

ABSTRACT

Title of dissertation: INDUCING SEMANTIC FRAMES
 FROM LEXICAL RESOURCES

Rebecca Joyce Green, Doctor of Philosophy, 2004

Dissertation directed by: Professor Bonnie J. Dorr
 Department of Computer Science

The multiple ways in which propositional content can be expressed is often referred to as the paraphrase problem. This phenomenon creates challenges for such applications as information retrieval, information extraction, text summarization, and machine translation: Natural language understanding needs to recognize what remains constant across paraphrases, while natural language generation needs the ability to express content in various ways.

Frame semantics is a theory of language understanding that addresses the paraphrase problem by providing slot-and-filler templates to represent frequently occurring, structured experiences. This dissertation introduces SemFrame, a system that induces semantic frames automatically from lexical resources (WordNet and the *Longman Dictionary of Contemporary English* [LDOCE]). Prior to SemFrame, semantic frames had been developed only by hand.

In SemFrame, frames are first identified by enumerating groups of verb senses that evoke a common frame. This is done by combining evidence about pairs of semantically related verbs, based on LDOCE's subject field codes, words used in LDOCE definitions and WordNet glosses, WordNet's array of semantic relationships, etc. Pairs are gathered into larger groupings, deemed to correspond to semantic frames. Nouns associated with

the verbs evoking a frame are then analyzed against WordNet's semantic network to identify nodes corresponding to frame slots.

SemFrame is evaluated in two ways: (1) Compared against the handcrafted FrameNet, SemFrame achieves its best recall-precision balance with 83.2% recall (based on SemFrame's coverage of FrameNet frames) and 73.8% precision (based on SemFrame verbs' semantic relatedness to other frame-evoking verbs). A WordNet-hierarchy-based lower bound achieves 52.8% recall and 46.6% precision. (2) A frame-semantic-enhanced version of Hearst's TextTiling algorithm, applied to detecting boundaries between consecutive documents, improves upon the non-enhanced TextTiling algorithm at statistically significant levels. (Previous enhancement of the text segmentation algorithm with thesaural relationships had degraded performance.)

INDUCING SEMANTIC FRAMES
FROM LEXICAL RESOURCES

by

Rebecca Joyce Green

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Committee:

Professor Bonnie J. Dorr, Chair
Professor Linda Coleman
Professor James A. Hendler
Professor Michael Israel
Professor Philip Resnik
Professor Amy S. Weinberg

©Copyright by
Rebecca Joyce Green
2004

Acknowledgments

Special thanks are due to my advisor, Bonnie Dorr, who has very patiently and kindly made helpful suggestions to encourage my research and to improve this dissertation, even when it was quite inconvenient to do so. My thanks also go to each of those who served on my committee for their time, encouragement, and enthusiasm; here special mention must be made of Philip Resnik, who not only served on the committee during his sabbatical, but also functioned as committee point person during the revision process. I also express my appreciation to Dekang Lin, who generously supplied his verb cluster data for comparative purposes.

Although somewhat removed in time and space, my appreciation for and dependence on the Berkeley linguistics community must also be acknowledged. Without Chuck Fillmore and George Lakoff, I would not have the understanding I have of how our cognition structures our language and how our language in turn reflects how we think. And were it not for Eastern transplants from the Berkeley linguistics community, I would not have been able to complete my dissertation committee.

I also acknowledge the ongoing support of friends, family, compatriots within the CS doctoral program, and the CLIS community, faculty and students alike.

Table of Contents

Acknowledgments	ii
List of Tables	v
List of Figures	vii
1 The Paraphrase Problem	1
1.1 The Nature of the Paraphrase Problem	1
1.2 Research Hypotheses	2
1.3 Research Approach	5
1.4 The Nature of the Solution to the Paraphrase Problem	7
1.5 Synopsis of Dissertation	12
2 Review of Previous Work	14
2.1 Paraphrase	14
2.2 Semantic Predicate Argument Structure	18
2.2.1 Thematic Roles	19
2.2.2 Semantic Verb Classes (Levin/MIT Lexicon Project)	23
2.2.3 Lexical Conceptual Structure	27
2.2.4 Frame Semantics	32
2.3 FrameNet	36
3 Lexical Resources	39
3.1 Lexical Resources as Data Sources	39
3.1.1 Definitions	40
3.1.2 Example Sentences	43
3.2 Longman Dictionary of Contemporary English (LDOCE)	44
3.3 WordNet	47
4 Identifying Semantic Frames by Extension: Task 1	51
4.1 Data Sources	54
4.1.1 LDOCE-based Data Sets	55
4.1.2 WordNet-based Data Sets	73
4.2 Mapping WordNet Verb Senses to LDOCE Verb Senses	80
4.3 Verb Pair Weights	83
4.4 Combining Verb Pairs to Form Verb Groups	85
4.5 Summary	87

5	Identifying Semantic Frames by Participant Structure: Task 2	89
5.1	Enumerating Noun Senses that Evoke Semantic Frames	92
5.1.1	Nouns Drawn from LDOCE Entries	93
5.1.2	Nouns Generated by Morphological Analysis	96
5.1.3	Nouns Mapped to WordNet Synsets	98
5.2	Computing Conceptual Density	99
5.3	Interpreting Conceptual Density	103
5.3.1	Identifying Names for Semantic Frames	104
5.3.2	Identifying the Internal Participant Structure of Semantic Frames	105
5.4	Results	106
5.5	Summary	109
6	Evaluation	111
6.1	Hearst's TextTiling Algorithm	112
6.2	Modification of TextTiling to Incorporate Semantic Frames	116
6.3	Evaluation Results	121
6.4	Summary	133
7	Discussion	134
7.1	Threats to the SemFrame Approach	134
7.1.1	Data Sparseness	135
7.1.2	Shallowness in Mapping Technique	136
7.1.3	Inadequacy of Merging Process	138
7.2	SemFrame <i>vis-à-vis</i> FrameNet	139
7.3	Summary	159
8	Contributions and Future Work	162
8.1	Other Resources	166
8.1.1	WordNet 2.0	170
8.1.2	FrameNet	177
8.1.3	Corpus Data	178
8.2	Other Contexts	181
8.3	Other Applications	186
8.3.1	Word Sense Disambiguation (WSD)	186
8.3.2	Information Retrieval (IR)	188
8.3.3	Information Extraction and Related Tasks	192
8.3.4	Machine Translation	193
8.4	Summary	196
	References	197

List of Tables

Table 1.	Relationship Types Underlying Paraphrase	16
Table 2.	Definitions for Verbs Evoking the COMMERCIAL TRANSACTION Frame . . .	41
Table 3.	Results of Frame Clustering Process	86
Table 4.	Glosses and Example Sentences in LDOCE for Verb Senses Related to Golf	87
Table 5.	Distribution of Files across Brown Corpus Subsets in SEMCOR	124
Table 6.	Token-Sequence Sizes and Count	125
Table 7.	Block Sizes and Count	125
Table 8.	Final Token-Sequence Sizes and Count	125
Table 9.	Final Block Sizes and Count	126
Table 10.	Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 1.0	128
Table 11.	Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 1.5	129
Table 12.	Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 2.0	130
Table 13.	Average Maximum Harmonic Mean (F-score) over 125 Trials, Cluster Threshold = 1.5 or 2.0	132
Table 14.	FrameNet-based Evaluation of SemFrame, Cluster Threshold = 1.5	153
Table 15.	FrameNet-based Evaluation of SemFrame, Cluster Threshold = 2.0	154
Table 16.	Comparative Results 1: FrameNet-based Evaluation of WordNet Verb Hierarchies	156
Table 17.	Comparative Results 2: FrameNet-based Evaluation of Verb Classes (Levin 1993b)	158

Table 18. Comparative Results 3: FrameNet-based Evaluation of Verb Clusters (Pantel and Lin 2002, Lin and Pantel 2001b)	160
Table 19. Comparative Final Clustering Results for LDOCE and WordNet	167
Table 20. Derivational Morphology Links among <i>Pace</i> Words	171

List of Figures

Figure 1. Representative Verb Definitions from LDOCE	45
Figure 2. Overview of Task 1	53
Figure 3. Algorithm for Generating LC.pairs	57
Figure 4. Group Average Link Clustering Method, Voorhees' Algorithm	59
Figure 5. Algorithm for Generating LI.pairs	65
Figure 6. Algorithm for Generating LM.pairs	67
Figure 7. Algorithm for Generating LS.pairs	70
Figure 8. Algorithm for Generating LT.pairs and LZ.pairs	72
Figure 9. Algorithm for Generating WC.pairs	74
Figure 10. Algorithm for Generating WN.pairs	77
Figure 11. Algorithm for Generating WR.pairs and WX.pairs	78
Figure 12. Algorithm for Task 2	92
Figure 13. Conceptual Density in Subtrees	100
Figure 14. Examples of Frame Structures Generated in Task 2	107
Figure 15. Algorithm for Correspondence of FrameNet and SemFrame Frames	147

Chapter 1

The Paraphrase Problem

This dissertation introduces SemFrame, a system that induces semantic frames automatically from lexical resources and demonstrates that such frames can improve the performance of a well-regarded text segmentation algorithm. These accomplishments are noteworthy for the following reasons: (1) Prior to SemFrame, semantic frames had been developed only by hand. (2) Previous enhancement of the text segmentation algorithm with thesaural relationships caused a deterioration, rather than an improvement, in performance. (3) Semantic frame enhancements of other knowledge-intensive tasks also have the potential of improving performance.

1.1 The Nature of the Paraphrase Problem

The relationship between words and meanings is many-to-many: As a word, phrase, or sentence may have multiple meanings, so may a concept or proposition be expressed in multiple ways. Although considerable attention has been given to the first of these phenomena, the inverse phenomenon is less well studied (Edmonds and Hirst 2002, p. 106). However, it is arguably the more prevalent. While some lexical forms have only one meaning, it is almost always the case that semantic content can be expressed in more than one way. Thus the uncertainty in how specific semantic content will be expressed is often greater than the uncertainty of what a given word, phrase, or sentence means. Moreover, the meaning of a word or phrase that is ambiguous in isolation is usually constrained within a specific linguistic context to have a single best interpretation or at

worst to have a well-defined set of appropriate interpretations. But even context often only partially constrains the ways in which a speaker or writer may convey a message.

The difficulties involved in being able to express semantic content in multiple ways has been called the paraphrase problem. The problem affects both natural language generation and natural language understanding. This is perhaps clearer in the case of natural language generation, where the choice of how to express semantic content is everywhere present, indeed, is the very heart of the task. But it is also crucial to natural language understanding, where it is important to be able to ascertain the underlying unity of multiple expressions of shared semantic content. These issues affect many knowledge-based applications, including, for example, information retrieval, information extraction, question answering, text summarization, and machine translation.

1.2 Research Hypotheses

Two overarching hypotheses govern this research. The first concerns the use of a specific type of language-based knowledge representation structure—the semantic frame—to help resolve the paraphrase problem; the second concerns how semantic frames can be generated:

- General hypothesis 1. The incorporation of semantic frames into knowledge-intensive tasks can help solve the paraphrase problem.
- General hypothesis 2. Semantic frames can be induced automatically from data in machine-readable lexical resources.

The specific hypothesis that drives the overall research effort intertwines the two general hypotheses: Semantic frames can be induced from such lexical resources as WordNet (<http://www.cogsci.princeton.edu/~wn/>) and the *Longman Dictionary of Contemporary English* (LDOCE; Procter 1978) sufficiently well to improve performance on a knowledge-intensive task.

Up to this point, semantic frames have been generated by hand (Fillmore 1977, 1982, 1985; Fillmore and Atkins 1992).^{1,2} Furthermore, it is only recently, with the FrameNet project (Johnson et al. 2002), that overall breadth has been achieved in the generation of semantic frames. But if we are ever to have a comprehensive inventory of semantic frames, whether for general use or especially for use in specific subject domains, we need to develop the capacity to generate semantic frames at least semi-automatically,³ since generating them by hand is labor-intensive and lacks systematicity.⁴

¹Generation of semantic frames here refers to the identification of a semantic situation *type*, with its participant structure. Automatic generation of semantic frame *instantiations* has been pursued to a limited degree, as in Gildea and Jurafsky (2002) and Erk, Kowalski, Padó, and Pinkal (2003). The upcoming Senseval-3 Automatic Labeling of Semantic Roles task (Litkowski 2004a) will further promote this enterprise.

²The work of Riloff and Jones (1999) in inducing a semantic lexicon and a dictionary of extraction patterns for an information extraction (IE) system may at first glance seem a counterexample, since IE templates and semantic frames are essentially the same thing. However, in their work both domains (which correspond to frames) and semantic categories (which correspond to frame slots) are predefined.

³Manual post-editing is likely to be required.

⁴The FrameNet (2004) FAQ includes the following question-and-answer pairs:

Q1: How many frames do you think there are in all? How will you know when you're finished? How do you measure your progress?

A1: We don't know how many frames there are, but the number is obviously very large, and we don't expect to "finish" with all of them. However, that does not make our task

SemFrame begins the development of such a capability. Specifically, SemFrame implements an automatic method for generating a broad array of semantic frames, where each frame has associated with it (1) a set of verb and noun senses that evoke the frame and (2) a set of semantically typed slots. The validity of the frames thus generated is demonstrated by using them to enhance and improve the performance of a text segmentation algorithm (TextTiling; Hearst 1997) and by comparing them against a gold standard (FrameNet).

The contributions of this research effort include:

- the confirmation of the two research hypotheses mentioned earlier in this section,
- the generation of two resources (sets of verb and noun senses evoking a large number of semantic frames, as well as sets of typed slots associated with these frames),
- the formulation of a similarity measure more appropriate for computing similarity between definitions than extant similarity measures,
- the modification of an existing conceptual density measure, to make it more apt for defining the participant structure of semantic frames,
- the development of a mapping technique that, when given a source word sense and a target word, identifies those senses of the target word that are most likely to evoke the same semantic frame(s) as the source word sense, and
- an algorithm for semantic frame disambiguation in running text.

hopeless, any more than the task of writing ordinary dictionaries is hopeless because they can't possibly include all the specialized vocabulary in a language.

Q2: Have you plotted out all the frames that you need for the whole language?

A2: Our approach is basically opportunistic and bottom-up. . . .

1.3 Research Approach

The basic SemFrame approach to developing a general inventory of semantic frames involves two tasks. The goal of task 1 is to identify conceptual structures automatically by finding sets of verb senses used to communicate about the same propositional content; these are verbs whose syntactic arguments will tend to have sets of lexical items in common. In task 1 we would want to find that specific senses of *buy*, *purchase*, *sell*, *cost*, *pay*, *spend*, *expend*, *charge*, and *price* belong to one such set (among other verbs). The goal of task 2 is to automate the identification of the (semantic) arguments of the conceptual structures that correspond to these sets of verb senses, by analyzing the semantic types of the sets of syntactic arguments they share. As a result of task 2 we would want to discover that the group of verb senses commonly referred to as COMMERCIAL TRANSACTION verbs have as their arguments a Buyer, a Seller, Merchandise, and Money.

It is because of the relationship between the set of syntactic arguments that a verb has and the set of semantic arguments of the corresponding semantic frame that SemFrame is designed in terms of verb senses and their arguments. Words representing other parts of speech may also take arguments and evoke frames, but no other part of speech is so closely connected to the semantic frame enterprise as verbs are.

Two types of data are readily available for identifying syntactic arguments for sets of verbs: corpus data and machine-readable dictionary data. While corpus data would provide a significant amount of data on syntactic arguments, they are subject to several disadvantages. The first disadvantage is actually a set of related disadvantages: The

amount of data available for each verb sense used could vary significantly, with some verb senses used many more times in a corpus than other senses. Indeed some verb senses (if we can pretend for present purposes that a stable set of verb senses somehow exists) probably would not occur at all in a specific corpus. Moreover, it is the rare corpus (especially of substantial size) that is word-sense tagged. On the contrary, machine-readable dictionaries are relatively comprehensive in the set of verb senses they recognize. Furthermore, they tend to provide a fairly consistent amount of data for all the verb senses they contain, with verb senses being readily isolated by the structure of the dictionary. Another disadvantage in using corpus data for this task is that syntactic arguments tend to be instantiated in natural language use in fairly specific ways; the task would be made easier if the semantic classes of instantiated values could be more readily discerned than would often be the case with corpus data. On the contrary, these arguments tend to be expressed at a more generic level in a verb's definition, at times naming the very semantic class whose identification is being sought. Using data from machine-readable lexical sources thus seems the more appropriate starting point for inducing semantic frames.

The research is conducted entirely with English language data. The motivation for doing so ties into a widely adopted tripartite model of paraphrase (Brun and Hagège 2003). The first component of this model involves detecting/recognizing paraphrases; the second component involves representing the semantic content shared across paraphrases in a normalized manner; the third involves generating paraphrases from a normalized representation. Both the first and third components have traditionally been

cast as monolingual phenomena: What are the multiple ways that a single language has for conveying specific semantic content? The second component, as a representation generalizing over language expressions, is, to a large extent, impervious to the idiosyncrasies of linguistic expression and, to that extent, language-neutral. The first and third components will, no doubt, increasingly come to be cast as multilingual phenomena: What are the multiple ways that Language has for conveying specific semantic content? But we would expect that the second component, which is a primary output of the work to induce semantic frames, will remain largely language-neutral.

1.4 The Nature of the Solution to the Paraphrase Problem

A fundamental premise underlying the development of SemFrame is that when such knowledge-intensive tasks as information retrieval, information extraction, question answering, text summarization, and machine translation are automated, it would be beneficial to make use of a knowledge representation structure of a specific kind, namely, one based on semantic predicate types. A semantic predicate type is a relational structure used to represent meaning and is defined by both the nature of the relationship it conveys and the types of its arguments; for example, EQUALITY is a particular comparative relationship between two Comparands, its semantic arguments.⁵

On the one hand, language-based tasks can convey reasonably fine-grained distinctions, inasmuch as they have at their disposal a rich array of syntactic structures

⁵The names of SEMANTIC PREDICATE TYPES will be shown throughout in small upper case letters; the names of their Argument types will be indicated with an initial upper case letter.

and lexical items for the expression of semantic content. Even though language as used rarely makes explicit all that speakers/writers mean to convey and that hearers/readers understand, still we commonly communicate relatively subtle distinctions through our use of language. On the other hand, underlying the fine-grained semantic distinctions reflected in our selection of one syntactic construction over another or our choice of one lexical item over another are semantic generalizations pertaining to the states, activities, processes, etc.—the semantic content—we are trying to communicate about.

While a given linguistic expression may be variously interpreted to convey different semantic content and thus may mean different things, ambiguity also inheres in the number of linguistic expressions we might use to talk about some particular semantic content. The proposal to make use of knowledge representation structures based on semantic predicate types addresses this type of ambiguity by tying the predicate types to the underlying situations (states, activities, processes) that are the subjects of linguistic expression. The semantic predicate types being proposed are thus meant to generalize over the various linguistic expressions available for communicating about some situation. For the purposes of the knowledge-intensive tasks mentioned, where expression of semantic content plays a key role, surface-level paraphrases should have (essentially) the same underlying representation. The semantic predicate types under consideration are thus perceived in terms of extralinguistic situations and not in terms of, say, syntactic behavior, which may or may not reflect the semantic generalizations.⁶

⁶Levin (1993a, 1993b), has repeatedly demonstrated a correlation between syntax and semantics. More specifically, she has shown that shared syntactic behavior tends to reflect

For example, consider the semantic predicate, TRAVEL, in which a Traveler moves from an Originating location to a Destination location along some Path, using some means of Conveyance. All of the following sentences convey, more or less, the same propositional content, while at the same time enjoying various pragmatic differences:

- (1) Chris traveled by plane from Boston to home.
- (2) Chris returned home from Boston by plane.
- (3) Chris took a plane home from Boston.
- (4) Chris flew home from Boston.

We note that as long as the means of Conveyance is an airborne plane, the verb *fly* is appropriate; if the means of Conveyance were, say, a car, the verb *drive* would be more appropriate. Further, we note that while the essential meaning remains the same for all four sentences, their syntactic structure varies: For example, in (1)-(2) the Conveyance *plane* shows up in a prepositional phrase; in (3) *plane* is the direct object of the verb; and in (4) the plane is implied by the verb *fly* and is thus absent from the surface structure of the sentence. Still, in all four cases, Chris is the Traveler, Boston is the Originating location, home is the Destination location, the Path between Boston and home is unspecified, but is presumed to involve air travel, and the Conveyance is a plane.⁷

semantic relatedness. She has not shown, however, that semantic relatedness (such as is evident between paraphrases) tends to be manifest in shared syntactic behavior. Examples provided herein indicate that the correlation is less strong when going from semantics to syntax than from syntax to semantics.

⁷One could also talk about Chris flying home from Boston by car, if the car were moving speedily, as if it were a plane. While the value associated with the Conveyance would change in this case (i.e., from plane to car), the overall relational structure would remain the same.

The basic solution being proposed is to address the paraphrase problem by using knowledge representation structures based on semantic predicate types. The structures should have, minimally, the following characteristics:

- Knowledge representation structures should correspond to semantic predicate types (which in turn correspond to situation types). (It should be noted, however, that a knowledge representation structure may embed other knowledge representation structures; see the remark below on extensibility.)
- Knowledge representation structures should focus on semantic/conceptual commonalities: When essentially the same propositional content is being conveyed, the same representation is used. (This responds to the invariance-under-paraphrase criterion suggested in Norman and Rumelhart 1975, p. 45.)
- Knowledge representation structures should represent gestalt relational patterns, with their internal organization corresponding to a predicate's semantic arguments. This semantic argument structure enumerates the types of participants that may be related by the predicate.
- When the argument positions of a particular knowledge representation structure are instantiated, the structure represents a specific state, activity, or process.

To these characteristics—which may be considered necessary features of semantic representations for knowledge-intensive tasks—may be added other desirable characteristics, as outlined in Norman and Rumelhart (1975, pp. 45-46) :

- Taken as a set, the knowledge structures should be able to represent the full range of human thought (or, minimally, the range that can be expressed linguistically).

- The representations should accurately reflect possible degrees of semantic relatedness between two items, that is, inclusion, partial overlap, complete overlap, or disjunction. It should be readily apparent from the form of the representations whether one state/activity/process is wholly contained within another, whether there is (only) some degree of shared meaning, or whether there is no semantic relatedness between them. (The case of complete overlap is already taken care of in the invariance-under-paraphrase criterion mentioned earlier.)
- In a similar vein, there should be continuity among the representations, such that small differences in propositional content are reflected by small differences in representation; at the same time similarities in propositional content are reflected by similarities in representation.
- The knowledge structures should be extensible, so that the representation of new semantic content can be linked into previously constructed representations.
- Lastly, as a reality check, the way in which knowledge is represented should be consistent with what is known of human information processing.

This dissertation will make the case for adopting semantic frames to fill the role just described. Indeed, semantic frames are quite accurately characterized as knowledge representation structures based on semantic predicate types. It will be argued that the incorporation of such structures in knowledge-intensive tasks will help resolve the paraphrase problem.

What it means to incorporate semantic frames into a knowledge-intensive task depends on the depth of knowledge that the semantic frames comprise. Various levels of

semantic frame knowledge include: (1) the set of semantic roles associated with a semantic frame; (2) the set of word senses that evoke a semantic frame; (3) the set of correspondences between word senses that evoke a semantic frame, their syntactic behavior, and their mapping to semantic roles in the frame; and (4) real-world knowledge of how the roles within the frame interact. Even incorporation of just the first and second levels will be shown to be of benefit.

1.5 Synopsis of Dissertation

Chapter 2 presents a review of work undertaken by others in three areas related to the work reported here. Section 2.1 offers a survey of work involving paraphrase; Section 2.2 is a comparative analysis of how four different approaches to semantic predicate argument structure—thematic roles, semantic verb classes (Levin), lexical conceptual structure (Jackendoff), and frame semantics (Fillmore)—could contribute to a solution to the paraphrase problem; Section 2.3 takes up the FrameNet project, which is developing an inventory of semantic frames by hand.

Since the hypotheses underlying the current research concern whether semantic frames can be induced from lexical resources, Chapter 3 examines in depth the two lexical resources—LDOCE and WordNet—at the center of that effort. Both of these resources are available in machine readable form and are highly regarded.

Chapters 4, 5, 6, and 7 present the work undertaken here, that is, the development and evaluation of SemFrame. The task of inducing semantic frames from LDOCE and WordNet is broken into two tasks: first, generating an extensional characterization of semantic frames by identifying sets of verb senses that evoke a common frame, and

second, generating an intensional characterization of semantic frames by identifying their internal participant structure. The first of these tasks is discussed in Chapter 4, while the second task is considered in Chapter 5. Chapter 6 presents an extrinsic evaluation of SemFrame, based on the incorporation of semantic frames into Hearst's TextTiling algorithm, which achieves improved performance on detecting boundaries between concatenated texts. Chapter 7 discusses some of the threats to SemFrame that arise from various design decisions. This chapter also presents an intrinsic evaluation of SemFrame, based on its coverage of FrameNet frames and the precision of the sets of verb senses associated with its frames; the results of this evaluation are shown to compare favorably with similar evaluations undertaken for other efforts to develop semantic verb classes.

The work presented here is of necessity preliminary and is subject to considerable expansion, modification, and improvement. Chapter 8 summarizes the contributions of the research already undertaken and discusses future directions in which SemFrame should progress. One direction that warrants exploration is the use of additional resources to detect frame semantic relationships between lexical units. Consideration is also given to the task of developing semantic frames for a specific subject domain. Another direction in which work should proceed is the incorporation of semantic frame information in other types of knowledge-based applications.

Chapter 2

Review of Previous Work

2.1 Paraphrase

Before the recent surge in the availability of machine-readable text, paraphrase received only sporadic systematic attention. Nolan (1970, p. 14) characterizes paraphrase in terms of sentence-level meaning equivalence or sentence synonymy. The criterion she offers for determining if two expressions are paraphrases of each other relies on grammatical transformations and specific linguistic formulae. Her conception of paraphrase is thus too restrictive to capture the significant degree to which the example sentences in Section 1.4 convey essentially the same information. A broader characterization is given by C. I. Lewis (1946, p. 86), who suggests that two linguistic expressions are synonymous if they have the same (non-zero, non-universal) intension, where “the intension of a proposition includes whatever the proposition entails; it comprises whatever must be true of any possible world in order that the proposition should apply to it or be true of it” (Lewis 1952, p. 57). Again this characterization, based on logical entailment, is more restrictive than the looser notion of paraphrase adopted here, in which entailments of paraphrased sentences may not correspond exactly or fully.

Within the past several years the paraphrase problem has begun to receive concerted attention from the computational linguistics community, as demonstrated by the holding of two international workshops on paraphrasing. The first of these workshops was held as a part of the 2001 Natural Language Processing Pacific Rim Symposium (<http://nlp.nagaokaut.ac.jp/pub/NLPRS2001WS.html>); a second workshop took place in

conjunction with the annual meeting of the Association for Computational Linguistics in 2003 (<http://nlp.nagaokaut.ac.jp/IWP2003/>). Within this context a rather wider array of phenomena are usually accepted as paraphrase. Hirst, for instance, has defined paraphrase as “[talking] about the same situation in a different way,” but notes that “the same situation” does not necessarily mean that the same details are conveyed, the same truth conditions hold, or the same message is communicated across paraphrases.

A significant challenge to solving the paraphrase problem is the variety of ways in which paraphrases may be related to each other. Table 1, which is based on examples from Hirst (2003), Rinaldi et al.(2003), and Kozlowski, McCoy, and Vijay-Shanker (2003), summarizes common types of relationships that underlie paraphrase. Because the phenomena have been classified in different ways by different researchers—different means of classifying being itself a source of paraphrase—the collection of relationship types given are not necessarily exclusive; nor are they likely to be exhaustive.

A number of applications have been targeted by researchers working on paraphrase. For example, Zukerman, George, and Wen (2003) apply paraphrase analysis to document retrieval in the context of question answering, an application also addressed by Rinaldi et al. (2003) and Lin and Pantel (2001a). Brun and Hagège (2003) use paraphrase detection in the context of information extraction; indeed, it can probably be said that all work in information extraction involves paraphrase detection. Radev and McKeown (1998) and McKeown et al. (1999) use paraphrase detection and generation in connection with multidocument summarization, while Inui et al. (2003) employ paraphrase generation within a rewriting task that offers reading assistance by simplifying text.

