Semantic Indexing and Typed Hyperlinking

J. Pustejovsky[†], B. Boguraev[‡], M. Verhagen[†], P. Buitelaar[†], M. Johnston[§]

[†] Department of Computer Science, Brandeis University
Computer Science Department, Waltham, MA 02254-9110, {jamesp,marc,paulb}@cs.brandeis.edu
[‡] Intelligent Systems, Apple Computer, Cupertino, California, bkb@apple.com
[§] Center for Human Computer Communication, Oregon Graduate Institute,
Portland, Oregon 97291-1000, USA, johnston@cse.ogi.edu

Abstract

In this paper, we describe linguistically sophisticated tools for the automatic annotation and navigation of on-line documents. Creation of these tools relies on research into finite-state technologies for the design and development of lexically-intensive semantic indexing, shallow semantic understanding, and content abstraction techniques for texts. These tools utilize robust language processing techniques to generate multi-purpose data structures called LEXICAL WEBS, used in the system TEXTRACT, an automated semantic indexing program designed to parse, index, and hyperlink electronic documents.

Introduction

In this paper, we describe the construction of LEXICAL WEBS. Lexical webs are normalized structured representations of the core semantic content of a text. The creation of lexical webs accelerates the availability of information in a highly indexed and cross-referenced format, facilitating navigation and rapid information access. One result of this research is the development of TEXTRACT, an automated semantic indexing program designed to parse, index, and hyperlink on-line documents.

Central to this work is an integration of lexical semantic techniques with corpus-based approaches to lexical acquisition. Characterizing the statistical properties of texts that have been annotated with semantic tags permits inference of semantic relations from unmarked texts that are richer than mere collocational associations. The present work is directed towards techniques for automatically enriching the treatment of web-based documents by making use of and extending these technologies.

This work emerges directly from our current reasearch on the development of an intelligent and trainable lexical database for a substantial fragment of English, called CORELEX, a lexicon that is structured in such a way that it reflects the lexical semantics of a language in systematic and predictable ways as expressed by syntactic structure. These assumptions are fundamentally different from other lexical semantic databases like Roget and WORDNET which do not account for any regularities between senses and do not relate lexical semantic information to syntactic structure.

Our claim is that semantic tagging and lexical acquisition are the central components needed to derive the content abstractions necessary for supporting more efficient tools for navigating on-line documents.

underlying utilizing a

Our research takes advantage of the fact that the lexical representation system being employed, Generative Lexicon theory, has developed many valuable techniques in the context of work within the TIPSTER and MUC efforts; namely, empirically-driven semantic techniques for lexically-based document enrichment in the service of information extraction tasks. The development of Generative Lexicon theory within *Core Lexical Engine* collaboration with Apple Computer has further focused on refining our lexical acquisition techniques for robust knowledge extraction.

Attainable Language Technology for the World Wide Web

The current state of the art in natural language processing does not provide the necessary functionality for determining a complete understanding of the semantic content of a text. It is not possible to robustly parse unrestricted text, and furthermore, even if a parse can be derived for a given utterance it is not yet feasible to consistently derive an adequate semantic representation from that parse. Given these constraints on the language processing problem, the challenge we face is how to derive useful technologies from the current state of the art.

Our strategy is to push the state-of-the-art in natural language engineering by enhancing the application of finite state technologies with semantically-driven phrasal analysis. Our goal is to employ lexically intensive linguistic processing to documents in order to derive a shallow understanding of their content. This involves assigning linguistic features to words in the text, identifying their grouping into phrases, and establishing semantic dependencies between phrases. This analysis is used to construct a network which enumerates, rank-orders, and types the entities discussed in a document, and captures the relationships between them.

The data structure that we derive is a semi-instantiated 'semantic network'. It is not a semantic network in the traditional AI sense, but rather a network of objects that we have determined to be semantically significant in the domain — a network of representative concepts, which we call a LEXICAL WEB. It is semantic in that the connections between concepts are semantic in nature, while still corresponding to the linguistic structure as they appeared in the text.

Although this approach does not provide a full semantic interpretation, it is extremely useful as an abstraction of the content of the text from which it was derived. Moreover, by relating it to the original document source, the user is offered a powerful prompting device as an enhancement to their navigation and search capabilities. HyperText, specifically HTML in the context of the World Wide Web, provides a natural delivery vehicle for the results of this technology.

Comparison to TIPSTER/MUC

On one view, our approach constitutes a scalable, domain independent extension of work done in the context of the MUC and TIPSTER research programs (cf. Cowie *et al* 1993, Guthrie *et al* 1996, Lehnert and Sundheim 1991).

The MUC and TIPSTER extraction exercises involved identification of shallow syntactic objects grounded in simple lexical information. These objects are mapped onto a template which provides a semantic description of a very narrow domain, in essence, giving a very basic form of semantic typing. Although these are perfectly legitimate models of specialized domains, when the information is derived from the web and other sources, there may be no predetermined selection of the domain. In other words, once one leaves the narrow domain of a sublanguage, the semantics is no longer so restricted, and hence the limited phrasal processing and impoverished lexical information driving these approaches is insufficient.

