task is to understand the relationships between them. Unification takes place due to one of four reasons. (1) A pattern was matched and an attachment was made as specified in its rule. (2) An entity occurs as the subject, object, or complement of a verb event. (3) Proximity of a time or location is assigned to the nearest event. (4) Prepositional phrases are usually attached to the preceding noun or verb phrase.

The preceding three-sentence paragraph is represented here using this sequence. Indentation implies attachment to preceding (less indented) entity or event.

Sentence 1:

0: PERSON: Mrs. Tannager
0: (~~): female
0: SUBSTANCE(OF): Tenormin
0: QUANTITY: 50 mg
0: TIME: per day
0: TIME(PREP:about): two weeks

Sentence 2:

1: FEELING: noticed
1: ACT(OBJECT): no significant improvement
1: COGNITION(PREP:in): her headaches
1: PERSON(ACTOR:noticed): Mrs. Tannager

Sentence 3:

2: COMMUNICATION: the recommendations
2: ACT: made
2: PERSON(PREP:by): Dr. Hideto
2: (~~): doctor
2: SUBSTANCE: The Tenormin
2: PROCESS(==): used

Inter-sentence Unification

An attempt was made to extend the intra-sentence approach to do inter-sentence unification in the same structure. Although good results often occurred in the matching of entities to previous references, the approach didn't seem adequate for its major intended use which was to unify the entities with its events. Grouping the information across sentence boundaries in this way was error prone and unwieldy. A current effort is

exploring whether outputting data in a generic format would be useful. If an alternate approach such as this looks promising, an attempt will be made to reimplement the matching of references aspect of the earlier attempt.

Successful systems to date have overcome some of the inter-sentence unification problem by building the expected frames ahead of time. Fillers for each of the roles in the frame also have expected constraints. For example, at MUC-4 it was known at the outset that the item of interest was a terrorist event and that pieces of desired information included perpetrator, target, effect, instrument, date, location, etc. The 'date' role could be then constrained to be filled with elements such as day, month, year. The 'location' role might only be filled with country or city names; the 'perpetrator' role might favor terrorist groups; target might favor a 'human' filler; etc. "Knowledge" can then be built up, a priori, listing known terrorist organizations or instruments that terrorists might use in an attack. Knowing the names of countries and terrorist organizations is useful information and more on-line information is becoming available all the time.

In the above example if a frame with constraints on the role fillers were set up ahead of time then the goal could be to fill in the slots as follows:

Patient/Doctor Hypothesized Script (not implemented):

Patient_name(person): Mrs. Tannager

Drug_name(drug): Tenormin

Dosage(quantity): 50 mg

Frequency(time) : per day

Duration(time) : two weeks

Symptom / Illness(symptom / disease): headache

Effect(?): no significant improvement

Doctor_name(person / doctor): Dr. Kilder

What this LDRD is still grappling with however is how to organize the information into a frame or other format in a domain-independent way that still makes sense.

Entity-Relationship Tables

Moving information in natural language to a database or a template-with-roles format is a problem that even humans do not do using a known, repeatable algorithm. Language provides many ways of saying the same thing. Grammatical patterns that syntactically look the same often have different meanings due to the knowledge that humans bring to a situation. Work in this section is in an idea generation stage. Two examples are shown. The first table contains entity/entity relationships that might target the question, "Who is X?" The second table is event (or verb) driven and shows entity attachments.

Identifying "equalities"

The baseball text depended heavily on appositives that identified a person with a position. The number followed by a colon is the sentence number in which the information was obtained.

Sentence Fragments Providing the Information

- 19: Daniel Silverman, regional director of the NLRB's New York office, said . . .
21: Union leader Donald Fehr said he would . . .
28: However, Dodger owner Peter O'Malley said he . . .
31: Bud Selig, baseball's acting commissioner; Jerry McMorris, owner of the Colorado Rockies, and John Harrington, Boston Red Sox chairman, will meet with Fehr and his staff.
37: Sources said the owners are still tied to the proposals made by special mediator William J. Usery . . .
39: Union lawyer Eugene Orza said it is still the owners' goal to . . .
44: "This we believe . . . fairly," general counsel Fred Feinstein said.
52: While the clock ticked toward the season opener, Chicago White Sox owner Jerry Reinsdorf launched another verbal attack at Fehr, comparing him to the late cult leader Jim Jones.