Relationship type	Example
Lexical synonymy	The toddler sobbed, and he attempted to console her. <i>vs.</i> The baby wailed, and he tried to comfort her.
Syntactic variation	The gangster killed at least 3 innocent bystanders. <i>vs.</i> At least 3 innocent bystanders were killed by the gangster.
Morphological derivation	I was surprised that he destroyed the old house. <i>vs.</i> I was surprised by his destruction of the old house.
Clause subordination <i>vs.</i> anaphorically linked sentences	This is Joe's new car, which he bought in New York. <i>vs.</i> This is Joe's new car. He bought it in New York.
Noun-noun phrases	She loves velvet dresses. <i>vs.</i> She loves dresses made of velvet.
Comparatives <i>vs.</i> superlatives	He's smarter than everybody else. <i>vs.</i> He's the smartest one.
Different sentence types	Who composed the Brandenburg Concertos? <i>vs.</i> Tell me who composed the Brandenburg Concertos.
Different argument realizations	Bob enjoys playing with his kids. <i>vs.</i> Playing with his kids pleases Bob.
Inverse relationship	Only 20% of the participants arrived on time. <i>vs.</i> Most of the participants arrived late.
Head switching	Mike Mussina excels at pitching. <i>vs.</i> Mike Mussina pitches well. <i>vs.</i> Mike Mussina is a good pitcher.
Overlapping meanings	Lindbergh flew across the Atlantic Ocean. <i>vs.</i> Lindbergh crossed the Atlantic Ocean by plane.
Inference	The tight end caught the ball in the end zone. <i>vs.</i> The tight end scored a touchdown.
Viewpoint variation	The U.S.-led invasion/liberation/occupation of Iraq . . . You're getting in the way. <i>vs.</i> I'm only trying to help.

Table 1. Relationship Types Underlying Paraphrase

Paraphrase detection relies on paraphrase acquisition, which takes place on two different levels. The first level involves paraphrase recognition: Which set of linguistic expressions express similar meaning? The second level involves learning to recognize

the linguistic patterns that allow/produce paraphrase. A variety of approaches to both levels of paraphrase acquisition have been explored.

Perhaps the most frequent approach to paraphrase acquisition involves the manual extraction of phrases and sentences that are paraphrases and the manual generation of patterns that generalize over paraphrase sets (e.g., Tomuro 2003; Brun and Hagège 2003). Automatic approaches to paraphrase are sometimes narrowly constrained to consider a single type of paraphrase structure, for example, lexical paraphrase, typically using WordNet to identify potential synonyms (Zukerman, George, and Wen 2003; Rinaldi et al. 2003).⁸ The most promising automatic approaches to paraphrase acquisition rely on the existence of multiple documents with shared semantic content, for example, multiple translations of a single source document (e.g., Ibrahim, Katz, and Lin 2003) or newswire stories from different agencies that report on the same events (e.g., Barzilay and Lee 2003). Alignment techniques identify sentences that are likely to represent the same semantic content; this may include analyzing parser output (especially in the form of dependency trees) to investigate the array of patterns that produce paraphrase. Specifically, Lin and Pantel (2001a) and Ibrahim, Katz, and Lin (2003) examine dependency structures between aligned sentences that link the same two nouns. Features along the path between the two noun anchors are scored (e.g., using mutual information) to produce a measure of similarity. If the measure exceeds a threshold, the

⁸The emphasis on synonymy and similarity in paraphrase acquisition may be overdone. Budanitsky and Hirst (2001, p. 29) note that it is semantic relatedness and not just semantic similarity that is required for most NLP applications.

two sentences are considered paraphrases, and the dependency trees can be harvested for patterns that allow paraphrased expression.

2.2 Semantic Predicate Argument Structure

Rumelhart and Norman (1975, p. 41) define a predicate as “a general function that specifies the relations that might exist among some set of concepts.” They define an argument as a “type of concept . . . used within the predicate” and suggest that arguments be given names that set forth the selectional restrictions that apply to them within the context of the predicate. In a similar vein, Saint-Dizier (1999, p. 3) states that “the argument structure specifies the number of arguments the predicate has [its ‘arity’], the arguments being the participants which are minimally required for the activity or state described by the predicate to be understandable.” Predicates are thus primarily relational in nature; that is, a predicate is a kind of relationship. As relationships, predicates are structures, used to interrelate some number of participant types or arguments.

The literature on the predicate-argument structure of verbs is voluminous. The portion of that literature that pertains to the current research is limited to that which addresses the semantic generalization of predicate-argument structure. The paraphrase criterion which underlies this research can only be addressed where some degree of generalization permits multiple linguistic expressions to be represented uniformly.

On the one hand, there have been many attempts to generalize over the argument side of the relationship by introducing relatively small sets of semantic cases/roles—now more commonly referred to as thematic relations or theta roles; such approaches limit the number

of possible argument types to a tightly constrained set. This accords with the observation of Riemsdijk and Williams (1986) that “this [thematic relations] terminology implies a system of *argument types*” (p. 241; emphasis transformed). On the other hand, there have also been attempts to generalize over the predicate side of the relationship by identifying groups of verbs with essentially the same argument structure; these attempts differ among themselves in degree of generality and in whether the shared argument structure is evaluated predominantly from a syntactic point of view or a semantic point of view. Among such developments are the semantic verb classes of the MIT Lexicon Project, Jackendoff’s lexical conceptual structure (LCS) theory, and Fillmore’s theory of frame semantics.

The general flow of the ensuing review is toward the claim that, of the developments discussed here, only frame semantics provides an adequate basis for the research proposed. The discussion of other schools of thought or representations is intended to show commonalities they have with the frame-semantic approach and also to indicate wherein they fail to provide fully for the specific task of representing semantic predicate argument structure, given the criteria set out in Section 1.4.

2.2.1 Thematic Roles

The predominant view of predicate-argument structure in modern linguistics is one that generalizes over arguments by assigning them to a small set of thematic roles, for example, Agent, Patient, Theme, Instrument.

On the one hand, such roles constitute semantic argument types and thus might represent, or at least correspond to, participant types for classes of semantic predicates.

This is the view expressed by Saint-Dizier (1999) when he states that thematic roles are semantic labels that “name the semantic relation which holds between an argument and its predicative term” (p. 1). This view is echoed by Jackendoff (1987), when he affirms that “thematic relations are part of a level of semantic/conceptual structure, not part of syntax” (p. 372), where “conceptual structure is essentially the form in which thought is couched” (p. 374). For Jackendoff, “thematic relations are to be reduced to structural configurations in conceptual structure” (p. 378). In other words, thematic roles correspond to argument types for particular kinds of semantic functions. For example, he proposes that Source is no more or less than the argument of a FROM Path-function and that Agent just is the first argument of a CAUSE Event-function.

On the other hand, the context in which thematic roles have developed tends to focus on the syntactic behavior of specific lexical items rather than on the semantic characteristics of related classes of items—or at least on the interaction of syntax with semantics. Somers (1987, p. 214) characterizes semantic case (one of the predecessor names for thematic relations) as “an (albeit rickety) bridge between the two levels” of the syntactic behavior of words and language-independent conceptual modeling. In its most restrictive form, a theory of thematic relations holds that the lexical entries for verbs have associated with them certain thematic roles, which “roles determine the number and type of complements that a verb takes in syntactic constructions it occurs in. According to the Correlation Principle, the syntactic categories of the complements selected by a verb . . . are predictable on the basis of the type of thematic roles assigned to them” (Ravin 1990, p. 11).

Seen from this perspective, thematic relations have had many criticisms leveled against them. There is first the problem that no list of thematic relations has ever been considered complete. No sooner is an inventory of thematic relations proposed than sentences with different types of arguments not covered by the inventory are proffered, for example, Negative quality (“without humor”), Function (“as a club”), Reference (“about the war”), Price, (“for five dollars”), Extent (“for two miles”), etc. (Examples taken from Croft 1991, p. 158.) To include these new argument types could lead easily to a proliferation of thematic relations: Croft suggests that “the only way . . . to account for the wide variety of thematic roles found with different surface predicates is to analyze the meaning of the verbs that they are associated with. The more difficult to handle thematic roles are associated with very small but semantically coherent classes of verbs, such as verbs of commercial exchange for the price role” (p. 158). This proliferation of roles would, however, contradict the basic philosophy of such relations, that only a small and finite number of them is needed (p. 156). However, if there is no major expansion of the number of thematic relations, then specific thematic relations will be asked to cover a range of argument types and thus lose some of their definition. As Levin (1995, p. 58) argues, “Broadening the semantic characterization of roles to cover problematic examples is no solution since their value is diminished, while restricting them can result in the unnecessary proliferation of roles.”

Ravin (1990, p.159) presses for a theory of semantic predicates in which arguments are simply the entities represented in their meanings: “For example, *buy* is a predicate that expresses four relations among four arguments, grossly, of the following form:

Possession (seller, merchandise) at t^0
Possession (buyer, sum of money) at t^0
Possession (seller, sum of money) at t^n
Possession (buyer, merchandise) at t^n ,”

where t^0 represents the time of the initial state, when the Seller still has possession of the Merchandise and the Buyer still has possession of the Sum of money, and t^n represents a subsequent time, after the Merchandise/Sum-of-money exchange, when the Seller has possession of the Sum of money and the Buyer has possession of the Merchandise. The arguments of the COMMERCIAL TRANSACTION semantic predicate are thus a Seller, a Buyer, Merchandise, and a Sum of money.

Ravin points out that there are four possibilities open for the correspondence between semantic arguments and syntactic complements: “arguments that do not correspond to complements and are never syntactically realized; arguments that are obligatorily realized; arguments that are optionally realized; and syntactic complements that do not correspond to any semantic arguments” (p. 160). All four cases occur. However, a standard theory of thematic relations cannot deal with the first and last of these possibilities. In a similar vein, Saint-Dizier (1999) distinguishes between “linguistic arity,” based on verb complementation, and “cognitive arity,” which includes all “participants which are minimally required for the activity or state described by the predicate to be understandable” (pp. 3, 5). It is cognitive arity that is of interest to this research effort.

Most problems with casting thematic relations as semantic predicate types stem from trying to embed them into a lexically- or syntactically-based theory. Far fewer difficulties arise when thematic relations are instead given a cognitively-based reading (but then the story they tell is a different one). Jackendoff appears to have the only major proposal on the table in which thematic relations correspond to semantic argument types. Since this proposal is only incidentally a theory of thematic roles—indeed inasmuch as it depends critically on generalizing over predicates through the use of semantic primitives to express various functions—further consideration of this proposal will be delayed until we take up lexical conceptual structures theory in Section 2.2.3.

2.2.2 Semantic Verb Classes (Levin/MIT Lexicon Project)

A line of research intimately connected with predicate argument structure and semantics is that of the MIT Lexicon Project, summarized in Levin (1993b). Levin presents the fundamental premise of this work in these words: “The behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning. Thus verb behavior can be used effectively to probe for linguistically relevant pertinent aspects of verb meaning” (p. 1). Specifically this work proceeds by forming classes of verbs that pattern similarly with respect to various alternations—for example, the middle alternation, the conative alternation, the causative/inchoative alternation, the body-part possessor ascension alternation—and then trying to determine which meaning components the verbs have in common that would account for their shared syntactic behavior. For example, Levin has found that “the

middle construction is available only to . . . verbs whose meaning involves a notion of causing a change of state” (p. 5).

As an extended example of this line of research, Levin and Rappaport Hovav (1991) reports the investigation of a class of “apparently semantically related . . . verbs of removal,” which involve “the removal of a substance or physical object—the locatum—from a location” (p. 126). In-depth analysis of the syntactic behavior of these verbs revealed, however, not a single unified class of verbs of removal, but three distinct subclasses. One subclass includes verbs such as *clear*, *clean*, and *empty*. These verbs specify a locational state that results from the action named by the verb, but without making explicit how this state is achieved (pp. 129-130). Another subclass includes verbs such as *wipe*, *erase*, *rinse*, and *vacuum*. In contrast to the preceding subclass, these verbs make explicit the means by which the locatum is removed from the location, by specifying either the manner in which the removal takes place or the instrument used to effect it (p. 130). Lastly, they discuss a subclass of verbs of which *remove* is the prototype, which specify only that the locatum is caused not to be at the location, doing so without specifying either means or the resultant state (p. 132). These differences are reflected in different syntactic behavior associated with the verbs in the three subclasses. Furthermore, they suggest that when one of these verbs acquires additional senses through “lexical extension,” such that it takes upon itself the semantic character of one of the other subclasses, the verb also exhibits the syntactic behavior associated with that other subclass (pp. 137-138).

Levin notes that “the hypothesis that the syntactic behavior of a verb is fully semantically determined is not uncontroversial. Many researchers have argued that this hypothesis must be rejected, citing numerous purported counterexamples to it. Nevertheless, the meaning of a verb does have considerable predictive ability, as [many] examples . . . illustrate, suggesting that the ties between a verb’s meaning and its syntactic behavior cannot simply be ignored” (Levin 1993b, p. 12). Such controversy aside, the relevant issue here is whether the classes of verbs formed around common syntactic behavior also share semantic predicate argument structure.

In favor of answering the question positively is Levin and Rappaport Hovav’s characterization of verbs of removal, given primarily in terms of a locatum and a locational state, to which might be added an agent, and/or an instrument or a manner. There is no controversy here: These are semantic arguments of a removal predicate.

Even though there is ample evidence that the classes of verbs that pattern similarly with respect to diathesis alternations are semantic in nature, Levin is careful to note that these semantic classes are not the only semantic classes that could be formed: “Determining the appropriate meaning components is not easy, since a priori it is possible to classify verbs in many ways according to their meaning” (p. 13). Moreover, she acknowledges that “the important theoretical construct is the notion of meaning component, not the notion of verb class.” Rather, the grouping of verbs represented in Levin (1993b) has the pragmatic motivation of “[furthering] the isolation of meaning components.” In seeking this goal, “the classification system does not take into account every property of every verb” (p. 18).

As can be seen, the verb classes in Levin (1993b), while semantic in basis, are not explicitly based on the paraphrase criterion. (1) Verbs known to be related on this basis, for example, *send* and *receive*, on the one hand, and *buy* and *sell*, on the other hand, do not group together on the basis of syntactic behavior.⁹ (2) We saw in Section 1.4 that specific semantic arguments may not be expressed through a common syntactic means. (3) Moreover, Levin and Rappaport explicitly note the distinction between groups of verbs related through the paraphrase criterion and the semantic classes of verbs formed on the basis of syntactic behavior: “A single real-world event may be described in different ways, necessitating the use of verbs from different linguistically significant semantic classes” (1991, p. 134).

In short, even though the verb classes in Levin (1993b) are semantic classes and may at times be explained in terms of semantic arguments, they cannot be relied on to identify comprehensive sets of verbs related through the paraphrase criterion. The semantic argument structure of verb classes based on syntactic behavior parallels the semantic argument structure of verb classes based on the paraphrase criterion only in part.

⁹One of Levin’s general classes (13) groups together Verbs of Change of Possession. Subclasses within this general class include *Give* verbs (13.1), which include *sell*, and *Get* Verbs (13.5.1), which include *buy*. The most general of Levin’s classes correspond to very broad situation types and represent semantic groupings more than syntactic groupings. All breakdowns below the most general level are based on syntactic behavior. Note that there is no grouping within this classification that corresponds to COMMERCIAL TRANSACTION verbs *per se*.

2.2.3 Lexical Conceptual Structure

Saint-Dizier (1999) characterizes Jackendoff's theory of Lexical Conceptual Structure (LCS) as having been "designed to represent the meaning of predicative elements and the semantics of propositions" (pp. 31-32). According to LCS theory, language is organized into several autonomous levels of structure, of which one is the semantic/conceptual level (Jackendoff 1987, p. 372). Included within the model are a set of primitive conceptual categories such as Thing, Event, State, Place, Path, Property, Purpose, Manner, Amount, and Time. These primitive *types* are then "specialized" into lower-level primitives, for example, GO, STAY, BE, and ORIENT (Dorr 1993, p. 97). Formation rules expand the primitive types into more complex expressions involving lower-level primitives, as exemplified below (Jackendoff 1987, p. 375):

PLACE → [Place PLACE-FUNCTION (THING)]

PATH → [Path { TO/FROM/TOWARD/AWAY-FROM/VIA } ({THING/PLACE})]

EVENT → {[Event GO (THING, PATH)] / [Event STAY (THING, PLACE)] }

STATE → {[State BE (THING, PLACE)] / [State ORIENT (THING, PATH)] }

As can be seen, the expansion of certain primitives may include other primitives, a feature which is central to the compositional nature of Jackendoff's conceptual structures. Thus, as pertains to the issues of interest here, verbs in "a semantically coherent class [will] have common elements in their decomposition" (Levin 1995, p. 55).

As noted previously, Jackendoff adopts the usage of thematic roles in his model. However, they do not enjoy the status of primitives, but are considered instead as "relational notions defined structurally over conceptual structures" (Jackendoff 1987, p.

378). Thus, for example, a Source is the argument of the Path-function FROM, while a Goal is the argument of the Path-function TO. Jackendoff implicitly noted sizable gaps in the standard inventory of thematic roles when he asserted that although there is no conventional name for the arguments of, for example, OVER/ACROSS, TOWARD, and THROUGH Path-functions, “their conceptual roles are perfectly well defined and fall out of the general account of Path-functions” (p. 378). By treating thematic roles as derivative rather than primitive notions, Jackendoff appears to have resolved the problem of thematic role proliferation, since the expansion of the thematic role inventory is altogether principled.

Furthermore, as an outgrowth of the autonomy of the conceptual and syntactic levels of language, Jackendoff notes that lexical items other than noun phrases can fill thematic roles (p. 379); indeed “every content-bearing major phrasal constituent (sentence, noun phrase, adjectival phrase, prepositional phrase, and so on) corresponds to a conceptual constituent of some major conceptual category” (p. 376).

Jackendoff has borrowed from Gruber (1976) the premise that motion and location are conceptually primitive and can be generalized to other fields, for example, possession, thought transfer, and communication.¹⁰ Jackendoff has been criticized for forcing everything into a motion and location perspective, but such a criticism ignores Jackendoff’s division of thematic roles into different tiers. Motion and location are

¹⁰Although neither Gruber nor Jackendoff cast the phenomenon in quite this way, their premise is quite consistent with the view of metaphor in cognitive linguistics; Lakoff (1993, p. 27), for example, suggests that it is metaphor that rescues thematic roles from endless proliferation.

treated within the *thematic tier*, while agent and patient relations are dealt with within an *action tier* (Jackendoff 1987, p. 395). This relates to the point made by Levin (1985, as explicated in Dorr 1993, p. 111) that “a complete analysis of an event needs to take both the causal dimension and the motion/location dimension into account.”

Not only does Jackendoff distinguish between various tiers or dimensions, which give greater definition to thematic roles, but he also breaks down monolithic thematic roles into subroles. For example, he distinguishes at least three Agent subroles:

- (1) Actor, the first argument of ACT;
- (2) Volitional Actor, the first argument of [$\begin{smallmatrix} \text{ACT} \\ \text{VOL} \end{smallmatrix}$]; and
- (3) (extrinsic) Instigator of Event, the first argument of CAUSE.

The representation of semantic predicate argument structure can exist on various levels of granularity. On the one hand, the conventional view of thematic roles provides for a general/abstract level of semantic roles. On the other hand, if verb classes based on syntactic behavior used full-blown accounts of the syntactic behavior and meaning components of verbs, each verb could end up in a class all of its own (Levin 1993b, p. 18), and its predicate-argument structure could be defined very precisely. Neither approach accords with what is needed for the current task: Abstract semantic roles do not differentiate sufficiently to account for paraphrase relationships and highly specific semantic roles would tend not to allow for paraphrase relationships at all. What is needed instead is a level of representing semantic predicate argument structure at some in-between level.

The trend in LCS theory to distinguish between, for example, thematic and action tiers/dimensions, on the one hand, and between subroles of a general semantic role, on the other hand, produces a semantic representation of intermediate granularity. By itself, that characteristic is, of course, insufficient to prove the adequacy of LCS representations for expressing paraphrase relationships. Consequently, we consider the representation suggested by Busa et al. (1999, pp. 66-67) for two verbs, *buy* and *sell*, known to stand in a paraphrase relationship with each other:

buy (Fr. acheter):

[event CAUSE ([thing I],
 [event GO+poss ([thing J],
 [path FROM+poss ([thing K]), TO+poss (thing I)]),
 GO+poss ([thing VALUE-OF (J)]¹¹),
 [path FROM+poss ([thing I]), TO+poss ([thing K])]])])]

sell (Fr. vendre):

[event CAUSE ([thing I],
 [event GO+poss ([thing J],
 [path FROM+poss ([thing I]), TO+poss (thing K)]),
 GO+poss ([thing VALUE-OF (J)]),
 [path FROM+poss ([thing K]), TO+poss ([thing I])]])])]

¹¹Saint-Dizier gives “money” instead of “VALUE-OF (J)” in his representation of buy/acheter, but does so only because he is building up the representations incrementally. The representation “VALUE_OF (J)” is meant to be more expressive than “money” and applies equally to both verbs.

The representation for *buy* indicates that I causes the transfer of possession of J to go from K to I and the possession of money to go from I to K. The representation for *sell* indicates that I causes the transfer of possession of J to go from I to K and the possession of something of the value of J to go from K to I. What needs to be made clear here is that I, J, and K are variables, not constants. Comparing the relationships between the two representations reveals that the I of *buy* maps to the K of *sell*, while the K of *buy* maps to the I of *sell*; J remains the same across the two representations. The two representations do in fact reveal the shared semantic argument structure underlying the two verbs. Each includes four semantic roles: the thing (i.e., Merchandise) whose possession changes hands, the thing (i.e., Money) of value commensurate with the merchandise whose possession changes hands in the opposite direction, and two other things (i.e., the Buyer and the Seller) which are, respectively, the FROM pole of the money transfer/the TO pole of the merchandise transfer, and the TO pole of the money transfer/the FROM pole of the merchandise transfer.

There are several respects, mostly intertwined, in which these representations are, however, not optimal for the paraphrase task. First is the fact that there are multiple representations to be dealt with, not just a single semantic representation. The paraphrase criterion specifically posits that a single representation will suffice for equivalent meanings. Second, the reason there are multiple representations is that each lexical item has its own associated representation. These are, after all, *lexical* conceptual structures. The representations reveal that the Instigator (first argument of CAUSE) of buying is not the Instigator of selling. This suggests that the semantic representation is not completely

autonomous, but carries an internal link to syntactic behavior. Third, the semantic roles are not isolated in a way presumed to be intuitive for human understanding. However, in place of informative labels (e.g., Buyer, Seller, Merchandise, Money) the LCS representation includes explicit information about the logic of the roles. It is the encoding of these logical relationships that permits recognition of the significant degree of commonality between the representations. While LCS representations may not be exactly what the doctor ordered, they clearly have the potential of expressing something quite close. The number of operations required to map between LCS representations and a conceptual structure independent of specific lexical items might be quite small.

2.2.4 Frame Semantics

The approach to verb classification most consistent with the paraphrase criterion adopted here is Fillmore's theory of frame semantics. The basic notion underlying frames is an "appeal . . . to structured ways of interpreting experience, . . . an alternative to the view that concepts or categories are formed through the process of matching sets of perceptual features" (Fillmore 1976, p. 20). Elsewhere Fillmore describes a semantic frame as "any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits" and "a system of categories structured in accordance with some motivating context" (Fillmore 1982, pp. 111, 119). Thus, frame semantics is a theory of language understanding, situated within a world in which some types of experiences (for example, events, conditions, relationships) recur frequently and are structured. Fillmore suggests that experiences are

memorable “precisely because the experiencer has some cognitive schema or frame for interpreting it. This frame identifies the experience as a type and gives structure and coherence—in short, meaning—to the points and relationships, the objects and events, within the experience” (1976, p. 26).

The recurrence of prototypical experiences and the cultural or institutional meaning they have motivates the introduction of linguistic expressions to refer to the whole of these experiences, as well as to their salient parts or aspects. Such words or phrases are said to ‘evoke’ the frame and inherit its organization, while the frame structures the meaning of those words and phrases (Fillmore 1982, p. 117). While any given word will typically highlight (or profile) particular aspects of the frame, the whole of the frame structure is conveyed by each word or phrase that evokes/inherits it.

In the case of events, for which the richest frames exist, the internal structure of such scenarios typically includes some number of roles (that is, participants in the event, identified by the functions they play in the event) and may also include attributes of the event and subevents of the event. For example, Fillmore (1982, p. 116) analyzed the JUDGING frame as consisting of:

a person who formed or expressed some sort of judgment on the worth or behavior of some situation or individual (. . . the Judge); a person concerning whose behavior or character it was relevant for the Judge to make a judgment (. . . the Defendant); and some situation concerning which it seemed relevant for the Judge to be making a Judgment (. . . the Situation).

The Judge and the Defendant are examples of roles within the frame; the Situation is a subevent, requiring its own separate frame representation (depending on the semantic type of the situation being judged); attributes of the event include the degree of responsibility that the Judge ascribes to the Defendant for the Situation, as well as the nature of the Judge's judgment of the Situation—Was it positive or negative? In like manner, the COMMERCIAL TRANSACTION frame includes as its elements “a person interested in exchanging money for goods (the Buyer), a person interested in exchanging goods for money (the Seller), the goods which the Buyer could or did acquire (the Goods), and the money acquired (or sought) by the seller (the Money)” (Fillmore 1982, p. 116).

In a significant theoretical overview of frames, Barsalou proposes that they are “the fundamental representation of knowledge in human organization” (1992, p. 21); inasmuch as the various parts of a frame can also be described by frame structures, he further asserts that “human conceptual knowledge appears to be frames all the way down” (p. 40). Given the “combinability [of frames] into larger conceptual structures” (Fillmore 1976, p. 30), it may be that human cognition involves frames all the way up as well. Barsalou's overall thesis is that significant psychological evidence points to the inadequacy of “flat”, feature list representations of human cognition; frame representations not only avoid the problems encountered by feature list representations, but also support a wide array of cognitive processing tasks. Specifically, Barsalou argues that human knowledge evidences extensive use of attribute-value sets and relationships that are absent from feature list representations, but very much present in the basic

components of frames, which include attribute-value sets, structural invariants, and constraints.

What Barsalou calls “attribute-value sets” are simply sets of instantiated frame slots/roles/participants, which is the aspect of semantic frames of interest to us here. Barsalou notes that some of these attributes/slots/roles/participants can be said to constitute the frame’s core, in that they frequently have specific values and may indeed be conceptually necessary: How, for example, can one have a COMMERCIAL TRANSACTION event without considering a Buyer, a Seller, Merchandise, and Money as payment? Other attributes/slots/roles/participants may occur in the representation of some frame instantiations, but not all. Such frame elements would be considered optional.

By way of summary, the overall goal of the research being proposed here is to identify the predicate argument structure of semantic classes of verbs based on the paraphrase relationship. Given what has been established pertaining to possible theoretical approaches to semantic predicate argument structure, this goal can be made more specific: to identify semantic frames and their internal participant structure. This identification process involves both extensional and intensional subprocesses: Semantic frames will be investigated by attempting to identify sets of verb senses that evoke the same frame (the extensional subprocess); the semantic predicate argument structure corresponding to each frame will be sought through identification of participant types that tend to occur with the set of verbs that evoke the same frame (the intensional subprocess).

2.3 FrameNet

FrameNet (Baker, Fillmore, and Lowe 1998; Fillmore and Baker 2001; Fillmore, Wooters, and Baker 2001; FrameNet; Johnson et al. 2002) is an NSF-funded project to create a frame-semantics-grounded lexical resource for English. The aims of the project are threefold, each aim being tied to a specific deliverable. The first aim is to produce a sizable inventory of semantic frames, each including a frame name; a list of frame elements (a.k.a. slots) that comprise the human participants, inanimate props, and other roles within the frame; and a brief list of lexical items that evoke the frame. The second aim is to create an annotated subcorpus¹² in which the annotation documents, for specific word senses, the surface phrase in which it occurs, its grammatical usage, and the frame element it realizes, thus recording lexical-syntactic-semantic triples. The third aim is to create a lexical database based on the annotation results that records “the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses” (Johnson et al. 2002).