Our approach is different in two important ways. First of all, we have a different target. We try to generate something in the *direction* of a semantic interpretation, but not a full interpretation. We generate a reduced abstraction of the content of a document and then leave it to the user to apply their own model of context in order to construct an interpretation. The use of HyperText is an essential element of our approach, as it enables delivery of the results of our analysis to the user.

In order to perform the domain independent typing needed for such applications, it is necessary to exploit the particular syntactic expression and local context of text objects in a document. The nature of the linguistic expression is something that is shared among all textual documents and therefore can be used independently of a specific domain.

Once phrases describing particular objects have been identified, connectivity between those phrases can be established in a number of ways. Their interconnection across the discourse may be captured (anaphoric relations, discourse relations). The appearance of phrases in constructions with other phrases can be utilized in order to identify local dependencies. This also yields information regarding the properties of an object and the actions and events it is involved in. The dependencies between a phrase and other phrases, and properties and relations in the text enable a form of domain independent typing to be established.

The rest of this paper is structured as follows. First, we will describe the general theoretical framework underlying the focus of this research. We will then describe our work on the development of CORELEX and the lexical indexing system TEXTRACT-1, which together provide the basic infrastructure and tools that we apply to the task of automatic and machine-aided enhancement of texts. We then elaborate on the notion of LEXICAL WEB, which is an extension of our current indexing work, and describe how lexical webs can be acquired from texts, underpinning a richer, more sophisticated indexing system, TEXTRACT-2, and some initial experiments in lexical indexing.

Lexical Knowledge Representation

Since the lexical database in the Core Lexical Engine is an embodiment of Generative Lexicon theory (GL), we outline those aspects of the semantic representation language which are essential for the techniques of document enhancement outlined below. One of the central theses of GL is that the underlying meaning of an expression determines, in a systematic way, its syntactic expressiveness in the language. Another major goal of this work is to explain how words take on new meanings in novel contexts, while still being able to treating the core senses in a finite fashion. Important for our concerns in the present research is to acknowledge that lexical semantics is both an expressive and highly structured knowledge base, incorporating knowledge which linguists rarely associate with words (cf. Boguraev, 1992). This includes information relating, for example, not only what an object is, sortally, but also what it is used for, how it comes about, and what it is made of; these are termed an object's *qualia structure* (cf. Pustejovsky, 1991). Furthermore, words are also encoded with fairly rich information regarding the expression of arguments of various sorts, its *argument structure* (cf. Grimshaw, 1990), and a fine-grained encoding of how a word denotes events, i.e., its *event structure* (cf. Moens and Steedman, 1988, Passonneau, 1988). For more details on the theoretical model, see Pustejovsky (1995).

Generative Lexicon theory assumes that word meaning is structured on the basis of four generative factors, or *qualia roles*, capturing ways in which humans understand objects and relations in the world. They are: (1) CONSTITUTIVE: the relation between an object and its constituent parts; (2) FORMAL: that which distinguishes it within a larger domain; (3) TELIC: its purpose and function; and (4) AGENTIVE: factors involved in its origin or "bringing it about".

The qualia structure provides the structural template over which semantic transformations may apply to alter the denotation of a lexical item or phrase. These transformations are the generative devices such as type coercion, selective binding, and co-composition, which formally map the expression to a new meaning. For example, when we combine the qualia structure of an NP with that of a governing verb, a richer notion of compositionality emerges (i.e. *co-composition*), one which captures the creative use of words (Pustejovsky and Boguraev 1993).

The Current Technology: Lexical Indexing

The language of types described above is necessary for encoding lexical knowledge but is not in itself sufficient for the characterization of how words are used in specific domains with specialized senses. Domain tuning from the core lexicon is the next step for arriving at lexical structures that are sensitive to specific corpora. The second component provided by the Core Lexical Engine is a set of lexical acquisition tools for doing just this; namely, enriching the lexical structures of the core lexicon for a single domain. We illustrate this methodology briefly by considering the acquisition of technical terms such as complex nominals (e.g., *startup disk*, *system folder icon*) in the specialized technical domains of software applications, e.g., *Macintosh (Operating System) Reference* and *PowerTalk (Collaborative Environment) User Guide*.

Our work in this area is based on two major premises: the combination of statistical and knowledge-based techniques for lexical acquisition, and the realization that the lexicon must be a dynamically evolving database, rather than a static collection of definitions.

While recent work in lexical acquisition has demonstrated that statistically-based methods for mining large text corpora for lexical and world knowledge are both feasible and useful, the most successful techniques are proving to be those which integrate statistical methods with language-specific knowledge encoding techniques (cf. Boguraev and Pustejovsky, 1996 and Weischedel et al., 1994). This is the approach taken in the Core Lexical Engine project and the one we assume here.