Information Automatically Extracted in a Structured Format

Person	is a	Position
-----	-----	-----
19: Daniel Silverman	is/was	regional director of the NLRB's NY office
21: Donald Fehr	is/was	Union leader
28: Peter O'Malley	is/was	Dodger owner
31: Jerry McMorris	is/was	owner of the Colorado Rockies
31: John Harrington	is/was	Boston Red Sox chairman
31: Bud Selig	is/was	baseball's acting commissioner
37: William J. Usery	is/was	special mediator
39: Eugene Orza	is/was	Union lawyer
44: Fred Feinstein	is/was	general counsel
52: Jim Jones	is/was	the late cult leader
52: Jerry Reinsdorf	is/was	Chicago White Sox owner

Identifying "events"

Some semantic categories are event types to which object types through parsing can be connected or related. In a communication, for example, relevant roles might include the communicator, the recipient, what was communicated, and when and where the event took place. The current state of information can be output in a tablelike fashion. Remember that the second dimension to the table is the semantic category of each unit, though the "role" is still unknown except in a general way.

Example paragraph (slightly enriched from previous example):

Mrs. Tannager returns for a recheck. She continues to take the 20 mg of Feldene per day. Mrs. Tannager is also on 50 mg of Tenormin per day and has been on that for about two weeks. She has noticed no significant improvement in her headaches. The Tenormin was used in conjunction with the recommendations made by Dr. Kilder. I would try and avoid narcotics in this patient in the future, if at all possible.

SN: Agent	Action	Object
---	-----	-----
0: Mrs. Tannager	returns	
1: ---	take	Feldene the 20 mg per day
2: Mrs. Tannager	---	Tenormin 50 mg per day two weeks
3: Mrs. Tannager	noticed	no significant improvement her headaches
4: Dr. Kilder	made	the recommendations
5: ---	avoid	narcotics this patient the future

Longer example

This longer example from a treaty shows both the equality and event relationships.

1. The term "ballistic missile" means a missile that has a ballistic trajectory over most of its flight path. The term "ground-launched ballistic missile (GLBM)" means a ground-launched ballistic missile that is a weapon-delivery vehicle.

2. The term "cruise missile" means an unmanned, self-propelled vehicle that sustains flight through the use of aerodynamic lift over most of its flight path. The term "ground-launched cruise missile (GLCM)" means a ground-launched cruise missile that is a weapon-delivery vehicle.

1. If a ballistic missile or a cruise missile has been flight-tested or deployed for weapon delivery, all missiles of that type shall be considered to be weapon-delivery vehicles.

2. If a GLBM or GLCM is an intermediate-range missile, all GLBMs or GLCMs of that type shall be considered to be intermediate-range missiles. If a GLBM or GLCM is a shorter-range missile, all GLBMs or GLCMs of that type shall be considered to be shorter-range missiles.

GLBMs or GLCMs that have a range capability equal to or in excess of 500 kilometers but not in excess of 1000 kilometers shall be considered to be shorter-range missiles. GLBMs or GLCMs that have a range capability in excess of 1000 kilometers but not in excess of 5500 kilometers shall be considered to be intermediate-range missiles.

Results of equality relationship extraction:

SN:Object	is-a has-a	Object
-----	-----	-----
0: ballistic_missile	is/was	a missile
0: a missile	has	a ballistic_trajectory
1: GLBM	is/was	ballistic_missile
1: a ground-launched ballistic_missile	is/was	a weapon-delivery_vehicle
3: cruise_missile	is/was	an unmanned , self-propelled vehicle
4: GLCM	is/was	cruise_missile
4: a ground-launched cruise_missile	is/was	a weapon-delivery_vehicle
8: GLCM	is/was	an intermediate-range_missile
8: a GLBM or GLCM	is/was	an intermediate-range_missile
9: GLCM	is/was	a shorter-range_missile
9: a GLBM or GLCM	is/was	a shorter-range_missile

11: GLBMS or GLCMS	has	a range capability
12: GLBMS or GLCMS	has	a range capability

Results of event relationship extraction:

SN:Agent	Action	Object
0: a missile	has	a ballistic_trajectory
3: ---	sustains	flight (through) the use (of) aerodynamic lift
6: a ballistic_missile	---	a cruise_missile projectile
6: all missiles of that type	be considered	weapon-delivery_vehicles
8: all GLBMS or GLCMS of that type	be considered	intermediate-range_missiles
9: all GLBMS or GLCMS of that type	be considered	shorter-range_missiles
11: GLBMS or GLCMS	has	a range capability
11: GLBMS or GLCMS	---	excess (of) 500 kilometers
11: ---	be considered	shorter-range_missiles
11: ---	be considered	(not in) excess (of) 1000 kilometers
12: GLBMS or GLCMS	has	a range capability (in) excess (of) 1000 kilometers
12: GLBMS or GLCMS	---	be considered intermediate-range_missiles (not in) excess (of) 5500 kilometers
12: GLBMS or GLCMS	---	(in) excess (of) 1000 kilometers
12: ---	a range capability	(in) excess (of) 1000 kilometers
12: ---	be considered	intermediate-range_missiles
12: ---	be considered	(not in) excess (of) 5500 kilometers

Though the results of the second example still show holes, redundancies and some errors, most of the entities (nouns) and many of the relationships (verbs) are represented. More unification needs to be done and a fast-access index would need to sit on top before it could be used for information retrieval. Though not shown in this display, each unit still has an assigned semantic category as discussed earlier. The importance measures, such as f/b , have not yet been incorporated into the results of this section but is an anticipated future step.

Summary and Future Work

One thing that is still missing and is not clear how to accomplish is how to map agents and objects of events to a common format without domain specific or word specific knowledge. For example, the actor or agent of a 'buy' is the recipient, while the indirect object of a 'sell' is the recipient. Knowing that they are both transfer events is not enough information to answer the question, "Who purchased the book?" It might be enough information however to match to the sentence which involved the transfer or to bring up the structure containing the information about the transfer. Not having this content knowledge about the event is inhibiting progress in unifying events that occur over multiple sentences. Unifying instances of entities (e.g., unify all the references to "Mrs. Tannager") will need to be done but does not appear to present the same level of difficulties as does the frame and roles problem.

In order to expedite the research and algorithmic development pertinent specifically toward extracting information without domain knowledge and without pre-selection of

extractables, several intermediate tasks have been done in only a partial fashion. These tasks will need to be filled out (probably some through software purchases and integration) before the current program could be considered generally operational.

In summary, this project is in an experimental, idea-generation stage where some promising techniques for partial solutions have been identified. Some tools that will facilitate continued experimentation are also now in place. The solution will likely have several dimensions. This year's work concentrated on identifying key entities based on document level characteristics, identifying relationships between entities through sentence level syntax and verb occurrence, and integrating semantic knowledge. Second-year work will build on this two-pass approach with emphasis on solidifying the entity-relationship algorithms and then on the areas of data organization and query matching.