The first aim of the FrameNet project—to produce an inventory of semantic frames—is closely related to one of the objectives of the research reported here.¹³ However, the two approaches use vastly different means to generate frames. The frames

¹²Up to this point, texts for annotation have been drawn from the British National Corpus (<http://info.ox.ac.uk/bnc>) and the LDC North American Newswire corpora.

¹³The second and third aims of the FrameNet project reflect an interest with the interface between frames as conceptual structures and syntactic constructions. This line of research is crucial to the analysis and generation stages leading to and from frame semantic representations of propositional content. Gildea and Jurafsky (2002) describe a system using statistical classifiers, trained on FrameNet annotations, to identify which semantic roles specific constituents fill.

in FrameNet are initially based on “armchair analysis,”¹⁴ then modified as required by experience with using the proposed frames for corpus annotation.

Three observations arising from the FrameNet project are of especial interest here. The first concerns the primary role of verbs and the dependent role of nouns in the frame semantics context. The FrameNet investigators note that, while several semantic frames may be active in a particular sentence, it is usually the frame evoked by the verb that is the most prominent. As a result, analyzing the syntactic arguments of verbs bears the most direct relationship to the delineation of the elements within a frame. A further consequence of the primacy of verbs as predicing words is that they, along with deverbal nouns, typically receive significantly fuller description in the entries in the FrameNet database than do other (dependent) nouns and adjectives.

The second observation concerns the granularity of the semantic frames developed within the FrameNet effort. In the first stage of the project (FrameNet I), frames were organized around broad categories of human experience, for example, COGNITION, COMMUNICATION, JUDGMENT, and PERCEPTION; generalizations over such categories of experience were emphasized. In the second stage of the project (FrameNet II or FrameNet++), greater emphasis has been put on capturing the nuances of specific lexical items. The FrameNet investigators acknowledge that, with sufficient time for in-depth analysis, every word sense could end up with its own unique frame. In this second stage

¹⁴“Work on a new frame begins with the native speaker analyst’s intuitive judgment that some particular conceptual pattern underlies one or more lexical units in the language in a systematic way”(FrameNet (2004), FAQs).

of the research effort, greater consideration is given to splitting a proposed frame into multiple frames of a more specific nature than in generalizing over more specific frames to detect underlying unity. A parallel effort of identifying inheritance, composition, and use relationships among frames is ongoing, but appears not yet to have entered a period of systematic investigation. While lexically-specific frames will be required for natural language analysis and generation stages, the use of semantic frames for resolving the paraphrase problem will depend crucially on the identification of those frames (1) that are inherited by other, more specific, frames, (2) that compose other frames (i.e., that are parts of other, aggregate frames), and (3) that are used by other frames. These frames will tend to be the more general frames emphasized in FrameNet I.

The third observation flows naturally from the second and concerns the number of semantic frames that would exist in a comprehensive inventory. A truly comprehensive inventory of semantic frames, including frames on the most specific level, would be quite large. But there are two reasons why a smaller inventory of more general frames would itself be useful. First, as suggested above, the inventory of frames needed to address the paraphrase problem would not include the numerous collection of most specific frames. Second, semantic frames occur in a Zipfian distribution: A few semantic frames account for a much larger proportion of frame occurrences, while a vastly larger number of frames are used infrequently. A high degree of correlation between highly specific frames and infrequently used frames can be anticipated.

Chapter 3

Lexical Resources

3.1 Lexical Resources as Data Sources

Dictionaries and related lexical resources contain many data elements relevant to semantic frame identification and description and therefore warrant consideration as a major data source for this study. The data within a lexical resource entry often include differentiation of multiple word senses that a lexeme may have, and relative to each word sense, its part of speech, its definition, sentences exemplifying its use, and, in some cases, codes that reveal semantic class and syntactic behavior. This mix of semantic and syntactic information is especially relevant because of the “interdependencies between the meaning of a verb and its complement-taking properties” (Atkins, Kegl, and Levin 1986, p. 1).

Lexical resources have several advantages over corpus data for identifying both the existence of semantic frames and their internal participant structure. First, definitions and example sentences often mention their participants at a level of generality akin to a semantic type. Corpus data, however, are more likely to include instantiated participants. A lexical resource thus provides data that more directly support the identification of a frame’s internal participant structure. Second, lexical resources provide a fairly consistent amount of data for all words they include. And even though there is no consensus on where to draw the line between one sense of a word and another, it would not be egregiously wrong to say also that they contain a fairly consistent amount of data about all word senses they contain. The amount of data in a corpus for specific words or

word senses is likely to vary widely. Third, lexical resources provide their data in a more systematic fashion than do corpora. However, these resources may be years (even decades) in development; moreover, most machine-readable dictionaries were first prepared to serve as print tools for human users. These factors limit the consistency with which lexical resources in general and dictionaries in particular present their data (Atkins 1993, pp. 37-38; Levin 1993a, pp. 76-77).

3.1.1 Definitions

Perhaps the most basic of lexicographic rules is that one should define a lexical item in terms of its genus and differentiae; the genus should pinpoint the essence of the word, while the differentiae distinguish the meaning of the word from the meanings of other words of the same genus (Landau 1989, pp. 120-121). In the case of nouns, differentiae are typically characteristics of physical form or function, while in the case of verbs, differentiae more typically refine the genus in terms of manner (e.g., how an activity or process is done) or instrument (e.g., what device is used to achieve a state or carry out an activity or process).

Verbs are thus defined in terms of other verbs, in the infinitive (p. 141). The definition should make certain aspects of the verb's meaning and use clear: "If the sense defined is transitive . . . the form of the definition should leave little doubt that an object is called for to complete it and should suggest, if possible, the nature of the object. . . . An intransitive verb must be defined intransitively . . . by including the object as part of the definition, or by using another intransitive verb as a synonym" (pp. 141-142).

Syntactic arguments of the verb(s) used in a definition often correspond to semantic (and usually syntactic) arguments of the verb being defined. Note that when intransitive verbs are defined in terms of a transitive verb and object, the object in the definition will correspond to only a semantic argument of the verb being defined and not also a syntactic argument. For example, Table 2 gives the definitions of several verb senses¹⁵ that evoke the COMMERCIAL TRANSACTION frame, which includes as its semantic arguments a Buyer, a Seller, some Merchandise, and Money. Words corresponding to the Merchandise (*property, goods*), the Money (*money, value*), and the Buyer (*buyer, buyers*) are present in the definitions; however, no words corresponding to the Seller are present in these definitions.

Verb sense	LDOCE Definition
buy1.1	to obtain (something) by giving money (or something else of value)
buy1.2	to obtain in exchange for something, often something of great value
buy1.3	to be exchangeable for
purchase1.1	to gain (something) at the cost of effort, suffering, or loss of something of value
sell1.1	to give up (property or goods) to another for money or other value
sell1.2	to offer (goods) for sale
sell1.3	to be bought; get a buyer or buyers; gain a sale

Table 2. Definitions for Verbs Evoking the COMMERCIAL TRANSACTION Frame

¹⁵LDOCE verb senses are identified throughout in the form [*word form*][*word number*].[*sense number*]. While LDOCE uses numbers after word forms also to distinguish between words with different parts of speech, word numbers are used here only to differentiate among homonymous verbs.

However, while arguments appearing in a verb's definitions will generally correspond to semantic arguments of that verb, it is not necessarily the case that all semantic arguments of the verb being defined will show up in the definition. Indeed, certain arguments are typically only implied. Missing from a verb's definition in most cases is indication of the argument that fills the subject slot in active voice. Also missing from the definition of a transitive verb may be explicit mention of the argument filling the direct object slot.¹⁶ At times, however, the subject and/or direct object arguments are included parenthetically in a definition for the purpose of revealing selectional restrictions that pertain to them. However, parentheses are also used in definitions for purposes totally unrelated to predicate argument structure.

Depending on how such definitions are written, missing arguments of verbs might be recovered by searching for nouns which use those verbs in their definitions in the appropriate form. Consider a noun that corresponds to the subject of some verb, as, for instance, *buyer* corresponds to the subject of *buy*. Such a noun might be defined as "*a person who Vs.*" Another heuristic for identifying the semantic argument corresponding to the subject of a verb might be to probe for the existence of nouns generated through suffixation of such morphemes as '-er' onto the verb whose subject argument is sought.

¹⁶Fortunately, in at least some cases, a transitive verb whose definition lacks mention of the argument filling the direct object slot may have an intransitive polysemous counterpart that does mention the argument. Fellbaum (1998b, p. 86) gives as an example "the transitive superordinate *drink* [, which] takes noun arguments that are hyponyms of *liquid* or *beverage*. The intransitive subordinate verb has the more specific sense 'to drink alcoholic beverages.'" The definition of the intransitive sense of the verb is likely to explicitly contain a noun within the *liquid/beverage* hierarchy.

Somewhat parallel to this situation is the case of verbs whose direct object argument is sought, as, for instance, *merchandise* corresponds to the direct object of *buy*. Such a noun might be defined as “*that which is Ved.*” Lexical items corresponding to a verb’s direct object are less likely to be related to the verb through productive morphological processes than is the case with a verb’s subject.

At the same time that some verbs are given analytical definitions, which include material corresponding to arguments of the verbs, other verbs are defined simply in terms of their synonyms. For example, *purchase* might be defined simply as ‘BUY’. In these cases, the argument structure (both syntactic and semantic) of the synonym used in the definition is assumed to be the argument structure of the verb being defined.

3.1.2 Example Sentences

Example sentences are intended to illuminate the user’s understanding of a word by placing it in one or more typical contexts. What may be ambiguous in a definition may be made clearer in a sentence.

Atkins, Kegl, and Levin claim that “example sentences often use nouns that are chosen to emphasize the semantic relation of an argument to the verb, as well as to represent the prototypical semantic restrictions on this argument” (1986, p. 17). For example, WordNet gives as an example sentence for *pay*, “I paid four dollars for this sandwich.” This demonstrates that the Buyer (“I”) is typically animate, indeed is typically a person, and that the two items involved in the exchange are prototypically a sum of money (“four dollars”) and an inanimate object (“sandwich”). However, only the

sum of money is expressed specifically enough in this particular example sentence to indicate much of the semantic argument structure *per se* of COMMERCIAL TRANSACTION verbs.

3.2 Longman Dictionary of Contemporary English (LDOCE)

Of available machine-readable dictionaries, a particularly useful one is the *Longman Dictionary of Contemporary English* (LDOCE); indeed Boguraev and Briscoe have characterized LDOCE as “uniquely suitable for computational lexicography” (1989, p. 2). The value of this resource stems from several specific advantages. First, in being made available in machine-readable form, LDOCE was reformatted (‘lispified’) (Alshaw, Boguraev, and Carter 1989, pp. 47-48) so as to make the internal structure of the dictionary entries more amenable to analysis.¹⁷ Second, LDOCE uses a restricted vocabulary of about 2000 words.¹⁸ The use of this vocabulary makes it more likely that words with closely related meanings will use the same words in their definitions and thus support the pattern of discovery envisioned. Cowie (1999, pp. 112-113) notes that the vocabulary of LDOCE’s example sentences also comes from this controlled vocabulary; he further notes that the example sentences are designed to illustrate predicate argument

¹⁷This is not to say that the machine-readable LDOCE is a machine-tractable dictionary, but simply to indicate that it is more tractable than the prototypical machine-readable dictionary.

¹⁸Complicating the use of this restricted vocabulary is the “the liberal use of derivational morphology to extend [it]” (Boguraev and Briscoe 1989, p. 16). Another problem is the common use of phrasal verbs in the definitions, although the verb and particle appear only separately in the controlled vocabulary list (Cowie 1999, p. 111). What the verb contributes to the meaning of a phrasal verb is often not a central sense.

patterns. Third, “LDOCE is unique amongst MRDs [machine-readable dictionaries] in providing a formal representation of some aspects of meaning” (Boguraev and Briscoe 1989, p. 17), which includes subject field codes and semantic codes, described below.

As a lexical database, LDOCE consists of a series of structured entries for each of approximately 40,000 words, many of which are further divided into a series of subentries for the various senses of the word. The examples in Figure 1 are representative of verb entries in the lispified version of the dictionary.

The lines marked ① and ② represent headwords for entries. The format in ① is by far the more usual, giving a single spelling for the word, while the format in ② indicates that two spellings are allowed: both *anglicize* and *anglicise*; the first form of the headword for each verb has been used here to construct identifiers for verb senses. With

①	((clarify)
③	(1 C0141600 !< clar *80 i *80 fy)
④	(3 !"kl *67 r9faI)
⑤	(5 v !<)
⑦	(7 100 !< I *DE !; T1 !< ---- !< ----H----T)
⑧	(8 to (cause to) become clearer and more easily understood : *46 When will the government clarify its position on equal pay for women !?)
⑦	(7 200 !< T1 !< FOZC !< ----H----E)
⑧	(8 to make (a liquid !, butter !, etc !.) clear or pure by removing unwanted substances))
②	((anglicize !, *45 - ise)
③	(1 A0105300 !< an *80 gli *80 cize !, *45 -ise)
④	(3 !" *67 NgI9saIz)
⑤	(5 v !<)
⑥	(6 T1 !; (I *DE) !<)
⑦	(7 0 !< !< LNAP !< ----H----T)
⑧	(8 to (cause to) become English in appearance !, sound !, character !, etc !.))

Figure 1. Representative Verb Definitions from LDOCE

the entry for (*anglicize* !, *45 - *ise*) we are introduced to one of the problematic aspects of the machine-readable dictionary: The machine-readable version of the dictionary has been created from the printers' tapes and hence includes various codes for bolding, italicization, small caps, etc., that are generally extraneous for our purposes; similarly we find that all characters (e.g., punctuation) with special meaning to Lisp are preceded by an exclamation point.

Lines ③ and ④, giving serial number and syllabification, on the one hand, and pronunciation, on the other hand, are of no concern for our task.

Line ⑤ indicates the word's part of speech, and it is the 'v' in "(5 v !<)" that identifies each of these entries as a verb.

Line ⑥ gives grammar codes for each word. Since the grammar codes are specific to individual words and often do not generalize over the set of verbs that evoke the same semantic frame, their usefulness for the research undertaken here has been found to be quite limited. As Cowie (1999) puts it, "the relationship between frame elements and grammatical categories or functions is sometimes far from straightforward" (p. 142).

Lines ⑦ and ⑧ occur in pairs for each sense of the word. When there are two or more senses, grammar codes are to be found in ⑦ instead of in ⑥; additionally, line ⑦ gives the sense number. Line ⑦ ends with a subject field code (e.g., "FOZC" in the second sense of *clarify*, "LNAP" in the entry for *anglicize*), which is of potential use in grouping together verb senses that evoke the same frame semantic structure. Subject field codes may indicate up to 2 general subject domains (e.g., "MDON" = "MD" for Medical and "ON" for Occupation) or a general subject plus a specialization, (e.g.,

“MDZA” = Medical Anatomy). It should be noted, however, that the coding has been criticized for being both incomplete and inconsistent (Boguraev and Briscoe 1989, p. 17); indeed, subject field codes are present for only 30% of the verb senses (EAGLES 1998, “The Longman Dictionary and Thesaurus”). For example, the first sense of *clarify* has no subject field code assigned, as indicated by “----” in the third field of line ⑦ for sense 100.

Line ⑦ also includes a semantic code that gives selectional restrictions for the subject and object(s) of the verb. From the semantic codes given—“----H----T” for the first sense of *clarify* and for *anglicize*, “----H----E” for the second sense of *clarify*—we know that the first sense of *clarify* has at least two semantics arguments, one which is Human (H), the other which is Abstract (T). Unfortunately, many of these selectional restrictions are too general to be of much use in identifying types of semantic arguments.

Line ⑧ gives the word’s definition and may also give example sentences or references to diagrams in the printed work.

3.3 WordNet

WordNet (Fellbaum 1998a, 1998b; Miller 1998) is a machine-readable lexico-semantic database covering English nouns, verbs, adjectives, and adverbs, a cross of sorts between a machine-readable dictionary and a thesaurus; version 1.7.1 of WordNet has been used for the work reported here.¹⁹ The dictionary-like qualities of WordNet include

¹⁹WordNet 2.0 has recently been issued and incorporates features that this description indicates are missing from WordNet. The incorporation of WordNet 2.0 into this research is addressed in Section 8.1.1.

its broad coverage of lexical items in the general vocabulary of English, organized around word senses. However, WordNet omits much of the information typically found in dictionaries (e.g., pronunciation, morphology, etymology), emphasizing instead the semantics of the English lexicon. Despite the original intent of building a purely semantic resource, syntactic information—for example, in the form of verb frames—is increasingly being added to WordNet. The thesaurus-like qualities of WordNet include its relational structure (which operates only within parts of speech up through version 1.7.1), with an emphasis on synonymy. Indeed, the synset—a set of synonymous word senses, or alternatively, a concept that has been lexicalized—is the primary organizational structure within WordNet.

A limited number of relationship types (e.g., antonymy, hyponymy/troponymy, meronymy/entailment), mostly applicable at the concept level, are also used to interrelate synsets within specific parts of speech. Perhaps the most important of these, organizing the noun hierarchy, is the subsumption relationship, represented in WordNet by two inverse relationships, hyperonymy and hyponymy. Hyperonymy is the lexical form of the IS-A and IS-A-KIND-OF generalization relationships,²⁰ in which every entity belonging to one category (named by a particular label/lexical form) belongs also to its subsuming, superordinate category (named by a more generic label/lexical form); hyponymy is the lexical specialization relationship. At the top of the (not mutually exclusive) tree structures formed by the many subsumption relationships linking noun synsets are a set

²⁰WordNet does not distinguish between the formal/taxonomic IS-A-KIND-OF relationship and the telic/functional IS-USED-AS-A-KIND-OF relationship (Miller 1998, p. 35).

of 25 “unique beginners” that have subsequently been collapsed into a group of 11 “top” nouns (entity, abstraction, psychological feature, natural phenomenon, activity, event, group, location, possession, shape, and state; most of these are unique beginners as well—only entity, abstraction, and psychological feature are aggregates of multiple unique beginners). The noun hierarchies thus formed may extend to as many as 10-12 levels deep.

Despite the many types of relationships in WordNet, the relationality at the heart of the semantic frame is not made explicit. This is what Roger Chaffin has dubbed the “tennis problem” (Miller 1998, p. 34):

Suppose you wanted to learn the specialized vocabulary of tennis and asked where in WordNet you could find it. The answer would be everywhere and nowhere. Tennis players are in the noun.person file, tennis equipment in noun.artifact, the tennis court is in noun.location, the various strokes are in noun.act, and so on. Nouns that co-occur in discussions of tennis are scattered around WordNet with nothing to pull them together. Other topics have similarly dispersed vocabularies.

Thus, nouns have not been organized in WordNet so as to help identify the various participants within a topic of discourse, as a semantic frame does. The hierarchical structure may, however, be used to good advantage in identifying various expressions that refer to the same participant: It is commonplace for a human to refer to the same entity at varying levels of specificity (for example, “The sentence *I gave him a good novel, but the book bored him* is easily understandable . . . because [of the] lexical knowledge that a novel is a book” [Miller 1998, p. 33]).

Verbs also suffer from the “tennis problem,” although perhaps not to the same extent as nouns. Verbs in WordNet have been organized into a group of 15 semantic domains

(Fellbaum 1998b, p. 70). Thus, we can anticipate to some extent that verbs that evoke the same frame will be in the same semantic domain; still there are many exceptions. (Of course, with only 15 domains, we could not assume that verbs in the same semantic domain evoke the same frame, except perhaps at the most general level.) After acknowledging that WordNet's organization by part of speech precludes a full frame semantic approach, Fellbaum suggests that the relationships in WordNet "reflect some of the structure of frame semantics. For example, WordNet relates verbs like *buy* and *sell*, which are part of a common frame (the "COMMERCIAL TRANSACTION" frame) . . . In fact, both frame semantics and the relational semantics in WordNet share a great deal with semantic field analysis in that they all naturally relate words and concepts from a common semantic domain" (1998a, p. 5). Through the relational structure of WordNet, *buy*, *purchase*, *sell*, and *pay* are related together, creating a "lexical chain" (Morris and Hirst 1991): *buy* and *purchase* comprise one synset; they entail *paying* and are opposed to *sell*. Their relationship to other COMMERCIAL TRANSACTION verbs—for example, *cost*, *price*, *charge*—is, however, not made explicit in WordNet. It is conceivable that other frame semantic relationships between verbs/lexical chains of verbs might be uncovered by taking into account terms used in the glosses and example sentences (see Harabagiu and Moldovan 1998).

Chapter 4

Identifying Semantic Frames by Extension: Task 1

Prior to the FrameNet project, identification of semantic frames with their attendant participant/slot structure was pursued entirely in an ad hoc manner, and the overall scope of those efforts was limited. Even within FrameNet the initial generation of semantic frames and slots (a.k.a. frame elements) has continued to be based on native speaker intuition, with empirical data being used for subsequent validation and revision. Optimally we would have a method for deriving semantic frames that is both systematic and reasonably universal in scope. One goal of the SemFrame research effort is to devise a methodology for automatically identifying an inventory of semantic frames, which can then be used to establish a basic thesaurus of such conceptual structures.

This work draws on the dual observations (1) that the various verb (sense)s that may be used to communicate about a given semantic frame (as *buy, sell, pay, cost, spend*, etc., could be used to communicate about a commercial transaction) tend to accept more or less the same sets of lexical items among their syntactic arguments and (2) that the semantic classes of these lexical items tend to define the frame's semantic arguments (in the COMMERCIAL TRANSACTION example, these would be a Buyer, a Seller, Merchandise, and Money).

The effort to induce semantic frames from such lexical resources as the *Longman Dictionary of Contemporary English* (LDOCE) and WordNet involves two distinct tasks corresponding to these two observations. In the first task, discussed in this chapter, semantic frames are identified extensionally, by enumerating for each frame the set of

(LDOCE) verb senses that evoke the frame; these are sets of verb senses that can be used to communicate about the same propositional content and that tend to accept the same set of lexical items among their arguments. Sets of word senses that evoke a common frame will be referred to as *framesets*. In the second task, discussed in Chapter 5, the frames proposed in task 1 are fleshed out: The frame is given a label/name; senses of nouns that evoke the frame are enumerated; and the internal participant structure of the frame, consisting of the set of its semantic arguments, is set forth.

Being able to enumerate the membership of a verb frameset, that is, the set of verb senses that evoke a common frame, depends on access to semantic information about verb senses. Much of that information is relational in nature. If, for example, verb sense *x* evokes a specific semantic frame, verb senses associated with *x* through such semantic relationships as hyperonymy, hyponymy, antonymy, and synonymy might also evoke the frame. But other verb senses evoking the same frame are semantically related in ways that stand outside the standard set of paradigmatic relationships (e.g., the relationship between the COMMERCIAL TRANSACTION senses of *buy* and *pay*, in which buying something entails paying money for it).

For task 1 SemFrame gathers evidence about semantic relatedness between verb senses by analyzing data in LDOCE and WordNet from a variety of perspectives, a key element of the SemFrame approach. The overall approach is represented graphically in Figure 2. At the top of the figure are data sets from ten different sources of information, drawn either from LDOCE (LC.pairs, LI.pairs, LM.pairs, LS.pairs, LT.pairs, and LZ.pairs) or from WordNet (WC.pairs, WN.pairs, WR.pairs, and WX.pairs). Each data

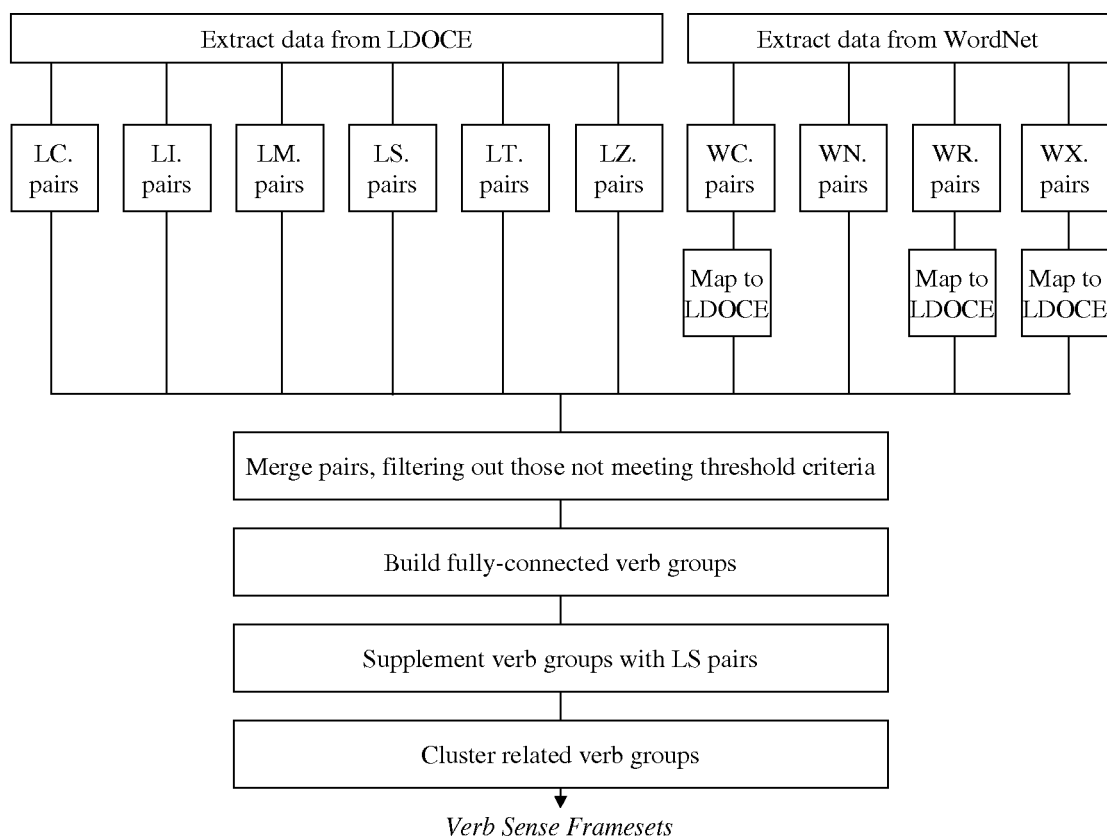


Figure 2. Overview of Task 1

set identifies verb sense pairs or verb synsets that potentially evoke the same frame; the data sets strive for high recall in the context of particular clues to semantic relatedness. The WordNet verb synset pairs undergo a mapping process to transform them into LDOCE verb sense pairs. All ten data sets are then merged into a single data set, and only those verb sense pairs whose cumulative support exceeds specific thresholds for either number of supporting data sources or strength of support are retained, thus achieving higher precision in the merged data set than in the ten input data sets. The interconnections among the verb sense pairs in this data set are analyzed to build

complete-link verb sense groups: The membership criterion for each of these groups is that a verb sense must have been paired with every other group member in the single, merged data set. The membership of these groups is supplemented with pairings from an especially strong (i.e., clean) data set. The verb sense groups are then subjected to a clustering process, which results in reducing the number of verb sense groups by merging together those groups whose similarity (due to overlap in membership) exceeds a clustering threshold. The verb senses within a group are hypothesized to evoke the same semantic frame, thus constituting a frameset.

4.1 Data Sources

Each of the six LDOCE data sets and four WordNet data sets includes some number (from several hundred to tens of thousands) of pairs: The LDOCE pairs each relate two LDOCE verb senses, while the WordNet pairs relate two synsets from the WordNet verb semantic network. In order to work with both types of pairs simultaneously, the WordNet synset pairs are mapped to LDOCE verb senses.²¹ Since a synset commonly unifies specific senses of multiple words, the mapping from WordNet synsets to LDOCE verb senses has the potential of producing a larger number of LDOCE-equivalent pairs than were originally generated between WordNet synsets.

²¹The WordNet-to-LDOCE mapping is built into the original generation of the WN.pairs data set.

4.1.1 LDOCE-based Data Sets

LC.pairs Data Set

The LC.pairs data set consists of LDOCE verb senses related through clustering, a technique that groups items together that are similar to each other, based on specific features/attributes associated with the items. Here the items grouped together are verb senses, and the co-occurrence of terms in their definitions and example sentences is used for forming clusters. The assumption underlying that choice is that words that evoke a common semantic frame will tend to include the same concepts in their definitions and example sentences, viz., those that refer to the overall situation and the participant structure associated with the frame. The restricted vocabulary used in LDOCE promotes the likelihood that when a given concept occurs in multiple definitions or example sentences, the same lemma will be used to refer to the concept.