Recent application domains and projects such as MUC and TIPSTER show that lexicons and dictionaries must evolve from static sources of word definitions to dynamic knowledge bases which function as resources for natural language technologies (such as NL-based information access and retrieval, machine translation, and natural language understanding). In particular, it will be increasingly important to provide adequate lexical support for processing of technical documentation. Given the frequency with which complex nominals are coined, for example, it is impractical to rely on a manual listing of them, and so it becomes critical that online lexical resources be able to acquire and interpret them dynamically as text is processed.

The approach to lexical indexing outlined here instantiates the model of corpus-driven lexical semantics acquisition presented in Pustejovsky *et al.* (1993) (cf. also Grishman and Sterling, 1992). This model is based on the notion of targeted acquisition, driven by an appropriately selected corpus, and is applicable to a range of lexical types, all of which require richly instantiated lexical entries. In our work, we have found that the semantics derived by this process for the interpretation of a class of complex nominals facilitates other aspects of linguistic processing, including the resolution of anaphors and the interpretation of possessives, postnominal complements, and conjoined nominals. This work suggests that a dynamic approach to lexicon and dictionary design, given the appropriate processing and analysis of suitable corpora, is a realizable goal and of great utility in subsequent linguistic analysis.

Complex nominals and other specialized phrases are identified in the text through the application of a grammar suitably tuned for technical terminology. The grammar is not dissimilar to the one presented in Justeson and Katz (1995). However, it should be stressed that a complex nominal subsumes the strong notion of a *technical term* as defined by Justeson and Katz — the primary difference is that for the purposes of acquisition and interpretation of complex nominals it is essential to identify all nominal objects which map to entities in the domain. In order to pick out the relations and properties which apply to each complex nominal, their appearance in a variety of basic syntactic patterns in the text is exploited. In order to identify these patterns within the text, the system needs a syntactic base which is more expressive than that provided by part-of-speech tagging alone. As it turns out, "shallow" syntactic parsing (as carried out for instance by the constraint grammar of Voutilainen *et al.*, 1992) offers enough in its system of syntactic function labels to drive the analysis of complex nominals in context.

The initial stage of the process is to run a window over the tagged text and identify matches between the text and a series of patterns which are tuned to identify higher syntactic constituents in which the nominals appear. Each of these constituents contributes a relation or property with which a complex nominal composes. One of the local ontologies established in our initial investigation of this methodology, using the Macintosh Reference corpus, is the disk ontology. We will present here a small portion of this ontology, which includes the complex nominals *floppy disk* and *hard disk*. The relation sets established from the corpus for these forms are as follows.

(A) *floppy disk*: `name [], erase [], access [], initialize [], test [], save file on [], repair [], eject [], insert [] into disk drive, lock [], unlock [], protect []`

(B) *hard disk*: `name [], save file on [], repair [], erase [], access [], initialize [], test [], attach [] to printer, reinstall system software on [], copy disk to []`

In this operational environment the system successfully established that floppy disks are things which can be inserted into a disk drive, ejected, locked, unlocked, and protected, while hard disks cannot, and that hard disks are things which can be attached to printers, have system software reinstalled on them, and have program disks copied to them, while floppy disks are not. Given this information, the interpretation of *floppy disk* and *hard disk* proceeds as follows. In order to identify the set of relations which are appropriate for *disk* we take the intersection of the relations sets for the complex nominals which are headed by *disk*. This gives us the set: `access []`, `erase []`, `name []`, `initialize []`, `test []`, `save file on []`, `repair []`. This set defines what can happen to all disks and to a large extent defines extensionally the meaning of the term *disk* in this domain.

This kind of methodology underlies our lexical indexing system, TexTract-1. Applying it to a technical manual in a given domain, we are able to construct a flat index of the objects in this domain. In the case of the *Macintosh Reference* technical documentation, enriching the term set with the relations, properties, and related objects of the terms (as illustrated above), yielded a domain catalog which, in effect, is an extensional description of the functionality of the domain. One task to which such domain catalogs have been applied is that of the automatic population of *Apple Guide* databases, which drive the delivery of context-sensitive on-line help in the Macintosh OS environment (cf. Boguraev, 1996).

Given the high precision of the linguistic analysis, the contextually-sensitive patterns targeting the syntactic environments of domain objects, and the closed nature of technical domains, this kind of system is capable of achieving very high degrees of recall and precision. For example, a comparative evaluation of the domain catalog derived from the *Macintosh Reference* manual against the manually crafted *Apple Guide* database for the same domain shows 94% recall and 89.5% precision for the identification of domain-relevant objects (terms); for relations, the figures are, respectively, 91.1% and 88.5% (cf. Boguraev, Pustejovsky, and Johnston, *forthcoming*).

What this discussion demonstrates is that with a combination of tagging and pattern matching, we can automatically acquire rich, domain-specific sets of structured relations for words in a given domain. The resulting lexical structures are the starting point for creating a lexical web for a document, and the material from which hotlinks between words in a document are constructed.