References

- [ANT90] Antworth, Evan L., "PC-KIMMO: A Two-level Processor for Morphological Analysis", Summer Institute of Linguistics, Dallas, TX, 1990.
- [CHU95] Chung, Minhwa; Moldovan, Dan I. "Parallel Natural Language Processing on a Semantic Network Array Processor", IEEE Transactions on Knowledge and Data Engineering, Vol. 7, No. 3, pp. 391-405, June 1995. (SMU, Dallas, TX)
- [DES95] Desonier, Lawrence M. "Relevance of Full Meaning Data Retrieval", Sandia National Labs, Internal Document.
- [JOH82] Johnson, R. A.; Wichern D.W. Applied Multivariate Statistical Analysis, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982.
- [KIM93] Kim, Jun-Tae; Moldovan, Dan I. "Acquisition of Semantic Patterns for Information Extraction from Corpora", IEEE 9th Conference on Artificial Intelligence for Applications. pp. 171-176. 1993. (USC, Los Angeles, CA)
- [KNI94] Knight, Kevin; Luk, Steve K. "Building a Large-Scale Knowledge Base for Machine Translation", Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94), pp. 773-778, July 1994. (USC/Information Sciences Institute, Marina del Rey, CA).
- [LEH94] Lehman, Jill Fain. "Toward the Essential Nature of Statistical Knowledge in Sense Resolution", Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), pp. 734-741. July 1994. (CMU, Pittsburgh, PA).
- [MCD94] McDonald, David D. "Trade-offs Between Syntactic and Semantic Processing in the Comprehension of Real Texts", RIAO-94, October 1994. (Brandeis Univ.)
- [RAU89] Rau, Lisa F.; Jacobs, Paul S.; Zernik, Uri. "Information Extraction and Text Summarization Using Linguistic Knowledge Acquisition", Information Processing and Management, Vol. 25, No. 4, pp. 419-428, 1989. (GE CR&D, Schenectady, NY)
- [RIL93] Riloff, Ellen. "Automatically Constructing a Dictionary for Information Extraction Tasks", Proceedings of the 11th National Conference on Artificial Intelligence, pp. 811-816, AAAI Press, Menlo Park, CA, 1993. (UMASS, Amherst, MA)
- [RIL94] Riloff, Ellen; Lehnert, Wendy. "Information Extraction as a Basis for High-Precision Text Classification", ACM Transactions on Information Systems, Vol. 12, No. 3, July 1994. pp. 296-333. (UMASS, Amherst, MA)
- [WOR93] WordNet 1.4 Copyright 1993, public domain software, Princeton University.

Appendix

Cluster Analysis

An article on baseball labor negotiations, given below, was analyzed for breadth and frequency of occurrence of each word.

Washington - On the eve of the start of spring training, bickering major league baseball players and owners brought their labor dispute to Capitol Hill on Wednesday. But instead of receiving any help, both sides got something of a tongue-lashing from irate senators.

The venue was a Senate Judiciary subcommittee hearing on proposals to repeal part of baseball's antitrust exemption, which would enable players to sue if owners impose unilateral terms that are different from the terms of the expired collective bargaining agreement.

And even though the players have said that they would end their strike if such an exemption were repealed, subcommittee members left little doubt that such action is at best remote.

"We are not going to legislate on this matter before spring training starts," said Sen. Arlen Specter (R-Pa.). "This is too complicated to have that happen. We're too busy on other matters."

Much of the three-hour hearing was taken up by lawyerly colloquy over the finer points of labor law. But the exchanges were interspersed with rambling statements from members of the subcommittee, who chastised the players and owners for everything from their intransigence to the price of tickets and hot dogs at ballparks.

"People are angry," fumed Sen. Joseph R. Biden Jr. (D-Del.), the ranking minority member of the Judiciary Committee. "Neither one of you is very popular," he said, looking out at the gathering of players, owners and their representatives.

Those attending included players Eddie Murray and David Cone; Donald Fehr, head of the Major League Baseball Players Assn.; and Bud Selig, acting baseball commissioner and president of the Milwaukee Brewers.

The f/b (frequency/breadth) ratio was computed for each word in the article. Those words with high f/b ratio included player (1.67), major_league (2.0), owner (2.0), and subcommittee (3.0). For comparison, the cluster analysis approach was applied to the same data, compiled in the following two-way table, where words from the article are tabled according to their frequency and breadth values.

Table 1. Breadth and Frequency Data from Example Article

	<u>Frequency</u>				
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>
B r e a d t h	<u>1</u>	w_1	w_6	w_{11}	.
	<u>2</u>	w_2	.	.	w_{13}
	<u>3</u>	w_3	w_7	.	w_{14}
	<u>4</u>	w_4	w_8	.	.
	<u>5</u>	.	w_9	w_{12}	.
	<u>6</u>
	<u>7</u>
	<u>8</u>
	<u>9</u>	.	w_{10}	.	.
	<u>10</u>
	<u>11</u>
	<u>12</u>	w_5	.	.	.