Clustering algorithms differ in several respects, including the structure of the clusters they generate, the similarity measures they use, and the clusters they form. As to cluster structure, text-based clustering often uses a hierarchical, agglomerative approach. In a first pass every item (here, verb sense) is placed in its own cluster. In each subsequent pass the two most similar clusters are linked, forming a new cluster that replaces its input clusters. The clustering continues until either a single cluster is formed or a similarity threshold fails to be met. Commonly used similarity measures include, for example:

$$Dice \ sim = \frac{2C}{(A+B)} \quad Jaccard \ sim = \frac{C}{(A+B-C)} \quad cosine \ sim = \frac{C}{\sqrt{AB}}$$

where A equals the number of features/attributes in one cluster, B equals the number of features/attributes in a second cluster, and C is the number of feature/attribute values that A and B have in common (Salton 1989, p. 318).²² The similarity measure used, however, does not produce as significant a difference in the resulting clusters as does the choice of which two clusters to merge at each stage. Four methods for determining this are common in text retrieval, namely, the single link, complete link, and group average link methods and Ward's method (Rasmussen 1992; Voorhees 1986; Willett 1988). Of these, the group average link method, which forms new clusters based on the average values of pairwise similarity measures, and Ward's method, which forms new clusters based on minimizing the increase in variance of distance from cluster centroids, have been found the more successful for information retrieval (Rasmussen 1992). The group average link method has been chosen for use in SemFrame.

The algorithm for generating LC.pairs is given in Figure 3. It takes as its input several data sets: (1) *RV*, the set of terms in LDOCE's restricted vocabulary, (2) *SW*, a set of stop words, (3) *M*, a set of (word, stem) pairs,²³ (4) *F*, a set of (word, frequency)

²²For example, the list of words (omitting only the most obvious stop words) in the LDOCE definition of *abduct* is "take away person unlawfully often force kidnap", while the list of words in the LDOCE definition of *kidnap* is "take someone away unlawfully order demand money something else safe return". A = 7 (the number of words in the definition of *abduct*), B = 11 (the number of words in the definition of *kidnap*), and C = 3 (the number of words in the intersection of the two definitions, i.e., "take away unlawfully"). According to Dice's coefficient, the similarity value between the two definitions is $(6 / 18) = .33$. For Jaccard's coefficient, the similarity value is $(3 / 15) = .2$. The cosine coefficient is $(3 / \text{sqrt}(77)) \approx .34$.

²³Stemming is implemented using both the Porter (1980) stemmer and a handcrafted set of (word, stem) pairs. For purposes of outlining the LC.pairs algorithm, the operation of the Porter stemmer is treated as if it also took the form of (word, stem) pairs.

Input.	RV , the set of LDOCE restricted vocabulary terms; SW , a set of stopwords; M , a set of ($word$, $stem$) pairs; F , a set of ($word$, $frequency$) pairs; DE , a set of ($verb_sense_id$, $def+ex$) pairs, where $def+ex_d$ = the set of words in the definitions and example sentences of $verb_sense_id_d$
Step1.	forall $d \in DE$, append to $def+ex_d$: $verb_sense_id_d$ and any subject field codes assigned to $verb_sense_id_d$
Step2.	forall $d \in DE$, remove from $def+ex_d$ any word $w \in W$
Step3.	forall $d \in DE$ forall $m \in M$ if $word_m$ exists in $def+ex_d$ substitute $stem_m$ for $word_m$ in $def+ex_d$
Step4.	forall $r \in RV$ $SR_r \leftarrow \text{strong_sense}(r, \text{threshold})$
Step5.	forall $a \in RV$ with a non-null value for SR_a forall $b \in RV$ with a non-null value for SR_b if ($a R b$) exists in WordNet for $R = \{\text{hyponymy, antonymy, entailment, cause}\}$ place a and b in the same $\text{strong_relationship_set}$, $SRS_s \in SRSS$
Step6.	forall $d \in DE$ forall $s \in SRSS$ forall $t \in SRS_s$ if t exists in $def+ex_d$ substitute combined SRS_s for t in $def+ex_d$
Step7.	forall $f \in F$ if $frequency_f > 1$, $wgt_{word_f} \leftarrow \frac{1}{frequency_f}$ else if $frequency_f = 1$, $wgt_{word_f} \leftarrow .01$ forall $s \in SRSS$ forall $t \in SRS_s$ if $t = word_f$ $wgt_t \leftarrow 2 \left(\frac{\sum_{p \in SRS_m} wgt_p}{\sum_{p \in SRS_m} 1} \right)$
Step8.	O -Voorhees' average link clustering algorithm applied to D , with initial weights forall t in $def+ex$ set to wgt_t
Step9.	forall $o \in O$ return all combinations of two members from o

Figure 3. Algorithm for Generating LC.pairs

pairs, stop words, and (5) *DE*, sets of words occurring in the definitions and example sentences of LDOCE's verb senses. The output it produces consists of sets of LDOCE verb sense pairs. While it uses Voorhees' algorithm for group average link clustering at the center of the generation process, both pre-processing of the clustering input (steps 1-7) and post-processing of the clustering output (step 9) are required. Voorhees' (1986, p. 471) algorithm²⁴ is outlined in Figure 4.

To build the initial data set *D*, 12,663 non-phrasal verb senses in LDOCE were extracted; phrasal verbs (mostly comprising a verb and a particle) were eliminated since they are seldom recognized as independent lexical items in other resources. Definitions have been cleaned up to exclude usage notes, references to pictures, typographical codes, etc. To the remaining words of the definition and example sentences are added the word being defined and any subject field codes assigned to the verb sense (as shown in step 1).

The next pre-processing step (step 2) eliminates from this data set any words occurring on one of two stop word lists²⁵, including such general terms as *somebody*, *person*, and *thing*, as well as such relator words as *and*, *to*, and *since*.

Complicating the use of LDOCE's restricted vocabulary is the "the liberal use of derivational morphology to extend [it]" (Boguraev and Briscoe 1989, p. 16). The

²⁴No general algorithm for the group average link method of clustering is known that runs in $O(N^2)$ time while using less than $O(N^2)$ space. However, in the special case that between-item similarity (in our case, similarity between verb senses) is calculated as the inner product of appropriately weighted vectors, the space requirement drops to $O(N)$ because cluster centroids, as the mean of all vectors in their spaces, can be used for computing between-cluster similarities (Voorhees 1986, p. 469).

²⁵The shorter list consists of 93 stop words, while the longer list consists of 284 stop words.

```

/* Initialize */
MaxSim ← 0;
for (i ← 1 to CollectionSize) {
    create singleton cluster i for doc i; info[i].centroid ← document i;
    ComputeSim (i, nn, sim);
    info[i].nn ← nn; info[i].sim ← sim; info [i].size ← 1;
    if (sim > MaxSim) {
        id1 ← i; id2 ← nn; MaxSim ← sim;
    }
}

/* Merge clusters until only one cluster remains or remaining sims are 0 */
while (MaxSim > 0.0 and NumActive > 1) {
    smaller ← MIN (id1, id2); larger ← MAX (id1, id2);
    info [smaller].centroid ← MergeCentroids (smaller, larger);
    info [smaller].size ← info [smaller].size + info [larger].size;
    a ← index of larger in active;
    active [a] ← active [NumActive]; NumActive ← NumActive - 1;
    MergeClusters (smaller, larger, MaxSim);
    MaxSim ← 0;
    for (each cluster, a, in active) {
        if (info [a].nn = larger or info [a].nn = smaller) {
            FindMaxSim (a, nn, sim) ← sim;
            info [a].nn ← nn; info [a].sim ← sim;
        }
        if (info [a].sim > MaxSim) {
            id1 ← a; id2 ← info [a].nn; MaxSim ← info [a].sim;
        }
    }
}

```

Figure 4. Group Average Link Clustering Method, Voorhees' Algorithm

stemming of words from the definition, using the Porter (1980) stemmer, helps to collocate what morphological variation has scattered. As this stemmer is fairly conservative, it only occasionally produces the same stem for two word forms from different lemmas. However, it often fails to produce the same stem for two word forms from the same lemma. The relatively small size of the data sets used to characterize the

verb senses increases the importance of merging together the morphological variation that the Porter stemmer fails to account for. Some of this variation is evident in the list of restricted terms given in LDOCE, which lists, for example, both *anxiety* and *anxious*, both *music* and *musician*; the Porter stemmer likewise stems these to different roots and thus fails to pull together words that are morphological variants of the same lemma. But some of the variation is completely obscured, for example, when the list of defining terms gives *bitter* to stand also for *bitterly* and *bitterness*. The Porter stemmer correctly stems *bitterness* to *bitter*, but fails to capture the relationship of *bitterly* to *bitter*, stemming *bitterly* instead to *bitterli*. Also, since the Porter stemmer only treats English suffixes, it does not capture the systematic relationship between, for example, *tell* and *retell* or *willing* and *unwilling*, pairs of words that evoke the same semantic frames. A list of equivalences between word forms appearing on the restricted vocabulary list but not recognized by the Porter stemmer has been developed by hand and applied to the data sets. After being processed by the standard Porter stemmer and the shorter and longer stop word lists (in step 3), the verb sense data sets retained 7697 and 7581 different stems respectively. After processing based on the list of further stem equivalences, the data sets were reduced to 6784 and 6682 unique stem,²⁶ respectively.

Investigation of the restricted vocabulary has further revealed that it does not altogether control for standard lexical semantic relationships that would often occur

²⁶The number of stems in the verb sense data sets far outnumber the number of words in the restricted vocabulary for several reasons. First, the data sets include 219 subject field codes. Second, the data sets include the lemmas being defined. Third, the definitions, and thus the data sets, also include many single-use words (given in SMALL BLOCK CAPS).

within a shared semantic frame. For example, all members of the following pairs (and triple) occur within the defining vocabulary: *difficult, hard; correct, right; evil, wicked; bind, fasten; cheat, deceive; look, see; say, speak, talk; different, other; nice, pleasant; and complete, all*. It is obvious that some senses of the members of these sets of words are quasi-synonymous and should be treated as equivalent for clustering purposes. It is not certain, however, that it is the equivalent senses of those words that are being used across the definitions of LDOCE.

To add to the problem, the defining vocabulary terms have been characterized as being more like ‘basic English’ than like semantic primitives (Alshawhi 1989, p. 155). These terms are thus high-frequency terms (Vossen, Meijs, and den Broeder 1989, p. 174) and highly polysemous; Wilks et al. (1993, p. 349) show that the terms in the LDOCE restricted vocabulary have an average of twelve senses, which contrasts with an average of two senses for terms not in the defining vocabulary. However, LDOCE lexicographers were to use only ‘central’ senses of these terms. As a result, approximately one half were used in only one sense, another one quarter in two senses, and only one quarter in three or more senses (Wilks et al. 1993, p. 376).

Given this set of circumstances, I have developed the notion of a *strong sense* of a word, based on frequency data in SEMCOR, a WordNet-sense-tagged extract of the Brown Corpus. For these purposes, the frequency of groups of word senses is compared (in step 4), where a word group includes all WordNet senses of a given word within the same part of speech and belonging to the same semantic domain (as indicated by ‘lexicographic file’ assignment within WordNet). Where the SEMCOR frequency of one

group of word senses dominates that of all other groups for that word, that word group—identified by lexeme, part of speech, and semantic domain—is designated a strong sense of the word. Dominance is achieved when the percentage of all occurrences of a word group in SEMCOR exceeds a threshold. Thresholds of 50%, 70%, and 90% have been investigated. Where a word has a strong sense, it is assumed to be the sense used in LDOCE definitions.

Strong senses of words are then used to discover *strong relationships* (in step 5), which occur when strong senses of nouns and verbs are interrelated in WordNet through hyponymy, hyperonymy, antonymy, entailment, cause. The interrelating is recursive, with some exceptions: Co-hyponyms (other hyponyms of one's hypernyms) and co-hypernyms (other hypernyms of one's hyponyms) are not considered to be interrelated, and no further relationships are added for strong senses interrelated through the cause relationship (which occurs but rarely anyway). For example, strong relationships occur between the following sets of LDOCE verbs at threshold = 0.7:

<i>be, exist, live</i>	<i>hear, listen</i>
<i>bind, fasten</i>	<i>know, recognize</i>
<i>bless, curse</i>	<i>laugh, weep</i>
<i>borrow, lend</i>	<i>let, prevent</i>
<i>buy, give, pay, sell</i>	<i>like, please</i>
<i>cease, continue</i>	<i>pronounce, say, speak, talk</i>
<i>fail, succeed, try</i>	<i>ride, walk</i>
<i>float, fly, sink, swim</i>	<i>shout, whisper</i>
<i>forget, remember, think</i>	<i>sleep, wake</i>

The strong senses of the words in each of these sets are assumed to relate to a shared semantic frame. Therefore, each of the words in a strong relationship set is replaced in the verb sense data sets (in step 6) with a token common to the strong relationship

(normally a string of all words in the strong relationship set separated by hyphens, e.g., *laugh-weep*).

A final pre-processing step (step 7) involves assigning an initial weight to the terms used in the verb sense data sets. Generally speaking, a term's weight is set to its inverse definition frequency,²⁷ with the following exceptions: If a term occurs in only a single definition, its weight is set very low, since it has nothing to contribute to the clustering process. If a term is part of a strong relation set, its weight is set to twice the average weight of the individual terms, on the grounds that, since these terms have already undergone clustering, the contribution of this set of terms is to be trusted.

The clustering algorithm has been executed numerous times (step 8), varying the stop word list used, the threshold for establishing strong relations, and the threshold used for establishing clusters. The output resulting from the combination of the smaller stop word list, the lesser strong relations threshold, and a fairly low cluster threshold criterion has been used to generate the LC.pairs data set. The clustering output, which may group any number of verb senses together, has been manipulated (in step 9) to establish a direct relationship between every two verb senses in a common cluster.

By way of illustration, here are two groups of LDOCE verb senses that clustered together (example sentences and phrases are given in italics):

²⁷The general assumption underlying this weighting scheme is that the greater the number of definitions and example sentences a word appears in, the less likely that it is a good indicator of which semantic frame(s) the word evokes, and conversely that the fewer definitions and example sentences a word appears in, the more likely that it is a good indicator of the semantic frame(s) evoked.

To double1.2 is (of an actor or actress) to act (2 parts in one play). *Mary agreed to double the 2 parts of the dancer and the mother.*

To premiere1.1 is to give a PREMIERE of (a play or a cinema film). To premiere1.1 is (of a play or a cinema film) to have a PREMIERE.

To star1.2 is to have as a main performer; FEATURE. *one of my favourite old films starring Charlie Chaplin.*

To star1.3 is to appear as a main performer. *Humphrey Bogart starred in a number of fine films.*

To understudy1.1 is to act as understudy to (an actor or actress) in (a part). *She understudied (Maggie Smith as) Desdemona.*

To latinize1.1 is to translate into the LATIN language.

To latinize1.2 is to use LATIN words and phrases, as in poetry. *a Latinized style.*

To pulsate1.1 is to shake very regularly. *The air seemed to pulsate with the bright light. | the pulsating beat of LATIN AMERICAN dance music.*

On the one hand, on the basis of the occurrence of words such as *actor*, *actress*, *part*, *play*, and *film* in their definitions, the LC.pairs data set establishes relationships between the appropriate senses of *double*, *understudy*, *star*, and *premiere*. On the other hand, the LC.pairs data set also establishes a relationship between senses of *latinize* and *pulsate* on the basis of the use of the word *Latin* in both their definitions; however, the *Latin* of *Latin American dance music* in the definition of *pulsate* has nothing to do with the *Latin* language of *latinize*. And because *Latin* occurs only infrequently in the overall verb sense data set, the weighting it receives is fairly high.

The LC.pairs data set consists of 13,694 pairs of verb senses that are clustered together within the same group. Although the data set contains some erroneous relationships, as just seen, its overall precision is fairly high. The incidence of egregiously incorrect groupings appears felicitously low.

LI.pairs Data Set

The algorithm for generating LI.pairs is given in Figure 5. The LI.pairs data set consists of pairs of LDOCE verb senses that are defined in terms of the same verb, operationally identified as the first verb infinitive appearing in the definition of a verb sense, as identified by the Minipar parser (Lin 2001). The assumption underlying this data set is that the verb in terms of which a verb sense is defined indicates the generic process or state that the verb sense refers to. The more specific this generic process or state, the more likely that it corresponds to a semantic frame. Verb senses defined in terms of the same verb thus have some degree of likelihood of evoking the same frame.

Input. D , a set of $(verb_sense_id, def_verb)$ pairs, where def_verb_d = the verb in terms of which $verb_sense_id_d$ is defined (identified, operationally, as the first infinitive in the definition)

Step1. forall v that exist as def_verb in D , form $DV_v \subset D$, by extracting all $(verb_sense_id, def_verb)$ pairs where $v = def_verb$

Step2. Remove all DV_v for which $|DV_v| > 40$

Step3. forall v that exist as def_verb in D ,
return all combinations of two members from DV_v

Figure 5. Algorithm for Generating LI.pairs

By way of illustration, here are two groups of verbs that are defined in terms of a common infinitive:

To caterwaul1.2 is to quarrel loudly and angrily.

To jangle1.2 is to quarrel noisily.

To scrap2.1 is to quarrel or fight.

To bicker1.1 is to quarrel, about small matters. *The 2 children were always bickering (with each other) (over / about small matters).*

To squabble1.1 is to quarrel, noisily and unreasonably.

To row2.1 is to quarrel, often noisily or violently. They're rowing again. *She always breaks dishes after rowing with her husband.*

To bandy1.2 is, in bandy words (with) , to quarrel (with).

To blink1.1 is to shut and open (the eyes) quickly, once or several times. *She blinked (her eyes) as the bright light shone on her.*

To pen1.1 is to shut (animals) in a PEN.

To cloister1.1 is to shut away from the world in or as if in a CONVENT or MONASTERY. *He had led a cloistered life in one of our older universities, and knew little of practical affairs.*

To slam1.1 is to shut loudly and with force. *Please don't slam the door. The door slammed (shut).*

To confine1.2 is to shut or keep in a small space. *John was confined to bed for a week with his cold.*

To exclude1.3 is to shut out from the mind (a reason or possibility). *We can exclude the possibility that it was the baby who shot the President.*

The seven verb senses defined in terms of “to quarrel”—*caterwaul, jangle, scrap, bicker, squabble, row, bandy*—evoke the same semantic frame. But only some of the seven verb senses defined in terms of “to shut”—*blink, cloister, confine, exclude, slam*, and two senses of *pen*—evoke the same frame.

As noted previously, the terms in the LDOCE defining vocabulary are, on the one hand, highly polysemous, but, on the other hand, tend to be used in only one or two senses. I have assumed that the more verb senses that are defined in terms of a given verb, the more likely that that defining verb is used polysemously across the set of definitions; similarly, I have assumed that light verbs would account for many of the most frequently occurring defining verbs. The assumption underlying the LI.pairs data set—that verbs defined in terms of the same verb have some reasonable possibility of evoking the same semantic frame—is not upheld when a verb is used in multiple senses or when it is used a light verb. Consequently, those verbs defined in terms of a defining

verb used more than 40 times have been eliminated from further processing in this data set.

The LI.pairs data set consists of 41,862 pairs of verb senses defined in terms of the same verb. As illustrated above, some of the relationships are very useful for SemFrame's task, but (many) others are not.

LM.pairs Data Set

The algorithm for generating LM.pairs is given in Figure 6. The LM.pairs data set consists of pairs of LDOCE verb senses that share a common stem, by either the Porter stemmer or the hand-generated list of stemming equivalences discussed in connection with the LC.pairs data set. The assumption underlying this data set is that verbs that are morphologically related are reasonably likely to evoke the same semantic frame.²⁸

Input. D , a set of $(verb_sense_id, verb_stem)$ pairs, where $verb_stem_d$ = the stem for the verb on which $verb_sense_id_d$ is based

Step1. forall m that exist as $verb_stem$ in D , form $DV_m \subset D$, by extracting all $(verb_sense_id, verb_stem)$ pairs where $m = verb_stem$

Step2. forall m that exist as $verb_stem$ in D ,
return all combinations of two members from DV_v

Figure 6. Algorithm for Generating LM.pairs

²⁸The procedure used generates verb sense pairs only when a morphological relationship exists between two distinct verbs. When such a relationship holds, all of the senses of the verbs are paired together. This creates an inconsistency: Senses of polysemous verbs are only paired with each other in LM.pairs if the verb is morphologically related to another verb.

By way of illustration, here are several groups of verbs that share a common stem:

To allow1.1 is to let (somebody) do something; let (something) be done; permit.

They do not allow it / smoking / you to smoke. You are allowed into the room.

To allow1.2 is to make possible (for); provide (for). *This plan allows 20 minutes for dinner. Your gift allows me to buy a car.*

To allow1.3 is to permit to be or to come. *She wont allow dogs in the house.*

To allow1.4 is to give, money or time. *My father allows me some money for books.*

The bank allows 5% on money kept with it.

To allow1.5 is to admit; accept. *We must allow that he is a brave man.*

To allow1.6 is to permit as possible; admit. *The facts allow no other explanation.*

To disallow1.1 is to refuse officially to recognize or allow. *disallow a GOAL | disallow a claim.*

To equal1.1 is (of sizes or numbers) to be the same (as). *x = y means that x equals y.*

To equal1.2 is to be as good, clever, etc. (as). *None of us can equal her, either in beauty or as a dancer.*

To equalize1.1 is to make equal in size or numbers. *to equalize incomes.*

To equate1.1 is to consider or make (2 or more things or people) equal. *You cant equate his poems and / with his plays.*

To liquefy1.1 is to (cause to) become liquid. *Butter liquefies in heat.*

To liquidize1.1 is to crush (fruit or vegetables) into a liquid-like form or juice.

To liquidate1.1 is to get rid of; destroy.

To liquidate1.2 is to kill.

To liquidate1.3 is to arrange the end of business for (a company), when it has too many debts.

To liquidate1.4 is (of a company) to bring business to an end in this way, becoming BANKRUPT.

To liquidate1.5 is to pay (a debt).

To liberalize1.1 is to make or become LIBERAL.

To liberate1.1 is to set free (from control, prison, anxiety, duty, etc.).

To liberate1.2 is to cause or allow (gas) to escape from a chemical substance.

The assumption generally holds for such verb pairs as *allow, disallow; equal, equate;* and *liquefy, liquidize*. However, the assumption does not hold for such pairs as *liquefy, liquidate; liquidize, liquidate;* and *liberalize, liberate*.

The LM.pairs data set consists of 4,717 pairs of senses of verbs with common stems. As in other cases, some of the pairs of verb senses in LM.pairs do evoke a common frame, but many others do not.

LS.pairs Data Set

The algorithm for generating LS.pairs is given in Figure 7. The LS.pairs data set consists of pairs of LDOCE verb senses linked in either of two ways. The first type of link is the strong relation link discussed in connection with the LC.pairs data set. Thus, every verb in a strong relation set is paired with every other verb in the same set. The second type of link is based on a systematic semantic relationship given in LDOCE (Krovetz 1992). These semantic relationship links are, in turn, of two different types: On the one hand, a verb sense may be explicitly linked to another verb or verb sense, as in a COMPARE or OPPOSITE note. On the other hand, a definition for one verb may consist simply of a synonymous verb.

The following entries from LDOCE illustrate explicit semantic relationships in LDOCE:

To *doff*1.1 is to take off (clothes, outer garments, and hats). — opposite *don*
To *don*1.1 is to put on (clothing and hats). — opposite *doff*

To *dower*1.1 is to provide with a DOWER. — compare *endow*
To *endow*1.1 is to give (as to a school) a large amount of money which brings in a yearly amount for use. He spent all his large fortune on endowing a hospital.

To *hypnotize*1.1 is to produce HYPNOSIS in (someone).
To *mesmerize*1.2 is HYPNOTIZE.

Input. D , a set of $(verb_sense_id, def)$ pairs, where def_d = the definition for $verb_sense_id_d$

Step1. forall $d \in D$, if def_d contains *compare* or *opposite* note, extract *related_verb* from note; generate $(verb_sense_id_a, related_verb_a)$ pair

Step2. forall $d \in D$, if def_d defines $verb_sense_id_d$ in terms of a related standalone verb (in BLOCK CAPS) extract *related_verb* from definition; generate $(verb_sense_id_a, related_verb_a)$ pair

Step3. forall $(verb_sense_id_a, related_verb_a)$ pairs, if there is only one sense of *related_verb_a*, choose it and return $(verb_sense_id_a, related_verb_sense_id_a)$, else apply generalized mapping algorithm to return $(verb_sense_id_a, related_verb_sense_id_a)$ pairs where overlap occurs in the glosses of $verb_sense_id_d$ and $related_verb_sense_id_d$

Figure 7. Algorithm for Generating LS.pairs

In these examples, the entries for *doff* and *don* each give the other as its OPPOSITE. Similarly, the entry for *dower* includes a COMPARE reference to *endow*. Further, one sense of *mesmerize* is defined as *hypnotize*.

A semantic relationship linkage is always made from, i.e., in the context of, a specific, linking verb sense. It may or may not specifically indicate which sense of the linked-to verb is related. Where a specific sense of the linked-to verb is indicated or if there is only one sense of the linked-to verb, the semantic relationship can be established with assurance. Where the specific sense of the linked-to verb is not indicated and the linked-to verb has multiple senses, links are established with any sense of the linked-to verb with at least one not-frequently-occurring word in its definition that matches a word in the definition of the linking verb sense. Such links are also established where there is a match between words related to words in the definitions, where such relationships are generated from the strong sense relationships discussed for the LC.pairs data set or are based on the satellite adjective relationships given in WordNet. This disambiguation

procedure mirrors the approach used in the LDOCE-to-WordNet mapping discussed in Section 4.2.

The LS.pairs data set consists of 594 pairs. The precision of this data set is quite high.

LT.pairs Data Set; LZ.pairs Data Set

The algorithm for generating LT.pairs and LZ.pairs is given in Figure 8. The LT.pairs and LZ.pairs data sets consist of pairs of LDOCE verb senses that share a subject field code. The LZ.pairs are limited to those verb senses that share specific subject field codes, while LT.pairs may share either general or specific codes. General subject field codes consist of two characters, while specific subject field codes consist of four characters, of which the first two characters are a general subject field code, the third character is a 'Z', and the fourth character restricts the general subject field code to a more specific category. Four positions are set aside for subject field codes, so that a verb sense may have assigned to it either one specific subject field code, one or two general subject field codes, or no subject field codes at all. These codes may refer to subjects as general as science and business or as specific as billiards and Latin.

The verb sense pairs below are representative of subject relationships established from both general (two-character, LT.pairs) subject field codes and specific (four-character, LZ.pairs) subject field codes:

<p>Input. D, a set of $(verb_sense_id, subject_code)$ pairs, where $subject_code_d$ = any 2- or 4-character subject field code assigned to $verb_sense_id$ (LZ.pairs only uses 4-character subject field codes)</p> <p>Step1. forall c that exist as $subject_code$ in D, form $DV_c \subset D$, by extracting all $(verb_sense_id, subject_code)$ pairs where $c = subject_code$</p> <p>Step2. forall c that exist as $subject_code$ in D, return all combinations of two members from DV_v</p>
--

Figure 8. Algorithm for Generating LT.pairs and LZ.pairs

From LT.pairs:

To blacklist1.1 is to put on a BLACKLIST and to avoid, not give help or work to, not trade with, etc.. *blacklisted for non-payment of debts / for political reasons.*

To cash1.1 is to exchange (a cheque or other order to pay) for CASH. *Can you cash this postal order for that old lady please? Where can I get this cashed?.*

To prosecute1.2 is (of a lawyer) to represent in court the person who is bringing a criminal charge against someone.

To seal1.1 is to make or fix a SEAL onto. *an official statement signed and sealed.*

From LZ.pairs:

To pod1.2 is (of bean plants, PEA plants, etc.) to produce PODs.

To root1.1 is to (cause to) form roots. *Try to root this plant in the garden. Do roses root easily?.*

To oink1.1 is (to make) the sound that a pig makes.

To quack1.1 is to make the sound that ducks make.