Toward Richer Document Processing

The current indexing system has also been expanded in order to create hyperlinks from the index back to the documents. Each term structure is linked to all occurrences of the word or phrase in the text, and sub-structured to include relations and properties associated with terms that are present in the document. After running the system TexTract-1 over Darwin's *Origin of Species*, the result was a fairly rich term index including compounds and proper names, as well as standard index terms (see Figure 1). Individual index entries are further instantiated, as exemplified by the information brought together under the proper name *St. Hilaire* (Figure 3).

The index derived by TexTract-1 is already sub-

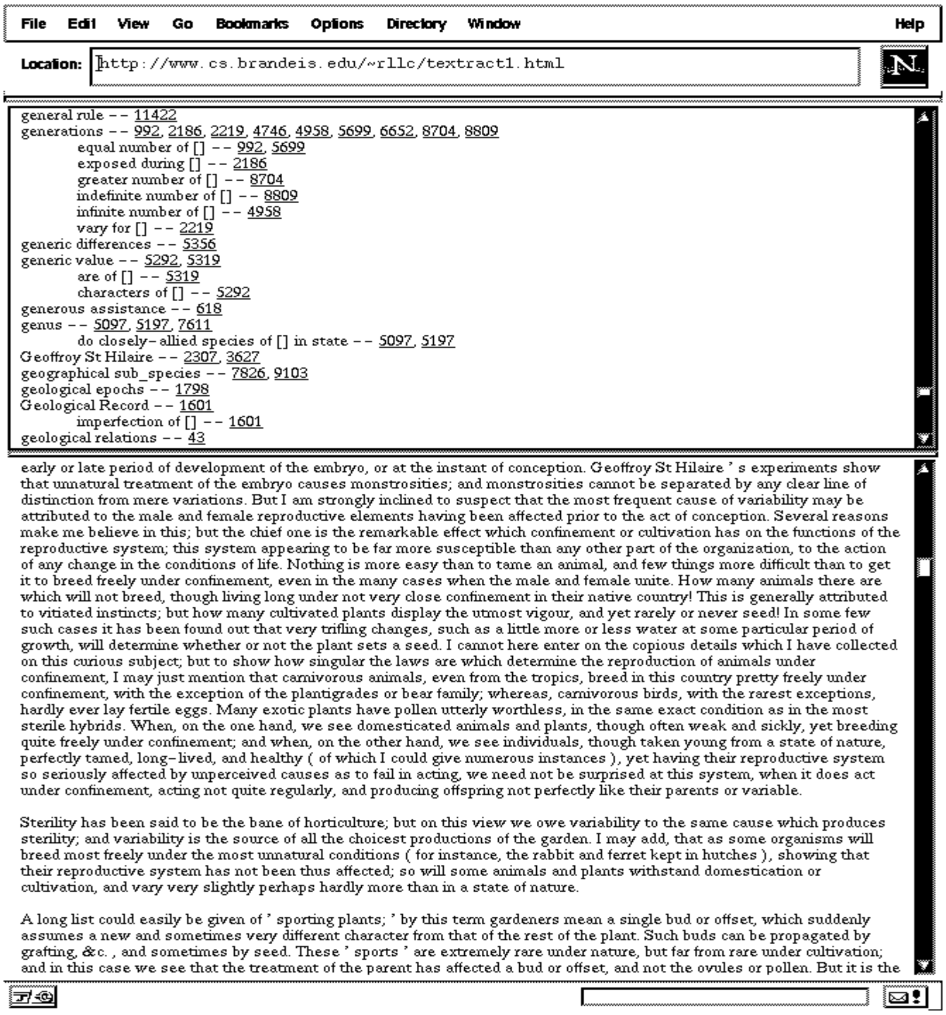

Figure 1: TEXTRACT-1

stantially richer than those created by under-informed noun phrase indexing engines, like the one marketed by Iconovex Corporation (cf. `www.iconovex.com`). However, there are several obvious problems with this structure, some due to omissions, others pertaining to spurious or irrelevant entries. Using semantic tagging and more refined pattern matching, however, enables us to create a more sophisticated version of the domain catalog, one that supports a richer index without creating the spurious entries produced by TEXTRACT-1.

CoreLex and Semantic Tagging

One of the major goals in our research has been the development of CORELEX (see Figure 2 for a sample), a semantic lexicon structured in such a way that it reflects the lexical semantics of a language in systematic and predictable ways, as expressed in the syntax. CORELEX embodies most of the principles of Generative Lexicon Theory, by representing how senses are related to one another as well as operating with underspecified semantic types. These assumptions are fundamentally different from existing sources such as Roget and WORDNET (cf. Miller 1990), both useful and extensive semantic lexicons, but which do not account for any regularities between senses nor do they relate semantic information to syntactic form.

Roget and WORDNET, however, are both used in the construction of CORELEX, since they are vast resources of lexical semantic knowledge which can be mined, restructured and extended (for discussion of this process, cf. Buitelaar 1997a and 1997b).