In the above table, w_i represents the i th word (or unique location) in the two-way table. For example, w_{10} is in Frequency column 2 and Breadth row 9. This means that the word (or words) represented by w_{10} appeared twice in the text and has 9 different senses in the WordNet. **Each w_i may represent multiple words appearing at that same location.** For example, the words "spring_training" and "major_league" both appeared twice in the article ($f=2$) and had only one sense in the WordNet ($b=1$). Both words are assigned equal importance and are represented by w_6 in the above two-way table. Cells in the table that are empty indicate that no word in the article had that combination of breadth and frequency. The objective of the cluster analysis is to cluster or group words in the table that are "close" to each other. Closeness must be defined in terms of a distance measure, which is defined below. Letting b_i represent the breadth of word w_i and f_i represent the frequency of occurrence of word w_i , then a natural distance measure between words w_i and w_j is given by

$$D(w_i, w_j) = |b_i - b_j| + |f_i - f_j|$$

This distance is simply the "city-block" distance between points in the two-way table. Using this distance measure, the following matrix of pairwise distances can be constructed.

Table 2. Initial Distance Matrix for Cluster Analysis Example

	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>
<u>1</u>	0													
<u>2</u>	1	0												
<u>3</u>	2	1	0											
<u>4</u>	3	2	1	0										
<u>5</u>	11	10	9	8	0									
<u>6</u>	1	2	3	4	12	0								
<u>7</u>	3	2	1	2	10	2	0							
<u>8</u>	4	3	2	1	9	3	1	0						
<u>9</u>	5	4	3	2	8	4	2	1	0					
<u>10</u>	9	8	7	6	4	8	6	5	4	0				
<u>11</u>	2	3	4	5	13	1	3	4	5	9	0			
<u>12</u>	6	5	4	3	9	5	3	2	1	5	4	0		
<u>13</u>	4	3	4	5	13	3	3	4	5	9	2	4	0	
<u>14</u>	6	5	4	5	13	5	3	4	5	9	2	4	2	0

The number in row i and column j of the above matrix represents the "distance," as defined above, between words w_i and w_j . For example, the 6 in row 12 and column 1 means that w_1 and w_{12} are 6 units apart. Note that the dimension of this symmetric matrix will be equal to the number of words (or word locations) included in the cluster analysis. The clustering algorithm is applied to this matrix.

Because we can rarely examine all possible groupings, a wide variety of clustering algorithms have been proposed for finding "reasonable" clusters without having to look at all configurations. The hierarchical clustering methods, which we are investigating, proceed by a series of successive mergers. Initially there are as many clusters as individual objects. Most similar ("closest") objects are first grouped, then these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster. With single linkage hierarchical clustering algorithms, groups are formed from the individual items by merging nearest neighbors, where the term "nearest neighbor" means smallest distance (largest similarity). The following are the steps in the hierarchical clustering algorithms for grouping N objects (items or variables).

Clustering Algorithm

1. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $D = \{d_{ij}\}$.
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters u and v be d_{uv} .
3. Merge clusters u and v . Label the newly formed cluster (uv) . Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters u and v and (b) adding a row and column giving the distances between cluster (uv) and the remaining clusters. (With single linkage, the distance measure is nearest neighbors, or smallest distance between items in the two clusters.)
4. Repeat Steps 2 and 3 a total of $N-1$ times. (All objects will be in a single cluster at termination of the algorithm.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

The first few steps of this algorithm are demonstrated below, applied to the distance matrix above.

In Step 1, w_1 and w_2 are grouped because they have the minimum distance (1) in the original distance matrix. The resulting distance matrix (applying the above algorithm) is shown in Table 3.

Table 3. Distance Matrix after Step 1 of Clustering Algorithm

	<u>(1,2)</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>
<u>(1,2)</u>	0												
<u>3</u>	1	0											
<u>4</u>	2	1	0										
<u>5</u>	10	9	8	0									
<u>6</u>	1	3	4	12	0								
<u>7</u>	2	1	2	10	2	0							
<u>8</u>	3	2	1	9	3	1	0						
<u>9</u>	4	3	2	8	4	2	1	0					
<u>10</u>	8	7	6	4	8	6	5	4	0				
<u>11</u>	2	4	5	13	1	3	4	5	9	0			
<u>12</u>	5	4	3	9	5	3	2	1	5	4	0		
<u>13</u>	3	4	5	13	3	3	4	5	9	2	4	0	
<u>14</u>	5	4	5	13	5	3	4	5	9	2	4	2	0

In Step 2, the (w_1, w_2) cluster is grouped with w_3 because they have the minimum distance (1) in the new distance matrix. The resulting distance matrix is shown in Table 4.