To extradite1.1 is to send (someone who may be guilty of a crime and who has escaped to another country) back for trial. *The English murderer was caught by the French police and extradited to Britain.*

To nick1.3 is ARREST. *The police nicked him before he'd gone far on the stolen bicycle.*

The inclusion of specific subject field codes in both the LT.pairs and LZ.pairs data sets reflects the high precision associated with them: With rare exception, two verb senses assigned the same specific subject field code evoke the same semantic frame. Thus the precision of the LZ.pairs data set is fairly high. But many of the verb senses

assigned the same general subject field code are only very indirectly related to each other, as can be seen with the examples taken from the LT.pairs data set.

The LZ.pairs data set consists of 8,440 verb sense pairs, while the LT.pairs data set consists of 189,343 verb sense pairs.

4.1.2 WordNet-based Data Sets

WC.pairs Data Set

The WC.pairs data set relates WordNet verb synsets through the same basic clustering technique outlined for the LC.pairs data set. The algorithm for generating WC.pairs is given in Figure 9. Data for 13,214 WordNet verb synsets were extracted²⁹ and pre-processed in much the same way as the LDOCE data, including the elimination of stop words from synset glosses, the stemming of remaining words, and the weighting of those stems. However, phrasal verbs were not eliminated, since synsets may include both phrasal and non-phrasal verbs. Nor was any attempt made to normalize the defining vocabulary of the glosses, as was done by applying equivalence sets based on strong relations to the LDOCE data. The clustering process used for the two data sets was the same, including the same threshold.

The results of clustering WordNet verb synsets can be seen in the following group of chess-related verbs; only the two senses of *promote* are related through WordNet's semantic network:

²⁹Since the verb-sense-to-synset ratio is 1.8 in WordNet 1.7.1, the number of WordNet verb senses analyzed is larger than the number of LDOCE verb senses. However, since the clustering is executed over definitions/glosses and since WordNet glosses are given at the synset level, the LDOCE and WordNet clustering steps are of equivalent magnitude.

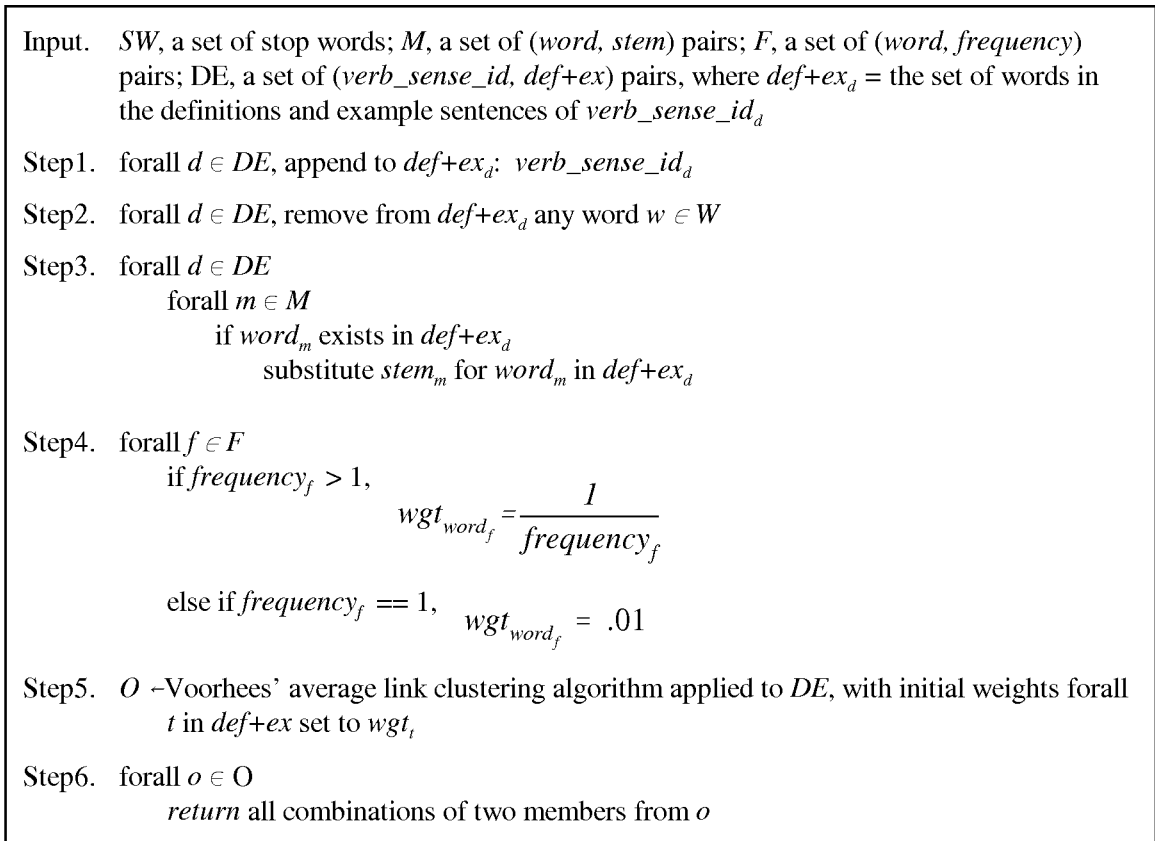


Figure 9. Algorithm for Generating WC.pairs

castle | move the king two squares toward a rook and in the same move the rook to the square next past the king; in chess
 checkmate, mate | place an opponents king under an attack from which it cannot escape and thus ending the game; in a game of chess; "Kasparov checkmated his opponent after only a few moves"
 promote | board games: change a chess pawn for a king by advancing it to the eighth row, or change a checker piece for a more valuable piece by moving it the row closest to your opponent
 promote | board games: be changed for a superior chess or checker piece

The efficacy of WordNet clustering can also be seen in its ability to discriminate, more or less correctly, among the following four computer-related clusters:

write | record data on a computer; "boot-up instructions are written on the hard disk"
format, initialize, initialise | divide (a disk) into marked sectors so that it may store data; "Please format this disk before entering data!"

interleave | intersperse the sectors on the concentric magnetic circular patterns written on a computer disk surface to guide the storing and recording of data

cascade | arrange (open windows) on a computer desktop so that they overlap each other, with the title bars visible

close | cause a window or an application to disappear on a computer desktop

initialize, initialise | assign an initial value to a compute program

hack, hack_on | fix a computer program piecemeal until it works; "Im not very good at hacking but Ill give it my best"

debug | locate and correct errors in a computer program code; "debug this program"

download | transfer a file or program from a central computer to a smaller computer or to a computer at a remote location

upload | transfer a file or program to a central computer from a smaller computer or a computer at a remote location

program, programme | write a computer program

compile | use a computer program to translate source code written in a particular programming language into computer-readable machine code that can be executed

run | execute a program or process, as on a computer or a machine; "Run the dishwasher"; "run a new program on the Mac"

input | enter (data or a program) into a computer

thrash | move data into and out of core rather than performing useful computation; of overloaded paging systems; "The system is thrashing again!"

port | transfer data from one computer to another via a cable that links connecting ports

ftp | use the File Transfer Protocol to transfer data from one computer to another; "You can ftp these data"

offload | transfer to a peripheral device, of computer data

spool | (computer science) transfer data intended for a peripheral device (usually a printer) into temporary storage³⁰

³⁰Unfortunately, because the machine-readable version of LDOCE dates from 1978 and also because its scope is limited to fairly general concepts, none of the relationships in the four computer-related clusters generate counterparts in the corresponding LDOCE mapped data set.

As with LC.pairs, the WordNet clustering output, which may group any number of verb synsets together, is manipulated to establish a direct relationship between every two synsets in a common cluster. The WordNet-LDOCE verb sense mapping (see below) is then applied to convert pairs of WordNet verb synsets into pairs of LDOCE verb senses. The resulting data set consists of 2,772 pairs of LDOCE verb senses corresponding to WordNet verb synsets that clustered together within the same group.³¹

WN.pairs Data Set

Building on the synonymy of words within a synset, the WN.pairs data set consists of all pairs of LDOCE verb senses mapped to the same WordNet synset in the process outlined in Section 4.2. The algorithm for generating WN.pairs is given in Figure 10.

For example, WordNet synset 00397696 includes senses of *impregnate*, *infuse*, *instill*, and *tincture*, for which the gloss and example sentence “fill, as with a certain quality; *The heavy traffic tinctures the air with carbon monoxide*” are given. The following senses of the verbs are given in LDOCE:

To impregnate1.1 is to make PREGNANT.

To impregnate1.2 is to cause a substance to enter and spread completely through (another substance).

To impregnate1.3 is (of a substance) to enter and spread completely through (another substance).

³¹Because the clustering algorithm used produces a hierarchical output, a given synset will appear in at most one cluster. The hierarchical structure could be used to group more specific clusters into more general clusters, but the data have not yet been analyzed in this way.

<p>Input. mapping of LDOCE verb senses to WordNet synsets</p> <p>Step1. forall lines in input file return all combinations of two LDOCE verbs senses mapped to same WordNet synset</p>

Figure 10. Algorithm for Generating WN.pairs

To infuse1.1 is (of a substance such as tea) to stay in hot water so as to make a liquid of a certain taste. To infuse1.1 is to cause (a substance such as tea) to do this.

To infuse1.2 is to fill (someone) with (a quality). *His speech infused the men with eagerness. He infused eagerness into the men.*

To instill1.1 is to put (ideas, feelings, etc.) into someone's mind by a continuing effort. *I instilled the need for good manners into all my children.*

To tincture1.1 is to give a slight colour or taste to (something).

The mapping between LDOCE verb senses and WordNet verb synsets relates WordNet synset 00397696 to the following LDOCE verb senses: impregnate1.2, impregnate1.3, and infuse1.2, all of which are correct. But the mapping misses tincture1.1, whose meaning clearly corresponds to the use of *tincture* in WordNet's example sentence. Absent a deeper analysis than is undertaken in the mapping process, the relationship between 'fill, as with a certain quality' and 'give a slight colour or taste to' goes undetected. The mapping also misses instill1.1, which is metaphorically related. In this case, the spelling difference between the British *instil* and the American *instill* precludes such a mapping even being considered. Further details are given below in Section 4.2.

There are 8,061 pairs of LDOCE verb senses in the WN.pairs data set.

WR.pairs Data Set; WX.pairs Data Set

The WR.pairs and WX.pairs data sets are based on semantic relationships between verb senses encoded within WordNet. The algorithm for generating WR.pairs and WX.pairs is given in Figure 11. The WR.pairs data set consists of pairs of verb senses directly linked through the hyperonymy/hyponymy, antonymy, entailment, and cause-to relationships within WordNet. The direct linkage criterion limits the relationships to a single level. The WX.pairs data sets extends the hyperonymy/hyponymy relationship to all levels: Within this data set, all WordNet verb senses subsumed by a common unique beginner are paired together. As with the WC.pairs data set, the WordNet verb sense pairs are converted to LDOCE verb sense pairs using the WordNet-LDOCE verb sense mapping (see below).

By way of illustration, after conversion to LDOCE verb senses, the first verb sense of *abandon* is related to ten verb senses corresponding to entries in the WR.pairs data set:

To abandon1.1 is to leave completely and forever; desert. *The sailors abandoned the burning ship.*

To chuck1.1 is to throw (something), with a short movement of the arms. *Chuck me the ball. Let's chuck all these old papers away! Don't be so noisy, or the driver will chuck us off (the bus).*

To discard1.2 is (in card games) to give up (unwanted cards). *to discard the Queen of Hearts.*

Input. WordNet data file for verb synsets

Step1. forall synset lines in input file

return (synset, related_synset) pairs for all synsets directly related through
hyponymy, antonymy, entailment, or cause_to relationships in WordNet
(for WX.pairs, also *return* (synset, related_synset) pairs for all synsets within hyponymy
tree, i.e., no matter how many levels removed)

Figure 11. Algorithm for Generating WR.pairs and WX.pairs

To dispose1.1 is to put in place; set in readiness. disposing soldiers for the battle.
 To expose1.1 is to uncover, so as to leave without protection (from something). *to expose ones skin to the sun The soldiers were warned to remain hidden and not to expose themselves. Her youth and beauty will expose her to many dangers.*
 To fling1.1 is to throw violently or with force. *Don't fling your clothes on the floor, hang them up. Every morning he flings the windows open and breathes deeply. She flung her shoe at the cat.*
 To forgo1.1 is to give up; (be willing) not to have (something pleasant). *I shall be happy to forgo (the pleasure of) his company.*
 To leave1.6 is to allow to remain undone, perhaps until a later time. *Let's leave that for another day.*
 To maroon1.1 is to put (someone) off a ship in a place where no one lives.
 To toss1.1 is to throw. *The children tossed the ball to each other. I tossed him the ball / a catch.*
 To waive1.1 is to give up willingly (a right, a rule, etc.). *We cannot waive this rule except in case of illness.*

There are 22 additional verb sense pairs generated from the WX.pairs data set, among which are the following:

To decamp1.1 is (of soldiers) to leave a place where one has camped.
 To defect1.1 is to desert a political party, group, or movement, in order to join an opposing one.
 To delude1.1 is to mislead the mind or judgment of; deceive; trick. *He deluded everyone into following him. Don't delude yourself with false hopes.*
 To depart1.1 is to leave (a place). *The royal train departed from the capital at 12 o'clock.*
 To jilt1.1 is to refuse to see (a lover) any more; unexpectedly refuse to marry (someone) after having promised to do so.
 To linger1.3 is to be slow to disappear. *The pain lingered on for weeks.*
 To skedaddle1.1 is to run away; hurry off.
 To unlearn1.1 is to forget on purpose (something learnt, such as a fact or habit).

Some of the verb senses added using extended hyponymy relationships are altogether on target, while others are less so.

After the conversion process, there are 16,952 pairs of LDOCE verb senses that correspond to the WR.pairs data set and 17,544 pairs of LDOCE verb senses that correspond to the WX.pairs data set.

4.2 Mapping WordNet Verb Senses to LDOCE Verb Senses

Three of the data sets in which WordNet verb senses are paired together (WC.pairs, WR.pairs, and WX.pairs) undergo one more post-processing step to convert them to LDOCE verb sense pairs. This conversion uses a mapping between WordNet verb senses and LDOCE verb senses, generated on the basis of data gathered and processed from each of the two sources.

Available data for each WordNet synset include the words in the synset and the words in the gloss, in both stemmed and unstemmed forms. For example, WordNet gives the following (modified) entry:

00713582 v introduce, present, acquaint | cause to come to know personally; “permit me to acquaint you with my son”; “introduce the new neighbors to the community”

The following non-phrasal senses of *introduce*, and *present* are available in LDOCE (*acquaint* is only in LDOCE as a phrasal verb, *acquaint with*):

To introduce1.1 is to make known for the first time to each other or someone else, by telling 2 people each others names. *I introduced them. Let me introduce myself my name is Simpson.*

To introduce1.2 is to bring in for the first time. *They introduced the idea that children could learn to read as babies.*

To introduce1.3 is to produce the first part of (something), to suggest or explain the main part. *The first few notes introduce a new type of music.*

To present1.1 is to give (something) away, at a ceremonial occasion. *Now that the sports competitions are over, Lady de Vere will present the prizes. When Mr. Brown left the firm, he was given a silver teapot the director presented it to him. He presented her with a bunch of flowers. We were presented with a much larger bill than we had expected.*

To present1.2 is to offer or bring (something) to someone’s notice directly; put forward for consideration or acceptance. *This report ought to be presented in greater detail and in clearer language.*

To present1.3 is to offer (in present one’s apologies, present one’s compliments, present one’s respects, etc.).

To present1.4 is to introduce (someone) to someone of higher rank. He had the honour of being presented to the Governor. *May I present Mr. Jobbings.*

To present1.5 is to give a public performance of. To present1.5 is to give the public a chance to see and hear (a (new) singer, actor, etc.). *The theatre company is presenting Eric Williamson as Hamlet next year.* | *presenting, for the very first time on an English stage, Fanny la Motte.*

To present1.6 is to introduce and take part in (a television or radio show).

To present1.7 is to show; offer to the sight. *Although he may be troubled, he always presents a calm smiling face.*

To present1.8 is (of non-material things) to offer; be the cause of. *He's clever at scientific studies; they present no difficulty to him.*

The mapping procedure identifies present1.3, present1.4, and present1.6 as the LDOCE verb senses that correspond to the WordNet synset. This is based on data for each LDOCE verb sense, which include the words used in the definition, in both stemmed and unstemmed forms, and words related to the words in the definitions through strong relation sets or through satellite adjective sets. Present1.3 is erroneously identified as a corresponding sense, because the occurrence of idiomatic phrases based on *present* is not recognized as introducing extra-definitional material. Both present1.4 and present1.6 are correctly identified as corresponding LDOCE verb senses based on the occurrence of *introduce* in their definitions.

The mapping between WordNet verb synsets and LDOCE verb senses relies on finding matches between the data available for the various verb senses in each resource. Each mapping instance is first assigned a basic similarity coefficient value that reflects the mean proportion of data units (e.g., stems/words from the definition/gloss) in each resource that are matched by data from the other resource:

$$Green\ sim = \frac{C}{2A} + \frac{C}{2B}$$

This measure is most like Dice's coefficient, but differs from it in downplaying differences in the number of data units in the two comparands. For instance, if a WordNet gloss consists of a single word that is matched in an LDOCE definition, the WordNet gloss will contribute .5 to the overall value (the maximum that one resource can contribute); the LDOCE synset would then contribute some additional value to the measure, depending on how many words are in its definition.

This basic measure is then modified to take additional features into account. For instance, matches on words *related to* words appearing in the definition/gloss count as half a match and add to the measure, but the number of related words present does not enter anywhere into the similarity coefficient computation. This measure is reduced when a verb has more than the average number of senses and when the words that match occur frequently. (Note: The measure is not increased when a verb has fewer than the average number of senses and when the words that match occur infrequently.)

This mapping strategy is fairly loose in that it does not consider the subcategorization pattern associated with the verb senses, which in some cases is all that distinguishes two senses of a verb. But for the purpose of establishing a mapping along semantic frame lines, the looseness of the mapping works reasonably well, at least with respect to precision. Examination of all the senses of some number of verbs in both WordNet and LDOCE has determined, however, that recall is lower than precision. Upon reflection, this is not surprising: First, given the paraphrase phenomenon, there is no reason why LDOCE and WordNet would have to use the same words in defining or exemplifying the same or closely related senses of a verb. Second, some recall failures arise where a verb

sense recognized by WordNet does not appear at all in LDOCE. Because of this, it is not a reasonable strategy, when only one sense of a verb appears in LDOCE, to assume that it matches some sense of the verb in WordNet, even if there is only one sense of the verb in WordNet.

The WC.pairs, WR.pairs, and WX.pairs data sets are transformed through this mapping from pairs of WordNet synsets to pairs of LDOCE verb senses.

4.3 Verb Pair Weights

During the generation of the data sets, an initial weight is associated with each LDOCE verb sense or WordNet verb synset pair. In the case of the LC.pairs and WC.pairs data sets, the initial weight is the value of the similarity coefficient for the cluster from which the pair is derived. In the case of members of the LS.pairs data set that are based on semantic links from a specific verb sense to a specific verb sense, the initial weight is 1.0. The initial weight for verb synset pairings underlying the WN.pairs is based on the similarity coefficient introduced in Section 4.2. Weights for WR.pairs and WX.pairs entries are set to 1.0 for antonymy and causal relationships, to 0.7 for hyponymy and entailment relationships. In the case of the LI.pairs, LM.pairs, LT.pairs, and LZ.pairs data sets, and also in the case of the remaining members of the LS.pairs data set, the initial weight reflects the size of the group from which the pair is derived, such that the larger the group, the smaller the weight.

When WordNet synsets are mapped to LDOCE verb senses, the weight assigned to the corresponding LDOCE verb sense pair(s) is the product of three weights: The first is

the initial weight, as described above; the second and third weights are the values of the similarity coefficients associated with each of the two WordNet synsets and their corresponding LDOCE verb senses in the WordNet-synset/LDOCE-verb-sense mapping just described.

With the exception of the 1.0 weights given to explicitly linked members of the LS.pairs data set and to WordNet verb synsets underlying the WN.pairs, WR.pairs, and WX.pairs data sets, none of these weights has extrinsic meaning, nor are the weights comparable to one another. However, in each case the higher the value of the weight, the greater the likelihood that the verb senses paired together evoked a common semantic frame.

After the assignments of initial weight values but prior to the combining of the verb sense pairs data sets, the weights across all the data sets are normalized. Examination of each data set has revealed that pairings at the highest assigned weights are essentially perfect. For the top portion of the pairings, so long as all pairings inspected evoke a common semantic frame, the new weight is set to 0.95. Further examination has established a linear relationship between the weights and the percentage of generated pairs that are correct (i.e., that evoke the same frame). Thus new weights for other pairings can be assigned by interpolation, as soon as the percentage of correct pairings at some appreciably lower weight is determined. The weights assigned to all data sets at the time of the upcoming step are thus a rough reflection of the likelihood that the two verb senses within a pair evoke a common frame.

4.4 Combining Verb Pairs to Form Verb Groups

Altogether, ten sources of data are exploited to generate groups of LDOCE verb senses or WordNet verb synsets hypothesized to evoke the same semantic frame; WordNet verb synsets are then mapped to LDOCE verb senses. The combined data set consists of 25,063 distinctive pairs of semantically related LDOCE verb senses,³² accompanied by a count of how many data sources supported each pairing; all of these pairs meet either a minimum cumulative weight threshold (.0.5) or a minimum count threshold (4 of 10). Of these pairs, 11,654 are supported by LDOCE data alone and 10,904 by WordNet data alone; only 2504 pairs are supported by both LDOCE and WordNet. Just over half the pairs are supported by a single data source; just under 50 are supported by 5 or more data sources.

The next step consists of generating fully connected (i.e., complete link) verb sense groups. These groups are comprised of sets of verb senses, each of which is paired with every other verb sense in the group in the combined data set from the previous step. There are 8,092 fully-connected verb sense groups, averaging 3.0 verb senses per group and including 7,098 distinct verb senses. That the number of groups and the number of verb senses in the groups are close in number reflects both the small size of the groups and also the degree of overlap between groups. These sets of verb senses are then supplemented by pairings from the LDOCE semantic relationships data source

³²All of the pairs in the data sets are ordered orthographically so that if multiple data sets associate the same LDOCE verb senses, they will be easily recognized, as the two verb senses will be in the same order in all data sets.

(LS.pairs), an almost perfect set of LDOCE verb sense pairings. In this process, if one verb sense of an LDOCE subject field code pair appears in a verb sense group but the other does not, the second verb sense is added to the group.

The final step in producing frame semantic groupings of verb senses involves subjecting the verb sense groups to the same clustering algorithm used for generating LDOCE verb sense clusters and WordNet verb synset clusters (pre-processing does not apply in these circumstances). In this context, verb sense groups are clustered together on the basis of the specific verb senses occurring within them. Verb sense groups whose similarity values exceed an arbitrary threshold—the use of four thresholds (0.5, 1.0, 1.5, and 2.0) have been explored—are then merged. In general, the lower the threshold, the looser the verb grouping and the fewer the number of clusters produced. The number of verb senses retained and the verb sense groups produced by using these thresholds are recorded in Table 3. As can be seen, the result of the process includes somewhat less than half of all the LDOCE verb senses.

To illustrate this procedure, consider senses of verbs having to do with golf. At least 16 of these exist in LDOCE, as can be seen in Table 4. One or more of these verb senses

Threshold	Number of verb senses occurring in at least one verb sense group	Number of verb sense groups produced
0.5	6461	1338
1.0	6414	1759
1.5	5607	1421
2.0	5604	1563

Table 3. Results of Frame Clustering Process

To bogey1.1 is (in GOLF) to hit the ball into (a hole), taking one stroke more than is average.
To golf1.1 is to play golf. to go golfing.
To hit1.1 is to give a blow to; strike. He hit the other man. He hit the ball (with the BAT).
To hook1.4 is (of a ball) to travel in a HOOK. To hook1.4 is to hit (a ball) in a HOOK.
To loft1.1 is (in cricket and golf) to hit (a ball) high.
To mishit1.1 is (in cricket, GOLF, etc.) to hit (the ball) in a faulty way.
To par1.1 is (in the game of GOLF) to play the number of strokes for (a hole or all the holes) which is equal to PAR.
To pitch1.3 is (of a ball in cricket or GOLF) to hit the ground.
To pull1.11 is (in GOLF) to strike (the ball) to the left of the intended direction (or to the right if one is left-handed).
To putt1.1 is to strike (the ball) gently along the ground towards or into the hole.
To putt1.2 is to take a stated number of PUTTs to hit the ball into (the hole). Jacklin3-putted the 17th hole.
To shoot1.12 is (in GOLF) to make (the stated number of strokes) in playing a complete game. Miller shot a 69 today.
To sink1.10 is (in games like GOLF and BILLIARDS) to cause (a ball) to go into a hole.
To slice1.4 is to hit (a ball) in a SLICE.
To stymie1.1 is (in GOLF) cause (someone or oneself) to be stopped by positioning the balls in a STYMIE.
To top1.5 is (in GOLF) to hit (a ball) above the centre. He topped the ball and it went all along the ground.

Table 4. Glosses and Example Sentences in LDOCE for Verb Senses Related to Golf

appears in 56 different fully connected verb sense groups (after supplementation by pairings from the LDOCE semantic relationships data source). As many as 8 of the verb senses belong to a fully connected verb sense group, but many of the groups have only 2 members; the average membership size is 2.7. Clustering and merging reduces the number of golf-related clusters from 56 to 7. One of these clusters contains 13 of the 16 relevant verb senses (all but hit1.1, hook1.4, and putt1.1).

4.5 Summary

This chapter introduces ten sources of data, drawn from LDOCE and WordNet, that are used in identifying verb sense framesets, i.e., sets of verb senses that evoke a

common frame. The data sources include terms that occur in LDOCE definitions and WordNet glosses, various semantic relationships coded within LDOCE and WordNet, LDOCE subject field codes, morphological relationships between verbs, and the infinitives that LDOCE verb senses are defined in terms of. Data sets based on these sources identify from hundreds to tens of thousands of LDOCE verb sense pairs or WordNet verb synset pairs that might evoke the same frame. After initial extraction, WordNet synset pairs are mapped to LDOCE verb sense pairs through a general frame-oriented disambiguation procedure. Data pairs from all ten sources are then combined and filtered through a cumulative-weight-or-support threshold.

Ever more comprehensive groups of verb senses are generated by: (1) combining verb sense pairs into fully-connected sets of verb senses; (2) supplementing those sets with verb sense connections from the most trusted of the original data sets, and (3) subjecting the verb sense groups to a clustering procedure that merges groups with significant overlap. Roughly 1500 verb sense framesets were produced for each of four clustering thresholds.

Chapter 5

Identifying Semantic Frames by Participant Structure: Task 2

The final result from task 1 is 4 sets (one for each threshold) of weighted LDOCE verb sense groups. Ideally, these sets of verb senses evoke/inherit the same frame. Task 2 processing seeks to identify the internal participant structure of each verb sense group output in task 1. As no interaction occurs between the verb sense groups in task 2, the question is entirely that of how to transform a set of verb senses into a frame structure, including a frame name and a set of frame slots.

In task 1 a simplifying procedure is adopted, namely, to identify the membership of a frame by enumerating the set of verb senses that evoke it, although clearly other parts of speech also evoke frames. In task 2 a similar simplifying procedure is adopted, which is to identify the participant structure of the frame in terms of nouns closely associated with its set of verb senses. The associations exploited include (1) nouns within the LDOCE definition of a verb sense, (2) nouns in example sentences within the LDOCE entry for the verb sense, and (3) nouns that are morphologically related to the verb senses. Thus the first subtask in task 2 is to expand the extensionally-based identification of semantic frames, heretofore accomplished by enumerating a set of verb senses that evoke a common frame, by supplementing that list with a list of noun senses that evoke the same frame.

For example, take the cluster of golf verb senses represented in Table 4 (but eliminating hit1.1, hook1.1, and putt1.1). Minipar's (Lin 2001) parsing of the definitions identifies the following nouns (with their number of occurrences):

1 average	2 cricket	1 ground	2 right
10 ball	3 direction	6 hole	6 slice
1 billiards	1 equal	3 number	2 stroke
1 cause	2 game	2 par	2 stymie
1 centre	3 golf	1 putt	1 way

The following nouns are extracted from example sentences:

1 ball	1 ground	1 miller	1 putted
--------	----------	----------	----------

The following nouns are morphologically related to the verbs around which the cluster is built:

1 bogey	1 loft	1 pitch_shot	1 putt
1 golf	2 par	1 pitcher	1 stymie
1 golfing	1 pitch		

Additionally, on the basis of semantic typing done in Minipar, the presence of 1 general pronoun, 2 numbers, and 4 personal pronouns are identified as being associated with this group of verb senses. Finally, each of these nouns is mapped to a WordNet noun synset, based on the same kind of mapping process described in Section 4.2. Further explanation of the identification of nouns corresponding to verb sense clusters is given in Section 5.1. The list of nouns, their cumulative number of occurrences, corresponding WordNet synsets, and a mapping weight become input to the conceptual density program, which is explained in Section 5.2.