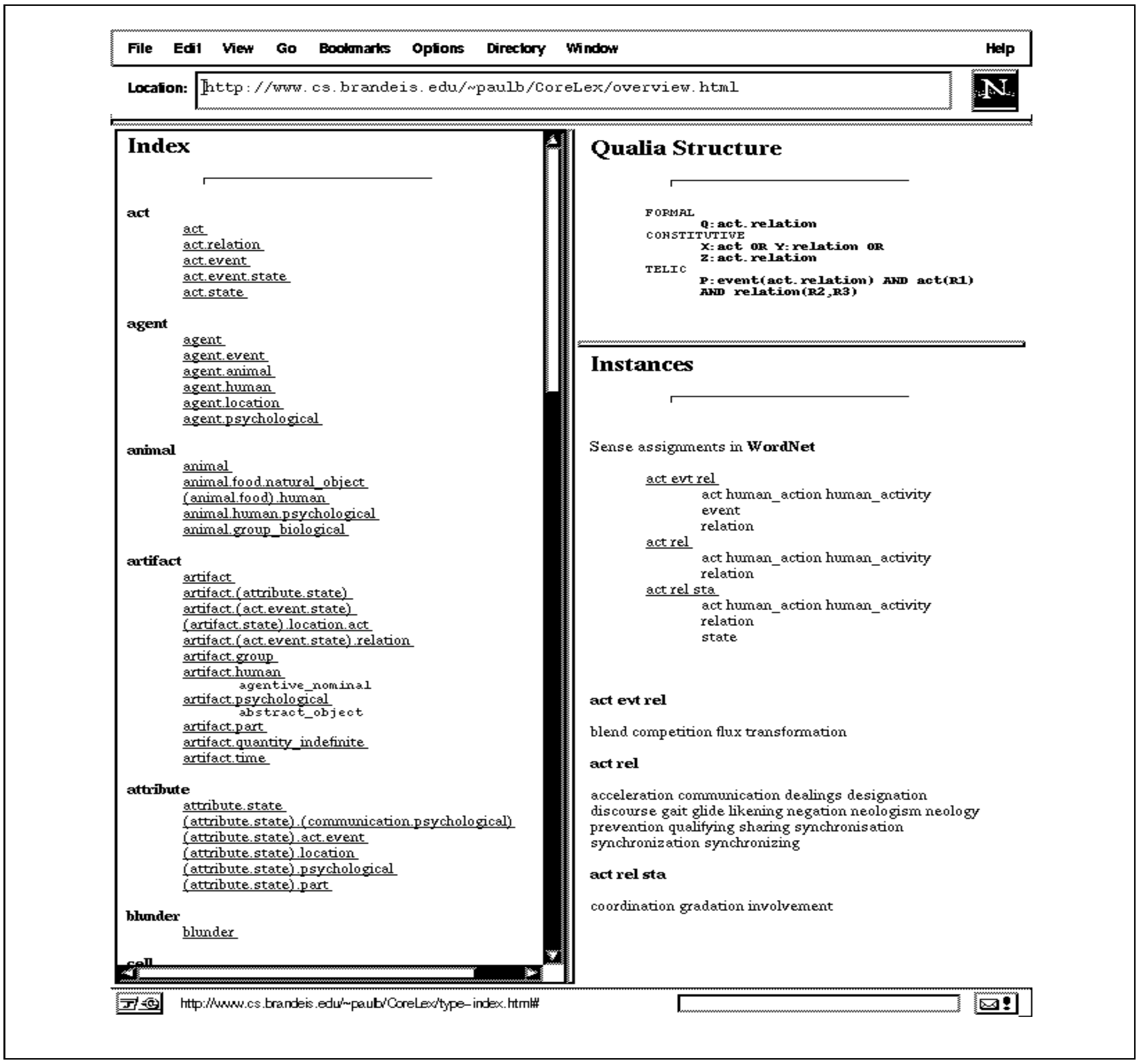

Figure 2: CoreLex

The top lattice structure is defined as those typed feature structures corresponding to the ten top underspecified categories, together with the categories generated by the application of *projective transformations* on this category, including $\neg$, $<$, $\circ$, $>$, *act*, and $=$. These transformations, together with qualia-based relations between types, define a circumscribed semantic field of concepts for each type. The top lattice types are: *Logical Types*: MODAL, QUANTIFICATION, RELATIONAL, and TEMPORAL; *Material Types*: EXISTENCE, SPACE, FORM, MOTION, SUBSTANCE, MENTAL.

The linguistic features exploited in the TIPSTER exercise were very limited. More sophisticated featural information is required to drive the identification of dependencies needed for the shallow understanding described here.

In order to assign this information to words, a core requirement is that of a more sophisticated tagging technology, a semantic tagger, which assigns grammatical function and semantic typing information in addition to part of speech. CoreLex provides the resources for augmenting current tagging techniques with semantic typing information: The Core Lexical Engine provides a rich language of types for lexical description as well as specific algorithms for defining how word senses relate to one another. This is the first step toward enriching the document for indexing and navigation. The type lattice is the formal backbone used to automatically assign semantic tags to relevant words and phrases within a document; these semantic tags are used as the first step in the construction of a typed semantic index of the document, or for typed hyperlink enhancement within a web-based browsing situation.