Table 4. Distance Matrix after Step 2 of Clustering Algorithm

	<u>(1,2,3)</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	<u>13</u>	<u>14</u>
<u>(1,2,3)</u>	0											
<u>4</u>	1	0										
<u>5</u>	9	8	0									
<u>6</u>	1	4	12	0								
<u>7</u>	1	2	10	2	0							
<u>8</u>	2	1	9	3	1	0						
<u>9</u>	3	2	8	4	2	1	0					
<u>10</u>	7	6	4	8	6	5	4	0				
<u>11</u>	2	5	13	1	3	4	5	9	0			
<u>12</u>	4	3	9	5	3	2	1	5	4	0		
<u>13</u>	3	5	13	3	3	4	5	9	2	4	0	
<u>14</u>	4	5	13	5	3	4	5	9	2	4	2	0

Continuing in this fashion, Step 10 produces the distance matrix in Table 5.

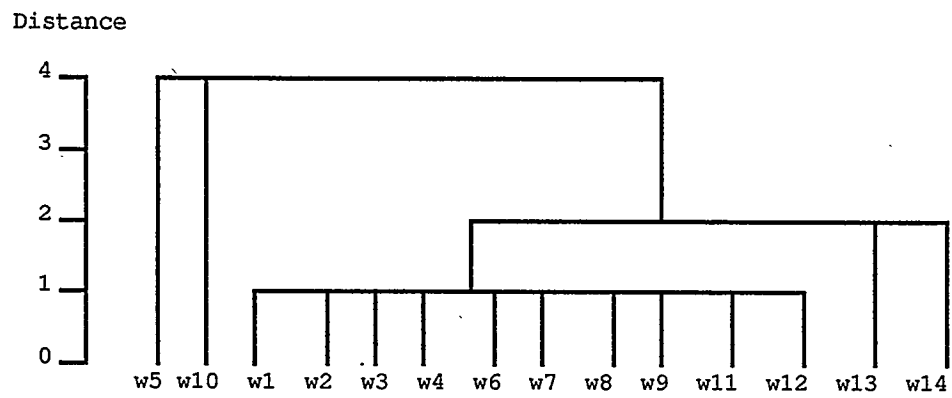
Table 5. Distance Matrix after Step 10 of Clustering Algorithm

	<u>(1,2,3,4,6,7,8,9,11,12)</u>	<u>(13,14)</u>	<u>5</u>	<u>10</u>
<u>(1,2,3,4,6,7,8,9,11,12)</u>	0			
<u>(13,14)</u>	2	0		
<u>5</u>	8	13	0	
<u>10</u>	4	9	4	0

At this stage, words $w_1, w_2, w_3, w_4, w_6, w_7, w_8, w_9, w_{11}$, and w_{12} all form a single cluster joined at distance 1. Words w_{13} and w_{14} form a cluster joined at distance 2, while words w_5 and w_{10} are still individual clusters.

In the following steps, the cluster (w_{13}, w_{14}) also joins the cluster $(w_1, w_2, w_3, w_4, w_6, w_7, w_8, w_9, w_{11}, w_{12})$ at distance 2. Words w_5 and w_{10} form a cluster joined at distance 4, which also joins with the other clusters at distance 4. The resulting dendrogram (graphical representation of the cluster analysis results) is given in Figure 2. Note that at distance 1, there are five distinct clusters, at distance 2 and 3 there are three distinct clusters, and at distance 4 all the data form a single cluster.

Figure 2. Dendrogram from Cluster Analysis Example



For more details on cluster analysis, see [JOH82].

DISTRIBUTION:

1	MS 0320	Chuck Meyers, 1011
1	0661	Olin Bray, 4503
1	0496	John Taylor, 5006
1	0971	William Cook, 9202
5	0971	Sharon Deland, 9202
1	0451	Sharon Fletcher, 9411
8	0451	Nancy Irwin, 9411
5	0829	Stephen Crowder, 12323
1	9018	Central Technical Files, 8523-2
5	0899	Technical Library, 4414
1	0619	Print Media, 12615
2	0100	Document Processing, 7613-2
		For DOE/OSTI