The overall idea of transforming such a list of nouns into a list of participants involves using the relationship structure of WordNet to identify an appropriately small set of concepts (i.e., synsets) within WordNet that account for (i.e., are superordinate to) as many of the nouns as possible; such synsets will be referred to as *covering synsets*. If the only constraint were to account for as many of the frame-associated nouns as

possible, the clear solution would be to pick out WordNet nodes at the top of their respective hierarchies, since the number of nouns covered by a synset is greater the higher the covering synset is in the WordNet hierarchy. However, nodes at the highest levels of the WordNet tree are more general and abstract than frame slots typically are. We want to identify synsets that characterize the participant structure of the frame as closely and accurately as possible. At the same time, the number of participants in a frame is generally small, perhaps only two, rarely more than four or five. This observation motivates the desire to constrain the number of covering synsets identified. It is to be taken for granted that some nouns on the lists will not be accounted for by the covering synsets.

The task of identifying the participant structure of the frame evoked by sets of verb senses relies on the hypothesis that the nouns associated with them will not be randomly distributed across WordNet, but will be clustered in various subtrees within the hierarchy. In essence, the task is to identify those clusters/subtrees and then to designate the nodes at the roots of the subtrees as covering synsets (subject to the aforementioned constraints).

It will be demonstrated that a conceptual density analysis of the WordNet data can lead to the identification of both an appropriate name for a semantic frame and a viable set of WordNet nodes/synsets that correspond to elements of the frame.

The overall algorithm for task 2 is illustrated in Figure 12.

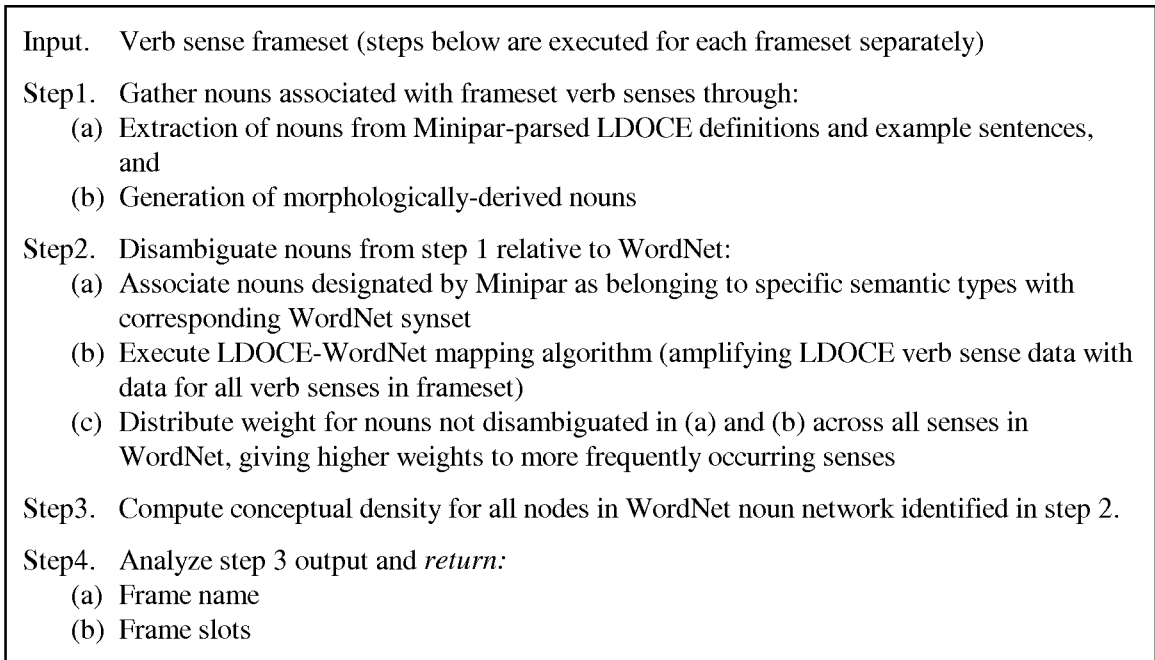


Figure 12. Algorithm for Task 2

5.1 Enumerating Noun Senses that Evoke Semantic Frames

Task 1 outputs sets of verb senses that evoke a common semantic frame. The corresponding sets of noun senses—that is, noun senses that evoke those same frames—can be identified through several different means. The set of nouns appearing in the LDOCE definitions or example sentences of the frame-evoking verb senses constitute one candidate set. Nouns that are morphologically related to the frame-evoking verb senses constitute another candidate set. Specific senses of the nouns are then identified, using a variant of the lexical entry mapping algorithm.

5.1.1 Nouns Drawn from LDOCE Entries

Syntactic arguments of the verb(s) used in a definition often correspond to semantic (and usually syntactic) arguments of the verb being defined. For example, Table 2 (repeated here for the sake of convenience) gives the definitions of several verb senses that evoke the COMMERCIAL TRANSACTION frame, which includes as its semantic arguments a Buyer, a Seller, Merchandise, and Money. Words corresponding to the Merchandise (*property, goods*), the Money (*money, value*), and the Buyer (*buyer, buyers*) are present in the definitions; however, no words corresponding to the Seller are present.

Verb sense	LDOCE Definition
buy1.1	to obtain (something) by giving money (or something else of value)
buy1.2	to obtain in exchange for something, often something of great value
buy1.3	to be exchangeable for
purchase1.1	to gain (something) at the cost of effort, suffering, or loss of something of value
sell1.1	to give up (property or goods) to another for money or other value
sell1.2	to offer (goods) for sale
sell1.3	to be bought; get a buyer or buyers; gain a sale

Table 2. Definitions for Verbs Evoking the COMMERCIAL TRANSACTION Frame

In order to isolate nouns in the definition of LDOCE verbs that might correspond to a semantic argument of the verb, LDOCE definitions for verbs have been systematically manipulated to cast them as full sentences.^{33, 34} These sentential definitions have been

³³One might ask why WordNet glosses were not similarly analyzed. The first answer is that the use of a restricted vocabulary in LDOCE made the use of words in its definitions more likely to be useful for the task. No particular attempt has been made in WordNet to restrict the words used in glosses. Moreover, the WordNet glosses have been designed not

processed by the Minipar (Lin 2001) parser, which returns dependency tree output. The Minipar output has been analyzed to identify nouns associated with each LDOCE verb sense. Where Minipar identified a noun as belonging to a particular semantic type (e.g., time, money, number, person, location), this information has been retained.³⁵

Various distinctions evident in the parser output are considered in assigning weights to the nouns. These weights are intended to reflect the likelihood that the nouns would correspond to a semantic argument of the verb. The first such distinction is whether a noun appears in the definition of the verb sense or appears in a sentence exemplifying the verb sense; nouns appearing in definitions are more likely to reflect the semantic type of a frame element than are nouns appearing in example sentences. A second distinction concerns the grammatical function played by the noun: Is it the subject (in an example sentence)? Is it the object of a verb or preposition (in an example sentence or a definition)? Nouns that cannot be determined to play one of these grammatical functions are less likely to reflect the semantic type of a frame element than those that do. A third distinction involves whether the noun appears parenthetically in the definition. Nouns

so much as definitions *per se*, but as indications for distinguishing between senses. Still, it is possible that an analysis of nouns used in WordNet glosses might make a positive contribution to the subtask.

³⁴For the most part this manipulation consisted of prepending the phrase “To *verb-sense* is to” to each of the definitions given within an entry; some additional editing was also undertaken. Thus, for example, the definition for the one sense of *abase*—‘to make (someone, esp. oneself) have less self-respect; make humble’—was recast as: ‘To abase1.1 is to make (someone, oneself) have less self-respect. To abase1.1 is to make humble.’

³⁵For example, dollar amounts (e.g., “\$5”) are designated as MONEY, cardinal numbers (e.g., “two”) as NUM, periods of the day (e.g., “night”) as TIME.

included parenthetically in a definition may (but do not always) reflect the prototypical semantic type of the subject or object of a verb. To determine how significant a role these distinctions should play in assigning weights, approximately 100 definitions were analyzed and an informal estimate of the correspondence between these variables and the usefulness of the noun as a reflection of a semantic argument of the verb noted. A default value of 1.0 was associated with each distinction; this value was increased to reflect the degree of usefulness of the distinction. An initial weight that is the product of the corresponding values is assigned to each verb sense/noun pairing.

In addition to making use of distinctions evident in the parser output, the weight assigned to nouns also made use of a *tf-idf*-like measure, where *tf* corresponds to the number of times a noun was associated with a verb sense in the LDOCE definition and example sentence data and where *idf* corresponds to the number of verb senses with which any given noun was so associated. On the one hand, if a noun occurs multiple times in a verb sense's definition or occurs in both its definition and in a sentence exemplifying its use, the noun is more likely to correspond to a semantic argument. On the other hand, if a noun occurs in the definition or example sentence for many verb senses, it is less likely to correspond to a semantic argument for a specific verb sense. Correspondingly, a noun that corresponds closely to a semantic argument of a specific verb is not so likely to occur with many other verbs. Thus the weight assigned to a verb sense/noun pairing takes into account both the parser distinctions outlined above and the *tf-idf* ratio.

Many of the nouns gathered through the parsing operation and being associated with specific verb senses are polysemous. Clearly when a noun reflects the semantic type of an argument of an associated verb sense, it is a specific sense (or set of senses) of the noun that is referred to. A preliminary disambiguation of these noun senses has been carried out by applying a mapping process similar to the procedure explained in Section 4.2. Further disambiguation steps will be set forth in Section 5.1.3.

5.1.2 Nouns Generated by Morphological Analysis

While arguments appearing in a verb's definition will often correspond to semantic arguments of that verb, not all semantic arguments of the verb being defined are likely to be reflected in the definition. Indeed, certain arguments are typically only implied. For example, unless supplied parenthetically, indication of the argument that fills the subject slot in active voice will often be missing from a verb's definition. Also missing from the definition of a transitive verb may be explicit mention of the argument filling the direct object slot.³⁶

Missing arguments of verbs may be recovered in part by probing for the existence of nouns generated through suffixation of particular morphemes onto the verb whose

³⁶As noted previously: In at least some cases, a transitive verb whose definition lacks mention of the argument filling the direct object slot will have an intransitive counterpart that mentions the argument. Fellbaum (1998b, p. 86) gives as an example “the transitive superordinate *drink* [, which] takes noun arguments that are hyponyms of *liquid* or *beverage*. The intransitive subordinate verb has the more specific sense ‘to drink alcoholic beverages.’” The definition of the intransitive sense of the verb is thus likely to explicitly contain a noun within the *liquid/beverage* hierarchy.

subject argument is sought. For example, let us take (a subset of) the verbs that evoke the COMMERCIAL TRANSACTION frame: *buy, purchase, sell, pay, cost, spend, price, charge*. Many of the nouns that are morphologically related to these verbs—e.g., *buy, buyer, purchase, seller, payee, payer, payment, cost, spender, spending, price, charge*—are especially good indicators of a semantic argument of this frame.

Generation of a list of nouns morphologically related to LDOCE verb senses has been produced by following three steps. The first step is to generate word forms that might correspond to morphologically related nouns by appending an appropriate set of suffixes to the root of each verb.³⁷ The product of this step is, for each LDOCE verb, a list of morphologically related word forms that might actually exist as nouns. The second step is to filter this list by checking if each generated word form corresponds to a noun in WordNet; if not, it is deleted from further consideration since subsequent analysis of the nouns takes place in WordNet. The product of this step is, for each LDOCE verb, a list of morphologically related word forms that exist as nouns in WordNet.³⁸ The third step is to match specific senses of the LDOCE verbs with specific senses of the WordNet nouns. This process exploits the same general mapping algorithm used for fixing correspondences between LDOCE verb senses and WordNet verb senses, discussed in

³⁷Prior to suffixation, alternative stems were produced by converting final *y*'s into *i*'s and dropping final *e*'s. The following suffixes were appended: *∅, -er, -or, -ee, -ed, -man, -men, -ness, -tion, -ation, -ion, -st, -ence, -ent, -ance, -ant, ship, -al, -ite, -ist, -ast, -ism*. If a verb ended in *-ize* or *-ise*, a further nominalization was generated by substituting *-ist* for that ending.

³⁸This list has been incorporated into CatVar (Habash and Dorr 2003), a large-scale lexical resource for English, which relates words, possibly of different parts of speech, that are morphologically related to each other.

Section 4.2. The product of this step is, for each LDOCE verb sense, a list of morphologically related WordNet noun synsets for which some evidence of correspondence exists in their associated definitions (LDOCE) and glosses (WordNet). As noted before, because this mapping strategy largely ignores the paraphrase phenomenon (except through its inclusion of related words), the precision of its results is higher than its recall. The weight assigned each of these LDOCE verb sense/WordNet noun synset associations incorporates both (1) a constant value for the relatively high usefulness of morphologically related nouns for identifying the semantic arguments of verbs and (2) a similarity coefficient value produced by the mapping process.

5.1.3 Nouns Mapped to WordNet Synsets

Through the processes just described, every verb sense in LDOCE has associated with it a weighted set of nouns, each of which (1) appears in its definition, (2) appears in a sentence exemplifying its use, or (3) is morphologically related to it. Some, but not all, of the nouns are WordNet-sense-disambiguated as a result of these processes. The remainder of the nouns are mapped to WordNet synsets through one of three further processes.

The first of these processes associates nouns that received semantic type designations in the Minipar parsing with specific (i.e., corresponding) WordNet nodes.

The second of these processes relies on the LDOCE-WordNet mapping algorithm, described in Section 4.2, one more time. This execution of the algorithm investigates word and stem matching between (1) data related to LDOCE verb definitions *for all verb*

senses in the frame semantic verb sense group and (2) data related to WordNet noun synset glosses.

The third of these processes completes the word sense disambiguation process by making default assignments for all nouns not associated with WordNet noun synsets through previous analysis. The strategy involved takes into account that WordNet senses are ordered by the frequency of their use in SEMCOR. Thus the first sense given for a word in WordNet has the highest *a priori* possibility of being the correct sense of the word. Specifically, the strategy is to assign half of the noun's weight to the first sense, a fourth of the weight to a second sense, an eighth of the weight to a third sense, and so on. In this manner, the original weight assigned to the noun is distributed across its various senses in WordNet, roughly proportional to the *a priori* likelihood of the sense being a correct assignment.

As a result of these two processes, all nouns associated with a verb sense group are mapped to specific WordNet noun synsets. All these mappings have an associated weight. The various nodes within WordNet's noun network that correspond to a verb sense group constitute *evidence synsets* for the participant structure of the corresponding semantic frame.

5.2 Computing Conceptual Density

After mapping all the nouns associated with a semantic frame to their corresponding WordNet noun synsets, the next step is to analyze the accumulated data from the evidence synsets. It is hypothesized that the WordNet subtrees with the highest

density—where density takes into account the number of evidence synsets present in the subtree, their weight, and their relative location in the subtree—are the most likely to correspond to frame slots. Intuitively, when evidence synsets cluster together, the subtrees in which they occur will be more dense than those subtrees where few or no evidence synsets occur. Figure 13 illustrates this notion of density, with evidence nodes occurring in 4 subtrees. Subtree 3 can be seen intuitively to be the most dense.

Agirre and Rigau (1995) use a similar notion of density for the purpose of word sense disambiguation. The specific application is not of interest here, but the measure they use is. They define conceptual density as the ratio between the expected area of a subtree containing a word sense and some number of context words (“marks”) and the actual area, where area is a function of the height of a node and the number of its descendants. More formally, Agirre and Rigau define a subtree (“subhierarchy”) in terms of its top, subsuming node (c), its height (h), the number of nodes within the hierarchy (used only indirectly, in the computation of mean number of hyponyms per node, $nhyp$), and the number of marks

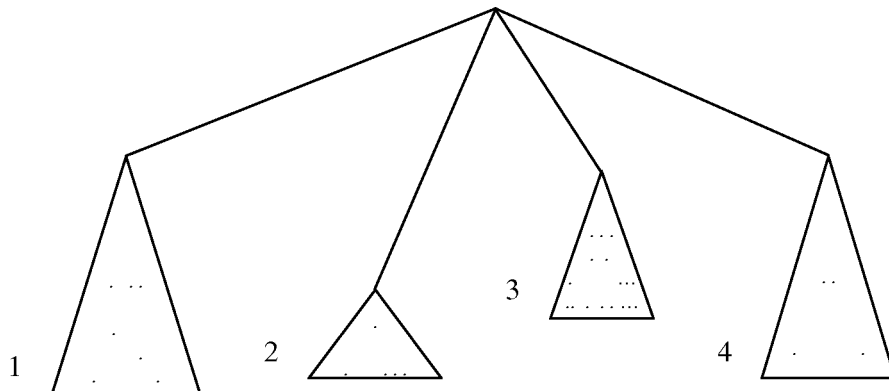


Figure 13. Conceptual Density in Subtrees

(m) in the subtree. They define the conceptual density (CD) of the subtree below c with m marks as:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \text{ where } descendants_c = \sum_{i=0}^{h-1} nhyp^i$$

Instead of using height and mean number of hyponyms per node to compute density, as suggested by Agirre and Rigau, the density measure adopted for use here incorporates the actual size of the subhierarchy below any given WordNet node. In addition, the evidence synsets, which correspond to Agirre and Rigau's marks, are weighted.

The density of a node in the WordNet noun network thus has two components in SemFrame: first, the occurrence of a weighted evidence synset at that node and, second, the occurrence of weighted evidence synsets at descendant nodes. For each node, record is initially made of (1) the cumulative weight associated with all nouns mapped to that node, (2) the number of nodes subordinate to that node within WordNet (its *treeSize*), and (3) the node's area, which is the product of its cumulative weight and its *treeSize*. Subsequently, a node's area is added to the area of all its ancestors. The conceptual density of a node n is computed in SemFrame as the ratio between its (cumulative) area and its (invariant) *treeSize*:³⁹

³⁹That is to say, a node's area is affected by evidence synsets below it, but its *treeSize* does not change on account of any of its descendants being evidence synsets.

$$CD(n) = \frac{\sum_{i \in \text{descendants}_n} (\text{wgt}_i * \text{treesize}_i)}{\text{treesize}_n}$$

Let us look at some of the ramifications of this measure. First, if a node is not an evidence synset and none of its descendant nodes are evidence synsets, the node's density is 0 (since all weights are 0). Second, if an evidence synset has no descendant nodes that are evidence synsets, its density will equal its cumulative weight.⁴⁰ Third, a node with descendant nodes that are evidence synsets will always have a higher density than the same node without descendant nodes that are evidence synsets.⁴¹ Fourth, if a node is not itself an evidence synset, but all its children have density x , the node will also have density x .⁴² It follows that an ancestor node that is an evidence synset, even if its weight is not very high, may have a higher density value than a more heavily weighted descendant node, especially if most or all of its descendants are heavily weighted.

⁴⁰In this case, the treesize that is a multiplicative factor in computing the node's area in the numerator of the density ratio will cancel out the treesize in the denominator.

⁴¹Under these circumstances, the addition of the descendant's area to a node's cumulative area will increase the numerator of the density ratio, while the denominator's value will hold constant.

⁴²Let S = the sum of the treesize of all the node's children. The node's cumulative area is, by the distributive law, the product Sx . Its treesize is S , and therefore its density is x .

5.3 Interpreting Conceptual Density

After all nouns associated with a particular semantic frame have been WordNet synset disambiguated, as described above, the density of each evidence synset and of each of its ancestor synsets is computed. Two thresholds have been set to eliminate WordNet nodes with either (1) high density values but little overall support or (2) low density values. Specifically, to receive further consideration, a node must have a distributed support count of at least 2.⁴³ A node receives support in the same proportion that it is awarded a noun's weight: If a noun is associated with a single WordNet synset, it contributes a count of 1 to that node, but if a noun's weight is distributed across multiple synsets in the final disambiguation process described in 5.2.3, then it contributes a count of .5, .25, .125, etc., according as it is the first sense, the second sense, the third sense, etc. This threshold helps minimize the effect of nouns that are erroneously associated with a semantic frame or that are mismapped to a WordNet synset.

The number of WordNet nodes with non-zero densities for a specific semantic frame depends in large part on the number of verb senses associated with the frame. But the complexity of the frame—that is, the number of slots in its internal participant structure—tends not to vary. It is desirable then to establish a relative threshold for density values rather than an absolute threshold. The threshold used here has been the mean density (of nodes within the processing of a particular semantic frame) plus 1

⁴³The thresholds and other heuristics involved in the interpretation of conceptual density values throughout this section are based on observations from a data set of one dozen frames.

standard deviation. This threshold retains nodes of interest while eliminating many spurious nodes.

Of the nodes that meet these threshold criteria, many are related hierarchically. In other words, there is often considerable conceptual redundancy in the set of nodes that remain. As a further filter, of nodes with a direct hierarchical relationship, only the one with the highest conceptual density is retained.

5.3.1 Identifying Names for Semantic Frames

The WordNet semantic network for nouns is divided into approximately a dozen or so subnetworks, as there is no single, top node dominating all other nodes. Instead WordNet noun network includes a group of unique beginners, nodes that are not subsumed by other nodes. Unique beginner nodes establish the semantic type of all their descendant nodes.

Semantic frames being sought in connection with the paraphrase phenomenon are of certain types. They correspond especially to events and processes, but also to some kinds of abstractions. We are not concerned, for instance, with the kinds of frames that describe concrete entities, although the notion of semantic frame still applies to such. Given the character of the frames sought, only nodes of certain semantic types will correspond to the overall frame.

The set of semantic types appropriate to frames includes abstractions, actions, events, phenomena, psychological features, and states. Accordingly, the node with the highest density value from among these subnetworks is designated as corresponding to the name of the semantic frame. The actual name is chosen from among the nouns in the synset. If

the synset occurs within SEMCOR's tagging, the noun that corresponds to that synset most frequently is designated as the frame name. If the synset does not occur within SEMCOR's tagging, the noun that is listed first in the synset is arbitrarily chosen as the frame name.

5.3.2 Identifying the Internal Participant Structure of Semantic Frames

The nodes that remain are all candidates for correspondence with slots in the internal structure of the frame. However, that set of nodes is often larger than the anticipated number of slots in the typical frame. The organization of the WordNet noun network by semantic types helps further restrict that number. Comparative analysis of sets of nodes has resulted in the insight that, with two exceptions, no more than one slot of a particular semantic type generally occurs in a frame; the exceptions are the entity and abstraction types. Thus, one final filter is implemented such that only the highest density node in any given subnetwork is output, except for the entity and abstraction subnetworks.

One of the major reasons for choosing to work with data drawn from lexical resources for the induction of semantic frames was that nouns associated with a verb sense in dictionary entries—in definitions and even in example sentences—are more likely to be at a hierarchical level corresponding to a frame element or slot than would be the case for nouns drawn from corpus data. Still, the nouns under investigation are not always at an appropriate level, and the data are sparse. Thus the nodes with the highest density are not always at an appropriate hierarchical level. To correct this situation, a set of approximately 40 nodes across the WordNet noun network have been picked out, which correspond most closely to standard frame elements. Nodes with high density scores that

are within any of the subtrees dominated by one of these 40-odd nodes are replaced by the “standard” frame element.

5.4 Results

As shown in Table 3 of Section 4.4, approximately 1500 framesets were output in task 1 for each of four thresholds. Each of these framesets is processed individually in task 2. This processing achieves two results: First, the list of LDOCE verb senses associated with a specific frame is expanded to include corresponding noun senses; both are given in terms of WordNet synsets. Second, the noun framesets are analyzed to produce a frame structure, including both a frame name and a list of frame participants. Several examples of resulting semantic frame structures are presented in Figure 14.⁴⁴

Examples (a)-(c) reveal problems as yet unresolved in task 2. Example (a) shows an incomplete frame structure, which is missing indication of the agent emitting the screech/scream. Example (b) illustrates the results of less than perfect disambiguation: The sense of *Wash* given does not correspond to a laundry-type frame; the appropriate sense of the word is in the same synset with the Laundry slot and could perhaps be weeded out on that basis. Similarly, only one of the two senses of *Washer* is correct. Example (c) has three slots—for Absentee, Deportee, and Emigrant—that correspond to

⁴⁴The examples in Figure 14 cannot be claimed to be representative of the output. A goodly proportion of the frame structures output are less satisfactory than those presented here. In some cases, there is insufficient evidence to output even a frame name. In other cases, it is unclear how the frame and slot names proposed correspond to a recognizable situation type. However, the issues/problems illustrated in these examples are representative of issues/problems that require further addressing and resolution.

- (a) Frame Screech (A-high-pitched-noise-resembling-a-human-cry;-):
Communication (Scream (Sharp-piercing-cry;-)) []
- (b) Frame Wash (The-work-of-cleansing-(Usually-with-soap-and-water)):
Washer (A-home-appliance-for-washing-clothes-and-linens-automatically) []
Laundry (Garments-or-white-goods-that-can-be-cleaned-by-laundrying) []
Wash (A-thin-coat-of-water-base-paint) []
Washer (Seal-consisting-of-a-flat-disk-placed-to-prevent-leakage) []
Washer (Someone-who-washes-things-for-a-living) []
Action []
- (c) Frame Emigration (Migration-from-a-place-(Especially-migration-from-your-native-country-in-order-to-settle-in-another)):
Absentee (One-that-is-absent-or-not-in-residence) []
Deportee (Expelled-from-home-or-country-by-authority) []
Emigrant (Someone-who-leaves-one-country-to-settle-in-another) []
Country (The-territory-occupied-by-a-nation;-) []
Organization []
Act []
- (d) Frame Killing (An-event-that-causes-someone-to-die):
Person1/agent []
Person2/recipient or Patient []
Event []
- (e) Frame Exultation (The-utterance-of-sounds-expressing-great-joy):
Success (A-person-with-a-record-of-successes;-) []
Communication (Exultation (The-utterance-of-sounds-expressing-great-joy)) []
Feeling []
- (f) Frame Emptying (The-act-of-removing-the-contents-of-something):
Empty (A-container-that-has-been-emptied;-) []
Filling (Any-material-that-fills-a-space-or-container;-) []
Relation (Content
(The-proportion-of-a-substance-that-is-contained-in-a-mixture-or-alloy-etc.)) []
Measure (Capacity (The-amount-that-can-be-contained;-)) []
Action []
- (g) Frame Repair (The-act-of-putting-something-in-working-order-again):
Mending (Garments-that-must-be-repaired) []
Mend (Sewing-that-repairs-a-worn-or-torn-place-in-a-garment) []
Patch (A-piece-of-cloth-used-as-decoration-or-to-mend-or-cover-a-hole) []
Location []

Figure 14. Examples of Frame Structures Generated in Task 2

the same semantic argument. Since they are not hierarchically related in WordNet, but are all entities, all are retained in the output. It is not immediately clear how this redundancy could be detected automatically. Also with relation to example (c), it may be noted that what the Organization and Act slots relate to is not intuitively clear.

Examples (d)-(g) represent more successful frame structures. Example (d) detects the presence of two distinct Person slots in a Killing frame (although it is unable to identify the semantic nature of the two slots as Killer and Victim); the presence of an Event slot is easily interpreted as that which was done by the Killer that resulted in the Victim's death. Example (e) nicely captures a relationship between a Person who has achieved a success, the associated Feeling, and a resulting exultant Communication; it thus suggests the embedding of a Communication frame instantiation within the Exultation frame, but has no means of indicating that the successful Person in this frame corresponds to the Speaker in the embedded Communication frame. In example (f), the description of the Relation slot is not altogether on target, but the related idea of a proportion of the container that is full/empty, suggested by the Relation slot, is not without benefit. The problems with example (g) are more subtle: The description given at the frame level is not as specific as the slots are. Moreover, indication is missing of the agent who effects the mending.