Linguistic Generalizations and Pattern Matching

The Core Lexical Engine also provides specific algorithms for defining how word senses relate to one another. The second step towards enriching the document for indexing and navigation, is to implement these algorithms as finite state pattern matchers. Linguistically distinct phenomena are associated with dedicated pattern matchers, each designed to identify salient semantic objects and relations in a text. In our previous research, we have developed a variety of knowledge extraction pattern matchers (cf. Johnston *et al.*, 1995, Boguraev *et al.*, forthcoming). In the current context, we wish to further extend and refine these pattern matchers to apply to a much broader class of linguistic phenomena. We identify the following pattern matchers needed for richer text indexing:

a. TERM IDENTIFICATION AND TYPING: These matchers are responsible for the determination of the technical terms used within a document and identification of the basic semantic types to which they are assigned.

b. RELATION AND PROPERTY PARSING: A further set of patterns

   identify the properties of the objects identified and the relations in which they appear.

c. REPORTED SPEECH: For certain genres of text, such as legal documentation and news reports a great deal of information regarding the content can be extracted by reference to expressions of reported speech. These are instances in which propositions are associated with particular individuals. Our work on this aspect draws on Bergler (1992), which provides detailed analysis of these constructions within the Generative Lexicon framework.

d. COMPLEX NOMINAL PARSING: Complex nominal parsing involves examination of the internal structure of complex nominals including: noun-noun compounds ('disk drive utility'), possessive's ('hard disk drive's SCSI ID number'), and nominal post-modificational constructions ('release hatch of disk drive') in order to extract knowledge about the domain (cf. Johnston

   et al., 1995).

Other issues currently being dealt with in the context of improving acquisition and shallow semantic analysis include: discourse anaphoric processing and sortal anaphora (cf. Kennedy and Boguraev, 1996a); and identification of salient topics and topic structure within text (cf. Kennedy and Boguraev, 1996b).

From Indexing to Lexical Webs

In essence, a Lexical Web is the next generation of the domain cataloging strategy, one which is made possible by the general functionality of the Core Lexical Engine technology. A Lexical Web is an interlinked structure of automatically generated index entries, embodying the semantic richness of the typing system from CORELEX. All fields are cross-indexed to other entries, where each field also points to the appropriate positions in the document.

The above-mentioned semantic tagging technologies as well as the more linguistically sophisticated pattern matching operations are essential resources in the acquisition of Lexical Webs, because it enables semantic tagging of the input text and it supplies conceptual prototypes which are inherited into entries in the Lexical Web. The Core Lexical Engine also provides descriptions of the canonical syntactic forms of particular semantic types, information which is employed in the identification and classification of the specific semantic types in a particular document collection.

Like the domain catalog in the Lexical Indexing system, a Lexical Web is essentially a network of the salient terms and relations within a given subdomain. The difference is that there is a far greater number of connections of a wider variety of types within the network, resulting in an overall higher degree of expressiveness and utility. Also, the connection of the term to its locations in the document collection is explicitly encoded. For each term, the following information will be captured:

1. The forms in which the term appears and their locations;
2. The basic semantic type of the term;
3. Relations appearing with the term;
4. Properties with which the term appears;
5. Typed relations to other terms (e.g., `has_part`, `is_a`, `part_of`);
6. Typical use and origin of the term (i.e., TELIC and AGENTIVE);
7. Definitional statements for the term.

The Core Lexical Engine plays a critical role in capturing this range of information. By examination of its phrasal and semantic context, each term is assigned to a lexical type from the core lexicon and inherits the basic structure associated with that lexical type. This basic structure is then filled out with more specific information derived through more detailed mining of the text to identify related forms of the same term, and the relations, properties, and related objects with which the term appears. The entry for *Mr. St. Hilaire* from Darwin's *Origin of Species* mentioned before, is shown in Figure 3.

We view the task of determining the Lexical Web for a text as a form of lexical acquisition. On the basis of corpus training, the relevant basic lexical types within the Core Lexical Engine are specialized into a set of data structures which together constitute a lexical web. One of the central concerns of our work is to develop a set of Lexical Web Acquisition Tools which are responsible for this training process.

Lexical Webs can be utilized to enable the automatic and semi-automatic annotation of web-based texts. There are two major modes of document enrichment one can envision. The first of these is close in nature to the help database application of the lexical indexing system and the task of automatically generating an index for a book. The information for a document in the Lexical Web can be processed to construct an index of the salient topics and subtopics within the document. This index can then be prefixed to the HTML form of the document with links from the index items to the relevant portions of the text. This can greatly facilitate access to the document.

The second of these facilitates non-linear navigation of the content of the text. Salient terms and phrases within the text are highlighted and enriched with links to related items within the document. These related items may in the simplest case be further occurrences of the same terms, but these items could also stand in `part_of`, `is_a`, and `has_part` relations.