Taking the last example (g), we may also examine briefly how frame and slot names are generated. The LDOCE verb senses in the corresponding frameset are: doctor1.2, patch1.1, patch1.2, mend1.1, mend1.2, repair1.1, and fix1.5. Nouns corresponding to these verbs include: *break, coat, elbow, fix, fixing, garment, hole, mend, mender, mending, old, part, patch, patching, radio, repair, road, sewing, shirt, and shoe*.

Conceptual density computations based on these nouns within WordNet identify a handful of prominent noun nodes, including those listed in Figure 14. The frame name, *Repair*, comes from the *Act* subnetwork of WordNet and is the highest scoring node from among the action/event/state-oriented portions of the WordNet noun network. Three of the slots, *Mending*, *Mend*, and *Patch*, come from the *Entity* subnetwork. Frame participants are often of the *Entity* semantic type, and the relevance of all three of these slots supports the decision not to limit the number of *Entity* slots retained. The *Location* slot is generated on the basis of a procedure applied to several semantic types: When the cumulative distributed weight assigned to the *Location* subnetwork (more specifically, to the highest node within that subnetwork) is as much as 2.5% of the overall node count, but no specific *Location*-oriented node emerges as a named slot, the generic semantic type slot is adopted instead.

5.5 Summary

This chapter has described how verb sense framesets—sets of verb senses that have been automatically identified as evoking a common frame—are expanded to include corresponding noun senses. This expansion is accomplished through the extraction of noun phrases from LDOCE definitions and example phrases and the generation of nouns that are morphologically-related to the verbs. This new set of frame-evoking words is mapped to the WordNet noun network. In the context of this mapping, the conceptual density of specific WordNet nodes is computed. This measure is used to identify noun nodes most likely to correspond to slots within the corresponding frame, as well as to

identify the WordNet node that best corresponds to the overall frame. Examples of frame structures generated in this task are presented and several outstanding issues—especially incomplete frames, imperfect disambiguation, and redundancy among slots—revealed.

Chapter 6

Evaluation

The underlying hypothesis driving this research is that, to the extent that semantic frames reflect cognitive organization and processing, their intelligent use should improve performance on computational tasks of a knowledge intensive nature. This expectation is examined and evaluated in the context of a text segmentation task, specifically that of detecting the boundary between texts when they appear consecutively one after the other. Text segmentation on this coarse level would be useful, for example, in isolating the individual stories of an audio news feed. Identification of distinct stories would be required prior to retrieval and analysis of individual texts.

Hearst's (1997) TextTiling algorithm for segmenting text into multi-paragraph units has been adopted as a baseline method for performing boundary detection between consecutive texts.⁴⁵ The TextTiling approach, which is based on term repetition, is supplemented with information about sets of verbs and nouns that evoke common frames. Semantic frame repetition substitutes for term repetition where (1) information is available on which semantic frame(s) the term evokes and (2) evidence for assignment of the term to a single semantic frame is sufficiently strong. The modification permits comparison between a straightforward implementation of the TextTiling approach and a semantic-frame-enhanced implementation of the approach. TextTiling has demonstrated strong

⁴⁵In the standard computational linguistics textbook (Jurafsky and Martin 2000), Hearst (1997) is one of only three text segmentation algorithms based on lexical cohesion to be cited and the only one of those to be mentioned beyond mere citation. Jurafsky and Martin (p. 663) characterize TextTiling as a "robust technique."

performance on text segmentation and has already been fine-tuned, making the comparison a meaningful one. Further, while modification of the TextTiling process to incorporate semantic frame information is readily accomplished, improved results are likely to be hard to come by. Hearst reports that in earlier work, incorporation of thesaural information into the algorithm degraded performance. Thus, the attempt to improve the performance of TextTiling by incorporating semantic information is not a matter of targeting low-hanging fruit.

Section 6.2 explains the TextTiling algorithm, while Section 6.3 discusses how that algorithm has been enhanced to incorporate semantic frame information. Section 6.4 presents the results of comparing the execution of an unenhanced version of TextTiling with SemFrame1 and SemFrame2 (which represent two levels of semantic frame enhancement of TextTiling). The performance of the three versions on detecting boundaries between concatenated tasks is evaluated. Both SemFrame1 and SemFrame2 are shown to achieve significantly better results than an unenhanced TextTiling, under a variety of specific conditions.

6.1 Hearst's TextTiling Algorithm

The gist of the TextTiling algorithm is as follows:

- Because sentences are not of uniform length, the input text is first divided into *token-sequences*. Hearst has used a token-sequence size of 20 to approximate the length of an average sentence.

- Words in the text are matched against a lengthy stop list; words not excluded on this basis are then analyzed to identify their root forms. In subsequent processing, terms from the text are represented by the corresponding lemma.
- A lexical similarity score is computed for gaps between token-sequences by comparing frequency data for lemmas occurring in a block of token-sequences before the gap with frequency data for lemmas occurring in an equal number of token-sequences after the gap. This score thus measures lexical similarity across blocks. Hearst has used a block size of 10 token-sequences to approximate the length of an average paragraph.
- For the similarity measure a normalized inner product is used:

$$score(gap_i) = \frac{\sum_t w_{t,b_1} w_{t,b_2}}{\sqrt{\sum_t (w_{t,b_1})^2 \sum_t (w_{t,b_2})^2}}$$

where i refers to the gap, t represents any term found in either or both of b_1 , the block of token-sequences before i , and b_2 , the block of token-sequences after the gap; the weight w for any term is its frequency within the block (the use of a *tf-idf* weight did not prove beneficial). The prototypical boundary is detected at a gap with a low lexical similarity score that is flanked on both its right and left sides by gaps with higher lexical similarity scores.

- The depth score of a gap is computed by finding the highest peak gaps to its left and right. (Hearst introduces a smoothing step to eliminate small perturbations of an extended slope.) The difference between each of the right and left peaks and the gap's valley is computed and the two differences summed; this gives the depth score for the

gap. Gaps with scores above a cutoff are hypothesized to occur at boundaries between text segments; Hearst uses the mean (\bar{s}) and standard deviation (σ) of the depth scores to set the cutoff.⁴⁶

The basic assumption underlying TextTiling is that text segments on some topic will be marked by repetitions of content words corresponding to that topic. A related assumption is that when the topic changes, as would occur between the segments of a single document, the vocabulary used will change concurrently. Consistent with these assumptions, the primary aim of TextTiling is to segment a single text into multi-paragraph discourse units by detecting corresponding changes in vocabulary. Exploiting TextTiling to detect boundaries between documents is therefore not its primary intended usage. Hearst points out that if two documents are on similar topics, the depth score for a gap between text segments in a tightly cohesive document may be greater than the depth score for the gap between the two documents.⁴⁷ To this may be added that, at least in theory, the words used at the beginnings and ends of documents might have some *a priori* probability of similarity, since both, for example, are likely to contain summary statements and thus to use summarizing language.⁴⁸

⁴⁶Hearst experimented with a liberal cutoff, $\bar{s} - \sigma$, and a conservative cutoff, $\bar{s} - \sigma / 2$. The liberal cutoff achieves higher recall; the conservative cutoff achieves higher precision.

⁴⁷Even if the two documents are on the same topic, the paraphrase phenomenon itself raises the possibility that different words would be used in the two documents, especially if authored by different persons.

⁴⁸The texts used for the evaluation do not necessarily start or end at document or even paragraph boundaries.

While circumstances contrary to the assumptions underlying TextTiling can be imagined, they are not expected to dominate a comparative evaluation of TextTiling and a semantic-frame-enhanced TextTiling. Because the number of distinct topics one may write on and the number of distinct roots in the general vocabulary are both large, the likelihood of two random texts being on the same subject or using the same content words is small, even given the potentially confounding influences of homonymy and polysemy.⁴⁹ Moreover, if there is a tendency to use summarizing words in both the beginnings and ends of documents, these words should be included on the stop list. In general, we may therefore assume a significant vocabulary shift between the end of one document and the beginning of another. Thus, especially when averaged over multiple trials, depth scores at the boundaries between concatenated documents can be expected to be high.

The only catch is our lack of knowledge of how frequently the vocabulary between the segments within a single document shift as dramatically as the vocabulary between documents. On the one hand, even between well-delineated segments of a text, some shared concept and/or vocabulary *may* carry over. On the other hand, there is no inherent reason why such carryover *must* occur. We therefore expect that the TextTiling approach

⁴⁹Homonymy is not expected to cause significant problems in falsely inflating lexical similarity scores because of the one-sense-per-discourse phenomenon (Gale, Church, and Yarowsky 1992). The occurrence of homonymous senses of a single word form in two randomly chosen documents is possible, but would not be frequent. Polysemy being more common than homonymy, the occurrence of different polysemous senses of a word in two randomly chosen documents is higher. At the same time, however, the polysemous senses of a word may or may not reflect a single topic. Thus the occurrence of polysemy might or might not prove detrimental to the success of TextTiling. In general, while homonymy and polysemy could have some affect on TextTiling, their adverse effects are assumed to be relatively insignificant.

should be able to detect boundaries between documents, using depth scores based on the lexical similarity of surrounding blocks of text, but we would not assume that these scores should be higher than all of the depth scores internal to a document.

Despite Hearst's expressed qualms, she reports having evaluated the TextTiling algorithm on a "document" consisting of 44 Wall Street Journal articles, using standard information retrieval (IR) performance measures—recall and precision—as they are applied to ranked retrieval. Specifically she recorded at each interval of 10 detected boundaries how many of those boundaries corresponded to between-document breaks (thus capturing information from which to compute precision) and how many between-document breaks had been identified as of that point (thus capturing information from which to compute recall). She identifies the break-even point between recall and precision occurring where both are 0.67.⁵⁰ She notes, on the one hand, that the score she achieved compares favorably with scores reported for other implementations of her algorithm. On the other hand, however, the evaluations were carried out over different texts, making the resulting comparisons only "suggestive."

6.2 Modification of TextTiling to Incorporate Semantic Frames

The paraphrase problem is relevant to the text segmentation task, much as it is to other tasks relying on natural language understanding. A writer often tries to create a

⁵⁰Given the typical inverse relationship between recall and precision, a break-even point between the two measures, if it occurs, would correspond to the highest value of their harmonic mean, or F-score, a measure commonly used to combine recall and precision into a single measure.

more interesting text by varying the words or phrases used to refer to a single concept. To the extent that the variation is syntactic in nature, the likelihood of the words and phrases used to talk about the concept being related morphologically is high. To the further extent that the morphological analysis of the input text is perfect, the shared roots identified for syntactic variants permit TextTiling's use of term repetition to recognize the cohesiveness of the text despite syntactic variation in how a concept is expressed.

Another manifestation of the paraphrase problem involves synonymy. As Hearst's lexical similarity measure only takes term repetition into account, it ignores the contribution synonymy makes to the cohesiveness of a text. Hearst's reliance on term repetition also overlooks the contribution made to textual cohesiveness by other lexical semantic relations, such as hyperonymy, hyponymy, antonymy, and branching relationships.

Semantics frames, however, do take such lexical semantic relationships into account. In running the TextTiling algorithm, if we substitute for a term a reference to the semantic frame(s) it evokes, a higher lexical similarity score should result in those circumstances where a concept is referred to in myriad ways and where that variation of expression resides in the lexical layer.

Trying to exploit the correspondence between words and the semantic frames they evoke is, unfortunately, not without its own set of dilemmas.

- The sets of semantic frames that have been identified are the product of a first round of development and are subject to further refinement and improvement. No doubt some frames are missing. At the same time, the frames that have been identified incur

both recall and precision failures: Some nouns and verbs that evoke a frame are not listed in association with it, while other nouns and verbs that are listed should not be. The extent of the recall failures is unknown; some number of the precision failures are erroneous word senses assigned on the basis of an imperfect mapping between LDOCE and WordNet verbs. All of these problems will affect the semantic frame enhancement of TextTiling.

- Since frames generalize over situation types, the number of available frames can always be expected to be rather smaller than the number of available lemmas. This means that in mapping words to the semantic frames they evoke, some degree of the assumed distinctiveness between texts that would register on the word level is likely to be lost on the frame level. Thus the substitution of semantic frames for terms may have a negative effect on precision at the same time that it should have a positive effect on recall. (Effective semantic frame disambiguation, however, could help improve precision without adversely affecting recall.)
- For the sets of frames that have been generated, the list of evoking verbs and nouns are specified in terms of WordNet senses. Input to a text segmentation process is highly unlikely to have undergone prior word sense disambiguation. This means that the translation of text word to semantic frame may need to undergo disambiguation in the frequently occurring event that a given *word* corresponds to multiple frames.
- But even if the input text undergoes word sense disambiguation, a semantic frame disambiguation process would still be needed, since a specific *word sense* may evoke multiple frames. Because of the precision quandary mentioned above, it is important

that this disambiguation be done; substituting multiple semantic frames for text words prior to TextTiling analysis would further minimize the distinctiveness of the input that permits the algorithm to be successful.

The disambiguation process that has been implemented takes advantage of part-of-speech (POS) tagging previously assigned to the input text. First, only nouns and verbs are selected to be mapped to semantic frames, since only nouns and verbs were associated with them in tasks 1 and 2. Second, nouns and verbs are differentiated from one another, so that only those frames associated with a word as the appropriate part-of-speech are considered in the disambiguation stage. Although texts are not routinely POS-tagged, highly accurate taggers are readily available; reliance on such tagging is therefore not a deal breaker.

In the evaluation, one version of semantic frame enhancement, SemFrame1, substitutes frame identifiers for only those text words with a single frame association; other text words are retained. In a second version of semantic frame enhancement, SemFrame2, all text words associated with one or more frames are replaced by the identifier for the most likely frame; again, other text words are retained. Semantic frame disambiguation takes into account that multiple words evoke the same frame and that when a frame operates within a text segment, it is not uncommon for more than one frame-evoking word to occur within that segment. Accordingly, semantic frame disambiguation proceeds by computing weights for frame possibilities over blocks of text and then making frame assignments on the basis of these weights. This process starts with the frame with the highest weight for the block, assigning it to all terms associated with that frame in the block, and then taking

up in turn each of the next most highly weighted frames that meet a weight threshold.

Frame assignments continue while frame weights meet or exceed the threshold and words remain in the block that are associated with some frame, but have not yet been assigned a specific frame. The blocks of text are processed using the same moving window approach used in the TextTiling algorithm. If a frame has been assigned to a term, the assignment remains fixed as the window slides forward to consider subsequent token-sequences.

The computation of a frame's weight takes into account how many potential invocations of the frame exist in the block, how many frames are associated with terms in the block, and how many terms evoke the frame. Specifically,

$$wgt_f = \frac{\sum_w freq_w}{(avg_frame_count * frame_size)}$$

where w represents any word in the block that evokes frame f ; $freq$ refers to the number of times w occurs in the block; avg_frame_count is, for all the words in the block that evoke the frame, the average number of frames they evoke; and $frame_size$ refers to the number of words that evoke the frame. This weight takes its inspiration from the $tf \cdot idf$ measure frequently used in information retrieval. The numerator is a measure of term frequency, summed over all terms in the block that are associated with the frame. The denominator parallels the concept of document frequency, bringing into play two measures which, as their values increase, the probability of the frame in question being a good assignment decreases. With regard to avg_frame_count : The larger the number of frames that could be assigned to a word, the less likely that any given frame is the correct assignment. With

regard to *frame_size*, the larger the number of associated nouns and verbs, the higher the *a priori* probability that a frame will be assigned. In order for such frames not to dominate the frame assignment process (thus effectively reducing the number of available frame assignments to those with many evoking words), word frequency and *frame_size* must balance each other by occurring in the numerator and denominator, respectively.

6.3 Evaluation Results

The effectiveness of semantic frames for dealing with the paraphrase problem is evaluated by comparing the performance of TextTiling using term repetition alone with two versions of semantic-frame-enhanced TextTiling, one using frame identifiers to replace terms when only one frame applies and the other using frame identifiers to replace all terms associated with one or more semantic frames. These three text segmentation strategies are referred to as TextTiling, SemFrame1, and SemFrame2, respectively.

Several versions of associating nouns and verbs with semantic frames emerged from task 1, differing on the basis of the threshold used for merging frames after the final clustering step in task 2; thresholds (1.0, 1.5, and 2.0) are more effective in the text segmentation task than others. These thresholds affect both SemFrame1 and SemFrame2. In addition, the disambiguation incorporates two variables that affect its results: (1) the block size over which disambiguation takes place and (2) a weight threshold which must be met for frame disambiguation to take place. Five values for both disambiguation block size and weight threshold have been examined. TextTiling is unaffected by the merging threshold and the disambiguation variables, and so is invariant. SemFrame1 is affected by

the merging threshold, but not by the disambiguation variables and thus exists in 3 variations. SemFrame2 is affected not only by the merging threshold, but also by the disambiguation block size and weight threshold; SemFrame2 thus has 75 variations.

The comparative evaluation has been executed over SEMCOR texts, a subset of the texts in the Brown Corpus, which have been WordNet sense tagged. The WordNet sense tagging is ignored in the evaluation, with the exception that the root lemma is used in lieu of the text word for all three approaches. The extraction of the root lemma parallels the morphological analysis of text words called for in the TextTiling algorithm. It differs, however, in one respect: WordNet often recognizes phrases as lexical units. Thus the extracted lemmas may be phrases as well as single words.⁵¹ The lists of nouns and verbs evoking semantic frames may include phrasal nouns (but not phrasal verbs); this occurs if a noun synset identified in task 2 as part of the participant structure of the frame includes within its set of synonyms a phrasal noun.⁵² In addition to extracting the root from the SEMCOR text, SemFrame1 and SemFrame2 also use the part-of-speech designation to do frame substitutions only for nouns and verbs, as previously discussed.

⁵¹If the phrase is used repeatedly in the text, TextTiling's use of term repetition is affected only in that a repetition of a phrase will count as only one repetition rather than counting as a repetition for each word in the phrase. However, if a text block uses both a phrase and a singleton word that is part of the phrase and has the same meaning as the entire phrase, the repetition of the term will not be recognized by TextTiling. The repetition can only be recognized by SemFrame1 or SemFrame2 if the phrase and the word evoke the same frame and/or are part of the same synset; the complexity of the frame assignment process is such that the repetition will not always be recognized when these conditions hold.

⁵²Approximately .2% of the input text are phrasal nouns that match nouns corresponding to semantic frames.

Within SEMCOR are two sets of texts taken from the Brown Corpus: Brown 1 has 103 texts; Brown 2 has 83 texts. All open class words in these 186 texts are WordNet-sense tagged. Table 5 (Francis and Kucera 1979) shows how many files of each genre are included in the two subsets. The numbers in square brackets indicate how many files (each approximately 2,000 words long) of each category are included in the entire corpus.

To promote greater efficiency in the evaluation step, SEMCOR texts are pre-processed. This pre-processing includes:

- extracting the root lemma for all nouns, verbs, adjectives, and adverbs, retaining only those words not on the stop list fine-tuned by Hearst for use with TextTiling,⁵³
- noting whether a text word is a noun or a verb, and
- recording which token-sequence each text lemma occurs within.

SemFrame1 input is further pre-processed to replace lemmas with frame identifiers if only one semantic frame is associated with the lemma. SemFrame2 input is further pre-processed to replace lemmas with frame identifiers according to the disambiguation process set forth above.

Token-sequence boundaries are defined in terms of numbers of words in the input text, regardless of their presence on the stop list. Accordingly, the number of units in token-sequences after processing against the stop list varies. The effect of this variation is

⁵³Closed class words—pronouns, conjunctions, and prepositions—are initially excluded from consideration by virtue of not having been analyzed in SEMCOR. They would likewise be excluded on the basis of their inclusion on the stop list.

Category	Brown1	Brown2
Reportage (political, sports, society, spot news, financial, cultural) [44]	7	0
Editorial (institutional, personal, letters to the editor) [27]	2	0
Reviews (theatre, books, music, dance) [17]	3	0
Religion (books, periodicals, tracts) [17]	4	0
Skills and hobbies (books, periodicals) [36]	6	8
Popular lore (books, periodicals) [48]	4	15
Belles Lettres, biography, memoirs, etc. (books, periodicals) [75]	3	15
Miscellaneous (government documents, foundation reports, industry reports, college catalog, industry house organ) [30]	1	11
Learned (natural sciences; medicine; mathematics; social and behavioral sciences; political science, law, education; humanities; technology and engineering [80]	33	10
General fiction (novels, short stories) [29]	29	0
Mystery and detective fiction (novels, short stories) [24]	2	9
Science fiction (novels, short stories) [6]	2	0
Adventure and Western fiction (novels, short stories) [29]	1	9
Romance and love story (novels, short stories) [29]	1	5
Humor (novels, essays, etc.) [9]	5	1
Total	103	83

Table 5. Distribution of Files across Brown Corpus Subsets in SEMCOR

minimized within the text segmentation task because a token-sequence is never processed alone, but is always processed within a block of token-sequences. Table 6 shows the distribution of token-sequence sizes (i.e., after processing against the stop list). Table 7 shows the parallel distribution of block sizes.

The sizes of token-sequences and of blocks takes on some degree of significance because in the pre-processing of the Brown texts, the recording of token-sequences is

1 21	4 302	7 2438	9 3375	11 1838	13 363	15 32	17 2
2 62	5 793	8 3211	10 2778	12 944	14 121	16 9	18 3
3 129	6 1433						

Table 6. Token-Sequence Sizes and Count

36 2	49 8	60 58	71 214	82 515	93 539	104 174	115 22
39 1	50 14	61 45	72 220	83 580	94 532	105 154	116 18
40 1	51 7	62 80	73 270	84 590	95 480	106 140	117 13
41 1	52 9	63 80	74 303	85 625	96 423	107 95	118 13
42 6	53 21	64 84	75 311	86 634	97 371	108 92	119 5
43 2	54 22	65 102	76 363	87 636	98 342	109 74	120 3
44 3	55 17	66 128	77 398	88 642	99 366	110 53	121 3
45 3	56 24	67 146	78 425	89 617	100 299	111 52	122 1
46 4	57 26	68 181	79 446	90 581	101 279	112 50	
47 3	58 31	69 163	80 513	91 639	102 207	113 38	
48 5	59 49	70 224	81 486	92 582	103 185	114 32	

Table 7. Block Sizes and Count

1 15	3 16	5 24	7 15	9 12	11 6	12 5	13 1
2 28	4 20	6 28	8 13	10 10			

Table 8. Final Token-Sequence Sizes and Count

done on a file-by-file basis. This means that each file starts with a full token-sequence, but often ends with only a partial token-sequence. Table 8 shows the distribution of sizes of final token-sequences; in comparison with Table 6, it can be seen that final token-sequences are shorter than the average token-sequence. Table 9 shows the parallel distribution of sizes of blocks ending with final token-sequences. In comparison with Table 7, it can be seen that even though final blocks are shorter than the average block, they are not the shortest of blocks. The manner of recording token-sequences is unlikely

50 2	67 6	73 1	79 5	85 7	91 11	96 4	101 3
59 1	68 3	74 7	80 6	86 7	92 3	97 2	102 1
60 1	69 3	75 4	81 5	87 10	93 5	98 4	103 1
63 1	70 2	76 9	82 6	88 6	94 6	99 4	104 1
65 2	71 2	77 7	83 3	89 10	95 2	100 2	110 1
66 1	72 3	78 6	84 4	90 6			

Table 9. Final Block Sizes and Count

to affect the text segmentation results in any significant manner.

During preliminary investigation there were 75 variations of SemFrame2, 3 variations of SemFrame1 (each of which corresponds to 25 of the SemFrame2 variations), and 1 version of TextTiling. Twenty-five texts were generated for each SemFrame2 variation. Each of these texts was formed by concatenating 51 randomly selected Brown2 texts; random selection was done without replacement. Each of the TextTiling, SemFrame1, and SemFrame2 algorithms were executed against the same concatenated texts.

Since a concatenated text comprises 51 Brown Corpus texts, 50 internal boundaries will be detected under ideal circumstances. The performance of each algorithm on each concatenated text was scored by examining as many as 150 of the gaps produced by each algorithm, in order by descending weight, and determining whether those gaps correspond to boundaries between the pre-concatenated texts. At each gap, recall and precision are computed, followed by a computation of the aggregate F-score, the harmonic means between recall and precision. The highest F-score represents the best tradeoff between recall and precision; this score is used to measure the goodness of the text segmentation algorithms.

Tables 10-12 show the average of the highest F-scores achieved over the 25 concatenated texts for each SemFrame2 variation and corresponding TextTiling and

SemFrame1 executions. In each cell, the highest F-score is bolded. In 8 of 75 sets of conditions, the highest F-score is achieved by TextTiling; in 48 of the conditions, SemFrame1 achieves the highest F-score, and in 22 of the conditions, SemFrame2 achieves the highest F-score; ties under several conditions account for the sum of best score counts totaling more than 75. SemFrame1's average best F-score is higher than TextTiling's average best F-score at each of the three weight thresholds over which SemFrame1 varies. While the actual improvement in performance is small, the results show generally that semantic frames can improve performance on a knowledge-based task and show specifically that the semantic frames induced from LDOCE and WordNet are sufficiently good to accomplish this improvement.

But SemFrame1 incorporates only limited semantic frame information, as it substitutes frames for terms only when no disambiguation is needed. The stronger case for semantic frame enhancement of TextTiling hinges on the degree to which SemFrame2 performs better than TextTiling based on term repetition alone. A visual examination of the initial tests shows that SemFrame2 performs better when the cluster threshold is 1.5 or 2.0, the disambiguation block size is 5 or higher, and the weight threshold is 2.0 or higher. Further text segmentation evaluation has been conducted over the resulting 24 variations. Besides this limitation on the number and identity of SemFrame1 and SemFrame2 variations being tested, these trials differ from the preliminary set of tests in two other respects: (1) In the second round of text segmentation trials, the text segmentation algorithms have been executed over a larger number of randomly generated texts, 125 instead of 25. (2) In the

Weight thresholds					
	1.0	2.0	5.0	7.5	10.0
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 1)	0.6380 0.6476 0.5900	0.6444 0.6476 0.594	0.6524 0.6540 0.5972	0.6400 0.6436 0.5844	0.6408 0.6408 0.5784
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 2)	0.6436 0.6492 0.6316	0.6440 0.6528 0.6276	0.6476 0.6552 0.6332	0.6216 0.6300 0.6112	0.6404 0.6544 0.6384
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 5)	0.6392 0.6472 0.6416	0.6360 0.6412 0.6296	0.6200 0.6276 0.6260	0.6556 0.6540 0.6324	0.6536 0.6692 0.6416
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 7)	0.6532 0.6452 0.6116	0.6468 0.6476 0.6232	0.6492 0.6564 0.6104	0.6496 0.6472 0.6312	0.6648 0.6688 0.6460
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 10)	0.6516 0.6624 0.6248	0.6456 0.6448 0.5992	0.6560 0.6476 0.6000	0.6404 0.6500 0.5972	0.6412 0.6548 0.6176

Table 10. Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 1.0

Weight thresholds						
	1.0	2.0	5.0	7.5	10.0	
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 1)	0.6356 0.6440 0.6112	0.6476 0.6612 0.6132	0.6420 0.6552 0.6320	0.6460 0.6552 0.6388	0.6384 0.6580 0.6468	
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 2)	0.6628 0.6736 0.6492	0.6372 0.6476 0.6248	0.6488 0.6588 0.6320	0.6420 0.6536 0.6568	0.6520 0.6516 0.6536	
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 5)	0.6340 0.6428 0.6296	0.6480 0.6608 0.6524	0.6468 0.6612 0.6640	0.6496 0.6628 0.6456	0.6232 0.6340 0.6256	
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 7)	0.6284 0.6344 0.6168	0.6428 0.6480 0.6604	0.6496 0.6508 0.6532	0.6304 0.6416 0.6492	0.6328 0.6464 0.6468	
TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 10)	0.6452 0.6488 0.6500	0.6296 0.6452 0.6332	0.6448 0.6540 0.6616	0.6380 0.6440 0.6584	0.6440 0.6632 0.6604	

Table 11. Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 1.5

	Weight thresholds				
	1.0	2.0	5.0	7.5	10.0
TextTiling	0.6508	0.6500	0.6484	0.6440	0.6464
TextTiling + SemFrame1	0.6548	0.6512	0.6564	0.6548	0.6516
TextTiling + SemFrame2 (disambiguation block size = 1)	0.6104	0.6052	0.6212	0.6224	0.6360
TextTiling	0.6320	0.6388	0.6492	0.6288	0.6528
TextTiling + SemFrame1	0.6424	0.6412	0.6540	0.6372	0.6484
TextTiling + SemFrame2 (disambiguation block size = 2)	0.6208	0.6164	0.6224	0.6124	0.6460
TextTiling	0.6408	0.6408	0.6436	0.6336	0.6420
TextTiling + SemFrame1	0.6384	0.6540	0.6492	0.6356	0.6556
TextTiling + SemFrame2 (disambiguation block size = 5)	0.6392	0.6540	0.6612	0.6448	0.6624
TextTiling	0.6308	0.6584	0.6372	0.6636	0.6452
TextTiling + SemFrame1	0.6308	0.6668	0.6480	0.6636	0.6480
TextTiling + SemFrame2 (disambiguation block size = 7)	0.6380	0.6668	0.6568	0.6696	0.6592
TextTiling	0.6532	0.6368	0.6420	0.6292	0.6584
TextTiling + SemFrame1	0.6600	0.6428	0.6456	0.6284	0.6664
TextTiling + SemFrame2 (disambiguation block size = 10)	0.6492	0.6260	0.6472	0.6388	0.6700

Table 12. Average Maximum Harmonic Mean (F-score) over 25 Trials, Where Cluster Threshold = 2.0

first round of trials, concatenated texts were drawn from Brown2 alone⁵⁴; in the second round, concatenated texts have been drawn from both Brown1 and Brown2. Table 13 displays the average maximum harmonic mean (F-score) for this second round of text segmentation trials. Here SemFrame1 achieves the best score 8 of 24 times, while SemFrame2 achieves the best score 16 of 24 times.