We envision both automatic and semi-automatic construction of web page texts. In an automated mode, the user will select the range of types of links that are desired and the system will automatically generate an index and links within the document. In a semi-automated mode the system will identify and suggest items which might potentially be linked and allow the user to choose whether or not they should be linked and how they should be linked. Similarly, the user can also be presented with an automatically derived index which can then be edited. Even in the semi-automatic mode, the use of this system will greatly facilitate the process of getting material onto the Web and will enforce a degree of comprehensiveness and consistency of style which is absent from hand-crafted HTML forms of documents.

content

used documents browsable the

Conclusion

This paper lays out a strategy for applying robust natural language processing and knowledge mining techniques to the automated semantic annotation of documents from plain text resources. Semantic tagging and corpus tuning of general lexical types from the Core Lexical Engine are employed to generate document content abstractions called lexical webs. Lexical Webs provide the information and structure necessary to automatically generate the index for and hyperlinks within a document and between documents. This research is important because it integrates techniques from corpus acquisition research and natural language technologies with the goal of document analysis and annotation, as well as information extraction for web-based documents.

The purpose of the language processing methodology that we have described

is to generate a data structure which provides a good indication of what a document is primarily *about*. Rather than just having a list of words or noun phrases, we have a network of meaningful connections among objects from the text. This provides a better indication of aboutness than a phrase list, but by no means should be taken as a representation of the complete meaning of a text. Although the representation of aboutness generated is only partially specified and lacks a specific semantics, it is extremely useful if we can relate it back to the document or document collection from which it was derived. HyperText, specifically HTML in the

context of the World Wide Web, provides a natural delivery vehicle for the results of this technology.

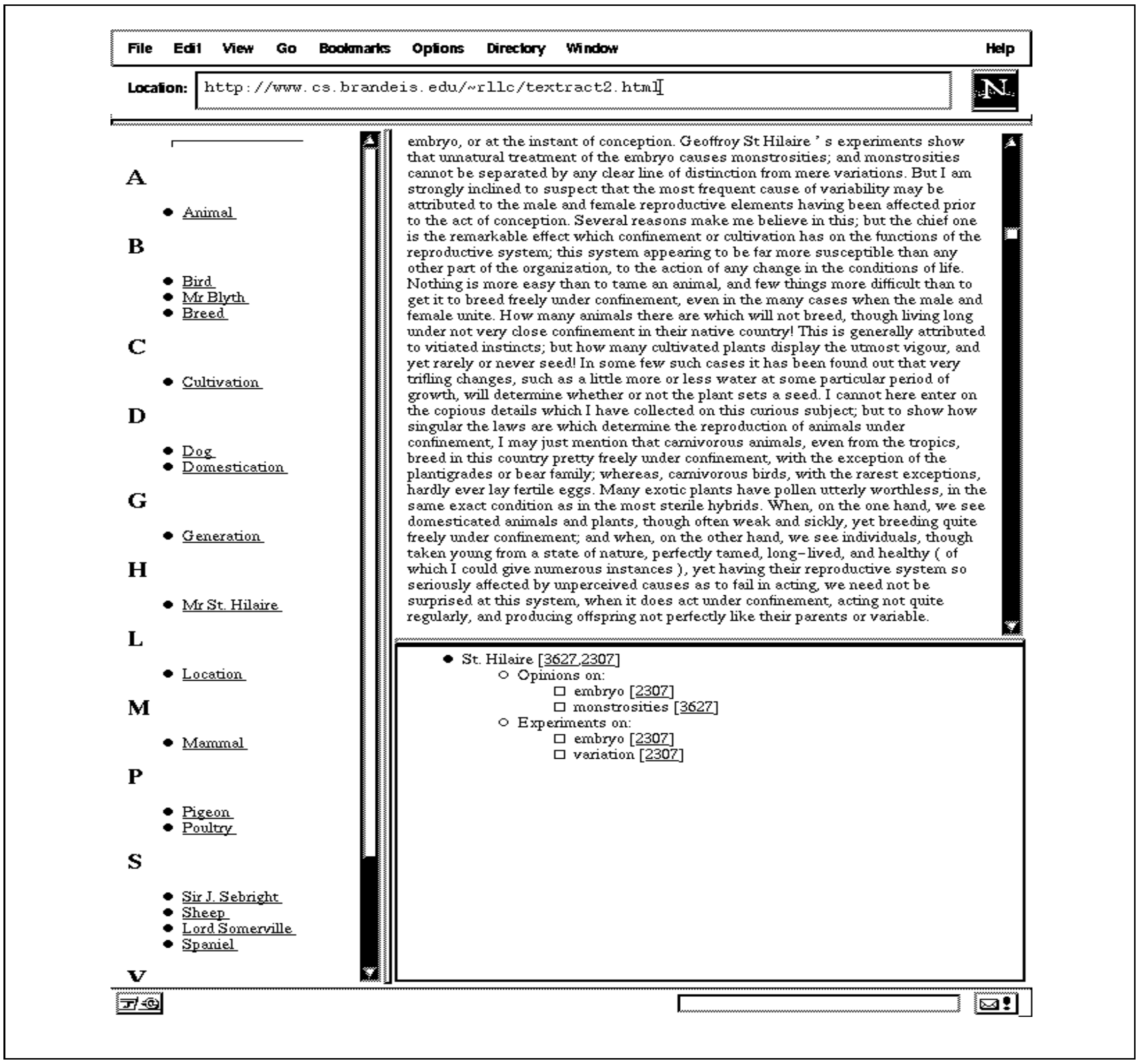

Figure 3: TEXTRACT-2

References

S. Bergler, *Evidential Analysis of Reported Speech*, Ph.D. thesis, Computer Science Department, Brandeis University(1992).

B. Boguraev, Building a Lexicon: The Contribution of Computational Lexicography, in: L. Bates and R. Weischedel, eds. *Challenges in Natural Language Processing* (Cambridge University Press, Cambridge and New York, 1992).

B. Boguraev, (1996). "WordWeb and Apple Guide: Comparative Indexing over Technical Manuals," Apple Research Laboratories, Technical Report TR-107, Cupertino, CA.

C. Kennedy and B. Boguraev, 1996a. Anaphora for everyone: Pronominal anaphora resolution without a parser. In The Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen, Denmark.

C. Kennedy and B. Boguraev, 1996b. Anaphora in a wider context: Tracking discourse referents. In W. Wahlster (ed.), The Proceedings of the 12th European Conference on Artificial Intelligence. London: John Wiley and Sons, Ltd.

B. Boguraev, J. Pustejovsky and M. Johnston, *forthcoming*, Content Abstraction and Indexing for Homogeneous Corpora.

B. Boguraev and J. Pustejovsky, *Corpus Processing for Lexical Acquisition* (Bradford Books/MIT Press, Cambridge, MA, 1996).

P. Buitelaar, 1997a, A Lexicon for Underspecified Semantic Tagging, to appear in: *Proceedings of ANLP 97, SIGLEX Workshop*, Washington DC.

P. Buitelaar, 1997b, CORELEX: A Semantic Lexicon with Systematic Polysemous Classes, submitted to: *ACL/EACL 1997.*

J. Cowie,T. Wakao,W. Jin L. Guthrie, J. Pustejovsky, S. Waterman, The *Diderot* Information Extraction System in *Proceedings of the First Pacific Conference on Computational Linguistics*, Vancouver, April 20, 1993.

J.L. Fagan, 1987. Experiments in Automatic Phrase Indexing for Document Retrieval: A Comparison of Syntactic and Non-syntactic Methods. PhD Thesis, Cornell University, September 1987.

J. Grimshaw, *Argument Structure* (MIT Press, Cambridge, 1990).

R. Grishman and J. Sterling, Acquisition of Selectional Patterns, in *Proceedings of the 14th Int'l Conf. on Computational Linguistics (COLING 92)*, Nantes, France, July, 1992.

Guthrie, L, J. Pustejovsky, Y. Wilks, and B. Slator. "The Role of Lexicons in Natural Language Processing," *Communications of the ACM*, 39:1, January, 1996.

M. Johnston, B. Boguraev, and J. Pustejovsky, The Acquisition and Interpretation of Complex Nominals", in *Working Notes of AAAI Spring Symposium on the representation and acquisition of lexical knowledge*, AAAI, 1995.

J. Justeson and S. Katz, 1995, Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text, *Natural Language Engineering*, 1.1.

W. Lehnert and B. Sundheim, 1991, A Performance Evaluation of Text-Analysis Technologies, *AI Magazine*, 81-94.

G. Miller, 1990, WordNet: An on-line Lexical Database, *International Journal of Lexicography*, 3, 235-312.

M. Moens and M. Steedman, Temporal Ontology and Temporal Reference, *Computational Linguistics*, 14(2):15-28.

R. Passonneau, A Computational Model of the Semantics of Tense and Aspect, *Computational Linguistics* **14** (1988).

J. Pustejovsky, The Generative Lexicon, *Computational Linguistics* **17** (1991a) 409–441.

J. Pustejovsky, The acquisition of lexical semantic knowledge from large corpora. In *Proceedings of the DARPA Spoken and Written Language Workshop.* Morgan Kaufmann, 1992.

J. Pustejovsky, *The Generative Lexicon: A Theory of Computational Lexical Semantics* (MIT Press, Cambridge, MA, 1995).

J. Pustejovsky, S. Bergler and P. Anick, Lexical Semantic Techniques for Corpus Analysis, *Computational Linguistics*, Special Issue on Corpus Linguistics, 19.2, 1993.

J. Pustejovsky and P. Boguraev, Lexical Knowledge Representation and Natural Language Processing, *Artificial Intelligence*, **63** (1993) 193-223.

A. Voutilainen, J. Heikkilä and A. Anttila, 1992, Constraint Grammar of English. A Performance-Oriented Introduction, Publications No. 21, Dept. of General Linguistics, University of Helsinki.

Weischedel R. et al., "Description of the PLUM system as used for MUC-5", *Proceedings of the 5th Message Understanding Conference*, Kaufmann, 1994.