But the average F-score tells only part of the story, especially since performance varies across the various combinations of concatenated texts (F-score values vary approximately 20% more or less than the reported averages). The real issue is how the algorithms match up on a trial-by-trial basis. For this, we use a paired-*t* test to investigate how the performance of the algorithms compare against each other. For the 24 variations tested, SemFrame1's F-scores are significantly higher than TextTiling's scores four times at $p = .05$, two of those at $p = .01$. At the same time, SemFrame2's F-scores are significantly higher than TextTiling's scores nine times at $p = .05$, six of those at $p = .01$. Again, while the actual improvement in performance is small, the results show that semantic-frame-enhanced TextTiling outperforms non-enhanced TextTiling at statistically significant levels.

As noted at the outset, TextTiling is recognized as a solid text segmentation algorithm; previous attempts to enhance it with semantic information similar to the semantic frame information used here degraded performance. Add to this that the semantic frame data

⁵⁴Initially, scoring for SemFrame2 incorporated training data from Brown 1. In the current version of SemFrame2, SemFrame2 scoring is completely unsupervised.

		Weight thresholds			
		2.0	5.0	7.5	10.0
1	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 5)	0.60688 0.60952 0.61664*	0.59992 0.60200 0.59664	0.59576 0.60024+ 0.59904	0.60136 0.59968 0.60216
2	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 7)	0.60248 0.60872* 0.60888+	0.60912 0.61120 0.61064	0.60592 0.60808 0.60800	0.59856 0.59992 0.60280
3	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 10)	0.60048 0.60808* 0.60728+	0.59480 0.59504 0.58968	0.60592 0.60576 0.60656	0.59280 0.59728+ 0.59480
4	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 5)	0.60464 0.60328 0.60736	0.59488 0.59232 0.60320*	0.60232 0.60192 0.60536	0.60552 0.60288 0.60872
5	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 7)	0.60024 0.60136 0.59944	0.60744 0.60712 0.61200	0.60696 0.60936 0.61456*	0.59032 0.59016 0.60080*
6	TextTiling TextTiling + SemFrame1 TextTiling + SemFrame2 (disambiguation block size = 10)	0.60136 0.59552 0.60328	0.59896 0.60056 0.60736*	0.61144 0.61160 0.62040*	0.59656 0.59688 0.60312+

Table 13. Average Maximum Harmonic Mean (F-score) over 125 Trials, Cluster Threshold = 1.5 or 2.0

Rows 1-3 = threshold 1.5; rows 4-6 = threshold 2.0.

SemFrame{1,2} significantly better than TextTiling at p = .05 (+) or p = .01 (*)

used here is still of a preliminary nature, and a picture of wider potential for significant semantic frame enhancements to knowledge-intensive tasks begins to emerge.

6.4 Summary

This chapter reports on the incorporation of semantic frame information in a knowledge-based task affected by the paraphrase problem, namely, text segmentation. A well-regarded text segmentation algorithm, TextTiling (Hearst 1997), has been enhanced with semantic frame information at two different levels. The resulting three versions of TextTiling are compared in their ability to detect boundaries between concatenated SEMCOR texts. Both semantic-frame-enhanced versions of TextTiling are shown to yield small, but statistically significant, improvement over the unenhanced version under specific conditions, using F-score (with equal weighting for recall and precision) as the evaluation criterion. Hearst's previous attempt to enhance TextTiling with thesaural relationships was not able to attain improved results. The success of semantic frame information in achieving improved performance for TextTiling, contrasted with the failure of thesaural information, justifies a place for frame semantics at the semantic relationships table.

Chapter 7

Discussion

The SemFrame enhancement of TextTiling achieved improvements in performance over unenhanced TextTiling, but the improvements were only modest. The moderate extent of this improvement suggests that, at the same time that SemFrame is able to provide valid semantic information, the framesets and frame structures that it generates are not without their weaknesses. In this light, Section 7.1 presents some of the limitations imposed on SemFrame by decisions originating in its design. Section 7.2 then examines the quality of the framesets generated by SemFrame in comparison with the corresponding output from the FrameNet project.

7.1 Threats to the SemFrame Approach

There are several threats to the SemFrame approach that may adversely affect the quality of its output. First, the input data set drawn from LDOCE and WordNet, while generally of high quality, is somewhat sparse for the uses to which it is put. Second, the mapping technique used for picking out specific word senses from WordNet and LDOCE is shallow. The first threat tends to curb recall at various turns in the development of SemFrame, while the second threat has a constraining effect on both the recall and the precision of its outputs. A third threat involves inadequacy in the multi-step process of merging smaller verb groups into verb groups that are intended to correspond to frames, but sometimes fail to achieve an appropriate degree of correspondence.

7.1.1 Data Sparseness

As discussed in Sections 1.3 and 3.1, SemFrame is based on data drawn from two lexical resources—LDOCE and WordNet—rather than on corpus data. This has allowed work to proceed in terms of a full general inventory of word senses (although crossing back and forth between the two sets of word senses in LDOCE and WordNet has caused its own problems) and has facilitated the identification of semantic argument types.

However, the amount of data available for any given word sense has been limited to an LDOCE definition, a WordNet gloss, a couple of example sentences, LDOCE subject field code data, and WordNet semantic relationships. In many cases, the amount of available data has been much less. Section 8.1 addresses the data sparseness problem by discussing the use of additional resources in future work.

The data analyzed have not been sufficient to reveal the full set of verbs that evoke a common frame. For instance, where previous frame semantic analysis has identified a broad set of verbs that evoke the same frame—as, for instance, with the COMMERCIAL TRANSACTION frame or the RISK frame—SemFrame fails to identify some of the verb senses known to evoke the frame and/or fails to identify that several verb framesets in reality evoke the same frame. It is presumed that this pattern carries over to other frames, especially to those with greater structural complexity (which are the very frames where the paraphrase problem is the hardest to solve). Since further elaboration of the frame—for example, the identification of noun senses that evoke the frame—relies on the verb framesets associated with a frame, recall failures at this stage have repercussions all the way down the line.

The recall ratio among noun framesets is a function not only of the recall ratio among corresponding verb framesets, but is also a function of the amount of data available about nouns associated with those verb senses. The generation of morphologically-derived nouns has played an especially important role in identifying the semantic arguments of a frame in SemFrame. However, some of a frame's semantic arguments—this is probably most often true of optional arguments—will not correspond to morphologically-related nouns. Thus, improving the identification of the full set of arguments for a frame cannot depend solely on improving the identification of morphologically-related noun senses, but must address their identification through sentences using the associated verbs. This particular limitation in the current SemFrame approach can be overcome by having access to a larger number of sentences for each verb sense.

7.1.2 Shallowness in Mapping Technique

As noted, the use of LDOCE and WordNet as data sources has had the advantage of being able to work with word senses rather than with words. Being able to work at this level of specificity has contributed significantly to the SemFrame venture. However, LDOCE and WordNet have different scopes, both at the word and word sense levels. Given a specific word sense in one resource, there are many difficulties in trying to identify the equivalent word sense in the other resource. First, the word may not be present in the other resource. Second, the word's specific sense may not be present in the other resource. Third, the two resources, while covering more or less the same semantic scope for a word, may divide up the word's senses in non-equivalent ways. These three

complications all constrain the possibility of finding equivalent word senses across the two resources.⁵⁵

Fortunately, it has not been necessary in working with the two resources to map a given word sense in one resource to a fully equivalent word sense in the other resource. All that has been necessary is to identify senses of the word in the target resource that evoke the same frame as the source word sense. Often word senses that evoke the same frame will have one or more shared words (more accurately, lemmas) in their definitions and/or example sentences; the more words they have in common, the more likely it is that they evoke the same frame. However, there are many reasons—starting with the paraphrase phenomenon itself—why two words that evoke the same frame may not have any shared words in their definitions and example sentences; thus, mapping between the two resources based on word (stem) matches will not necessarily have very high recall. For this reason, definition data from LDOCE have been supplemented with semantically related words to increase the possibility of common words. This move, however, increases the possibility that word senses that do not evoke the same frame will share word stems in their definitions and/or example sentences and thus decreases precision.

Since the mapping technique is used in a variety of contexts within SemFrame, there are many situations that arise in which incorrect word senses of either verbs or nouns can

⁵⁵While recall is normally characterized in terms of a system's relative ability to retrieve available units, we may also characterize the absence of relevant units as a type of recall failure. Ultimately, it does not depend on whether a mapping fails because the system is unable to recognize an equivalent word sense or because no equivalent word sense is present to be recognized.

be associated with a frame. Such a failure may show up in a number of ways: (1) The set of verb senses associated with a frame may include one or more senses that do not evoke the frame. (2) The set of noun senses associated with a frame—whether morphologically related or drawn from a verb sense’s definition or example sentences—may include one or more senses that do not evoke the frame. (3) The arguments associated with a frame may reflect an incorrect word sense. (4) The name associated with a frame may reflect an incorrect word sense. Such precision failures do occur, but they do not generally overwhelm the results. The incorporation of additional link types in WordNet 2.0 (discussed in Section 8.1.1) and the subsequent use of WordNet as the primary data source will address this limitation.

7.1.3 Inadequacy of Merging Process

SemFrame has built up verb framesets incrementally, starting with pairs of verb senses that are deemed to evoke the same frame, then, in a multi-step process, merging smaller verb sense groups into larger verb sense groups. Ideally, this merging process concludes when the verb sense groups correspond to frames. However, frames may be general or specific, so a consistent ideal stopping point does not exist in reality. That does not mean, however, that anywhere the process stops will yield good results. On the one hand, in SemFrame’s task 1 output are verb groups that span multiple frames, where the merging process has gone too far. On the other hand, SemFrame’s task 1 output also includes sets of verb groups with overlapping membership that evoke the same frame, where the merging process has not gone far enough.

Correction of this problem is likely to require more sophisticated analysis of, for example, word frequency distribution, levels of word senses within the WordNet network structure, and so forth. The exact nature of the remediating action needed depends on how much and what kind of progress is made in addressing the two difficulties outlined in Sections 7.1.1 and 7.1.2.

7.2 SemFrame *vis-à-vis* FrameNet

One of the motivations for the current research effort is the slow pace and intensive effort associated with FrameNet’s handcrafting of semantic frames. The development of frames in FrameNet is labor-intensive, on the one hand, and opportunistic, on the other hand; the overall approach lacks systematicity. As such, the FrameNet approach is unlikely to identify a comprehensive set of semantic frames in the near term. Furthermore, most of this effort would need to be duplicated in transferring the FrameNet approach to specific subject domains or to other languages.

The approach taken within FrameNet for the development of semantic frames does, however, have a significant advantage. Major aspects of the frames as initially proposed—the associated lexical units that evoke the frame, the frame elements or slots that constitute the frame’s structure, the interface between lexical units and frame slots—are verified against corpus data. FrameNet’s data are thus of reliably high precision.

The quality of the FrameNet output supports the possible use of FrameNet as a gold standard against which to evaluate SemFrame’s output. There are, however, significant

constraints imposed by FrameNet's opportunistic development strategy. (1) As of summer 2003, only 382 frames had been identified within the FrameNet project. In contrast, SemFrame had identified almost four times as many frames. Even if some of SemFrame's frames should be conflated, there are presumably many semantic frames that FrameNet does not yet cover. (2) Low recall affects not only the set of semantic frames identified by FrameNet, but also the individual sets of frame-evoking units listed for each frame: No verbs are listed for 38.5% of FrameNet's frames, while another 13.1% of them list only 1 or 2 verbs. (3) Although many of FrameNet's frames are semantically motivated, some of them are syntactically oriented (for example, EXPERIENCER-OBJECT, EXPERIENCER-SUBJECT). This attentiveness to syntactic behavior belies the premise that FrameNet's frames are intended primarily to address the paraphrase problem, as SemFrame's frames are.

Given FrameNet's low recall and its sensitivity to syntactic behavior, an evaluation of SemFrame using FrameNet as a gold standard should be limited to ascertaining whether there are corresponding frames in SemFrame for the frames in FrameNet and, for frames identified by both systems, ascertaining to what degree the lexical units (most especially the verbs) identified by SemFrame can be shown to evoke those frames.

An initial comparison of SemFrame's output with FrameNet's output targets those frames within FrameNet that could conceivably have been generated within SemFrame. Of the 382 frames currently in FrameNet, 147 have no verbs among their associated lexical units. Of the 235 frames with verbs among their list of associated lexical units, 38

have only non-LDOCE verbs. This leaves 197 frames for which SemFrame could conceivably have produced a corresponding frame and set of evoking lexical units.

Because FrameNet's frames include frames of varying levels of generality/specificity—with some semantic areas being covered only by one general frame, some other semantic areas being covered only by a combination of specific frames, and still other semantic areas being covered by a mix of general and specific frames—the SemFrame/FrameNet comparison has proceeded in terms of finding counterparts (i.e., corresponding frames) rather than finding equivalents (i.e., out-and-out matches). Counterpart frames are deemed to correspond if the semantic scope of one frame is included within the semantic scope of the other frame or if the semantic scopes of the two frames have significant overlap. Of necessity, frame correspondence between FrameNet and SemFrame is looser than frame equivalence.

For example, we get something akin to frame equivalence between FrameNet's ACTIVITY_FINISH frame, which includes *finish*, *tie_up*, and *wrap_up*, and SemFrame's ENDING frame, which includes specific senses of *discontinue*, *end*, *finish*, and *adjourn*.⁵⁶ This contrasts with what we might designate frame correspondence between FrameNet's ACTIVITY_FINISH frame and SemFrame's START frame, which includes specific senses of *apply*, *start*, *inaugurate*, *use*, *misuse*, *pervert*, *abuse*, *beatify*, *commence*, *begin*, *halt*, *stop*,

⁵⁶The definitions for the specific LDOCE verb senses designated in SemFrame are:
To *discontinue*1.1 is to stop or end.
To *end*1.1 is to (cause to) finish.
To *finish*1.4 is to arrive or end (in the stated place or way).
To *adjourn*1.1 is to bring (a meeting, trial, etc.) to a stop, for a particular period or until a later time.

*originate, cease, end, indicate, terminate, finish, and launch.*⁵⁷ Inclusion of the semantic scope of one frame within the semantic scope of another and significant overlap of semantic scopes are operationally defined in terms of the inclusion or overlap of the sets of words (specifically, of verbs) that evoke the FrameNet and SemFrame frames. Since FrameNet lists evoking words, without specification of word sense, the comparison

⁵⁷The definitions for the specific LDOCE verb senses designated in SemFrame are:
 To apply1.2 is to bring or put into use.
 To start1.7 is to (cause to) play on a team at the beginning of a match.
 To start1.6 is to begin using.
 To start1.5 is to go from a particular point; have a beginning or lower limit.
 To start1.4 is to begin doing a job or a piece of work.
 To start1.2 is to (cause to) come into being; begin.
 To inaugurate1.1 is to introduce (someone important) into a new place or job by holding a special ceremony.
 To use1.1 is to employ; put to use.
 To misuse1.1 is to use (something) in a wrong way or for a wrong purpose.
 To inaugurate1.3 is to be the beginning of (something, an important period of time).
 To pervert1.3 is to change or twist (the meaning of words).
 To abuse1.3 is to handle without care; treat badly.
 To start1.3 is to (cause to) go into (movement or activity); begin.
 To beatify1.1 is (in the ROMAN CATHOLIC church) to declare (a dead person) blessed.
 To commence1.1 is to begin; start.
 To begin1.1 is to start; take the first step.
 To start1.1 is to begin (a course, journey, etc.); set out (on).
 To halt1.1 is to (cause to) stop.
 To stop1.1 is to (cause to) cease moving or continuing an activity.
 To originate1.1 is to (cause to) begin.
 To cease1.1 is to stop (an activity).
 To end1.1 is to (cause to) finish.
 To indicate1.1 is to point out.
 To indicate1.5 is to show a need for; suggest.
 To terminate1.1 is to (cause to) come to an end.
 To finish1.4 is to arrive or end (in the stated place or way).
 To finish1.1 is to reach or bring to an end; reach the end of (an activity).
 To launch1.4 is to cause (an activity, plan, way of life, etc.) to begin.

between these sets of words proceeds on the word level rather than on the word sense level.⁵⁸

For these purposes, two general correspondence scenarios have been adopted. In the first scenario, a SemFrame frame is deemed to correspond to a FrameNet frame if there is some (even small) degree of overlap between the verbs listed for a FrameNet frame and for a SemFrame frame and the respective names of the two frames are semantically related. In the second scenario, a SemFrame frame is deemed to correspond to a FrameNet frame based on verb overlap alone, where a considerably higher threshold is used.

Since both of the general correspondence scenarios require some degree of overlap between the set of verbs listed for a FrameNet frame and the set of verbs listed for the SemFrame frame, a minimal verb overlap criterion equivalent to the verb overlap threshold of the first scenario is applied uniformly as an initial filter; this filter is set forth in the Minimal_Verb_Overlap algorithm of Figure 15. SemFrame frames whose verb framesets partially match the set of evoking verbs given in FrameNet have been identified for each of the 197 FrameNet frames that include at least one verb present in LDOCE. This partial match occurs in either of two ways, distinguished by the number of verbs listed for the FrameNet frame. The first partial match type occurs where the FrameNet frame lists four or fewer verbs (true of over one-third of the FrameNet frames that list associated verbs). For these frames, a partial match occurs when any verb associated with the FrameNet

⁵⁸The comparison is implemented as if LDOCE verb senses were not specified. Multiple senses of an LDOCE verb are counted only once.

frame is matched by a verb associated with a SemFrame frame. The second partial match type occurs where the FrameNet frame lists five or more verbs. For these frames a partial match occurs when two or more verbs associated with the FrameNet frame are matched by verbs associated with the SemFrame frame.⁵⁹ These partial matches produce 858 pairs of FrameNet and SemFrame frames for the version of SemFrame based on a stage 1 clustering threshold of 1.5 and 893 pairs for the version of SemFrame based on a stage 1 clustering threshold of 2.0.

This partial match criterion has several motivations. First, the partial match criterion has been implemented as a filter for efficiency's sake, so that only those pairs of SemFrame frames and FrameNet frames with some minimal possibility of correspondence are given further consideration (as opposed to considering all of the 2.75-plus million possible pairs of FrameNet and SemFrame frames).

Second, automating the identification of corresponding frames presupposes that the sets of lexical items associated with the frames will match in some way. If comprehensive sets of evoking lexical items are given, the overlap between corresponding frames should be extensive. However, given that the recall level of evoking verbs/verb senses is not guaranteed to be high on either of the FrameNet or SemFrame sides, it is conceivable that corresponding frames will have only a single verb (or no verbs) in common. (The `Minimal_Verb_Overlap` criterion will fail to consider frame pairs where there are no verbs in common.) The threshold implemented for the initial filtering is thus a minimal one.

⁵⁹The thresholds for partial matches were set prior to seeing the results.

Third, FrameNet sets of frame-evoking lexical units are not word-sense-disambiguated. In order to compensate for this lack of specification and to improve the likelihood that word matches between the FrameNet and SemFrame frames also match on word *sense*, the minimal threshold has been raised from one to two where the number of verbs present in the FrameNet frame is five or greater, since the more verbs the two frames have in common, the more likely it is that the verbs used in the FrameNet frames correspond to the verb senses used in the SemFrame frames. The partial match criterion used thus balances the desire to use a minimal threshold (because of low recall) and the desire to match on word senses, not just on words. The criterion used is biased somewhat against SemFrame, since it is conceivable that corresponding FrameNet and SemFrame frames will not pass this initial filter and thus not be given further consideration.

As just noted, this initial filter is intended to be minimal: It is more concerned with identifying FrameNet and SemFrame frames that may correspond with each other (i.e., with promoting recall) than with deselecting frame pairs that do not correspond with each other (i.e., with promoting precision). Therefore, in addition to the initial minimal verb overlap filter, the 850+ pairs of FrameNet and SemFrame frames that potentially correspond to each other must meet at least one of three precision-oriented criteria in order to be deemed corresponding frames. The overall correspondence criteria are summarized in the Correspondence algorithm of Figure 15.

The first of these criteria involves the first general correspondence scenario. To meet this criterion, the FrameNet and SemFrame frame names must be semantically related to each other, as set forth in the Framename_Semrel algorithm of Figure 15 and further

```

Minimal_Verb_Overlap (FrameNet_verbs, SemFrame_verbs) =
  if (|FrameNet_verbs| <= 4)
    if exists  $x$  s.t.  $x \in \text{FrameNet\_verbs}$  and  $x \in \text{SemFrame\_verbs}$ , then return true
  endif
  else if (|FrameNet_verbs| >= 5)
    if (exists  $x$  s.t.  $x \in \text{FrameNet\_verbs}$  and  $x \in \text{SemFrame\_verbs}$ ) and (exists  $y$  s.t.  $y \in \text{FrameNet\_verbs}$  and  $y \in \text{SemFrame\_verbs}$ ) and  $x \neq y$ , then return true
  endif
  else return false

Framename_Semrel (FrameNet_name, SemFrame_name) =
  FrameNet_set = the concatenation of: (1) FrameNet_name, (2) individual components of
    FrameNet_name, or (3) morphological variants of FrameNet_name or its components
  SemFrame_set = the concatenation of: (1) SemFrame_name, (2) individual components of
    SemFrame_name, or (3) morphological variants of SemFrame_name or its components
  if (exists  $x$  s.t.  $x \in \text{FrameNet\_set}$  and  $x \in \text{SemFrame\_set}$ ) then return true

Verb_Semrel (FrameNet_verbs, SemFrame_verbs) =
  forall  $x \in \text{FrameNet\_verbs}$ 
    if  $x \in \text{SemFrame\_verbs}$ 
      shared_count++
  forall  $x \in \text{FrameNet\_verbs}$ 
    forall  $y$  s.t.  $x \in \text{WordNet synset } y$ 
      FN [ $x, y$ ] = all WordNet synsets directly related to  $y$ 
  forall  $x \in \text{SemFrame\_verbs}$ 
    forall  $y$  s.t.  $x \in \text{WordNet synset } y$ 
      SF [ $x, y$ ] = all WordNet synsets directly related to  $y$ 
  round = 0
  while (round < 2)
    round++
    forall  $m \in \text{FN } [a, b]$ 
      forall  $n \in \text{SF}[c, d]$ 
        if (exists  $z$  s.t.  $z \in m$  and  $z \in n$ , then
          good_count++; FN [ $c, d$ ] = SF [ $c, d$ ]; delete SF [ $c, d$ ]
        endif
      endforall
    endforall
  endwhile
  precision = ((good_count + shared_count) / |SemFrame_verbs|)
  if precision >= .5 then return true else return false

Verb_Majority (FrameNet_verbs, SemFrame_verbs) =
  forall  $x \in \text{FrameNet\_verbs}$ 
    if  $x \in \text{SemFrame\_verbs}$ 
      shared_verbs++
  if (shared_verbs / |FrameNet_verbs| >= .5 || shared_verbs / |SemFrame_verbs| >= .5)
    then return true

```

Figure 15. Algorithm for Correspondence of FrameNet and SemFrame Frames

```

Correspondence (FrameName_name, FrameNet_verbs, SemFrame_name, SemFrame_verbs) =
/* Partial match criterion/filter */
if Minimal_Verb_Overlap (FrameNet_verbs, SemFrame_verbs)
/* General correspondence scheme #1; first additional criterion */
if Framename_Semrel (FrameNet_name, SemFrame_name) then return true
/* General correspondence scheme #2; second and third additional criteria */
else if (Verb_Semrel (FrameNet_verbs, SemFrame_verbs) or Verb_Majority
(FrameNet_verbs, SemFrame_verbs)) then return true
endif
else return false

```

Figure 15. Algorithm for Correspondence of FrameNet and SemFrame Frames—*Cont.*

explained and exemplified below. The second and third of these criteria correspond to the second general correspondence scenario. The second criterion is met if half or more of the verbs from the SemFrame frame are semantically related to verbs from the FrameNet frame, as set forth in the Verb_Semrel algorithm of Figure 15 and further explained and exemplified below. The third criterion is met if the set of verbs shared by FrameNet and SemFrame framesets account for half or more of the verbs on either side of the comparison; this criterion is given as the Verb_Majority algorithm of Figure 15.

For example, the APPLY_HEAT frame in FrameNet includes 22 verbs: *bake, blanch, boil, braise, broil, brown, char, coddle, cook, fry, grill, microwave, parboil, poach, roast, saute, scald, simmer, steam, steep, stew, and toast*, while the BOILING frame in SemFrame includes 7 verbs: *boil, coddle, jug, parboil, poach, seethe, and simmer*.⁶⁰ Five of these

⁶⁰The definitions for the specific LDOCE verb senses designated in SemFrame are:
 To boil1.2 is to cook (food) in water at 100 degrees C.
 To boil1.4 is to cause to reach the stated condition by cooking in water.
 To coddle1.1 is to cook (eggs, fruit, etc.) slowly in water just below boiling point.
 To jug1.1 is to boil (meat) in a closed pot or JUG.
 To parboil1.1 is to boil until partly cooked.

verbs—*boil*, *coddle*, *parboil*, *poach*, and *simmer*—are shared across the two frames and constitute over half of the SemFrame frameset. The two frames are deemed to correspond by passing both the Minimal_Verb_Overlap and Verb_Majority criteria.

In comparing FrameNet and SemFrame frame names for semantic relatedness, Framename_Semrel must take into account that, in contrast to SemFrame frame names, all of which correspond to WordNet synsets from specific subnetworks of the noun network, FrameNet frame names have no consistent underlying basis. Some of the names are nouns (especially gerunds), for example, ABUNDANCE, BIRTH, GROOMING, QUESTIONING. Some are syntactically regular verbal phrases, for example, CAUSE_TO_BE_WET, MOVE_IN_PLACE, SHOOT_PROJECTILES. Some are phrases that include a process and a phase, for example, ACTIVITY_START, COMMERCE_SELL, EDUCATION_TEACHING. Some indicate a syntactic basis to the frame, for example, COTHEME, EXPERIENCER-OBJ, INCHOATIVE_ATTACHING.

Because of the degree of variation found in the naming of FrameNet frames, the process of establishing semantic relatedness between FrameNet frame names and SemFrame frame names has been optimized for higher recall. This optimization includes expanding frame names to include individual components of the names and words that are morphologically related through CatVar (Habash and Dorr 2003) to the frame name or one of its components. This set of words for each FrameNet and SemFrame frame name is then searched in both the noun and verb WordNet networks to find all the synsets that

To poach1.1 is to cook (eggs or fish) in gently boiling water or other liquid, sometimes in a special pan.

To seethe1.2 is to cook by boiling; STEW.

To simmer1.1 is to (cause to) cook gently in liquid at or just below boiling heat.

might correspond to the frame name. The last part of the expansion pulls in all synsets that are directly related to the synsets that correspond to the frame names. If the set of synsets gathered for a FrameNet frame name intersects with the set of synsets gathered for a SemFrame frame name, the two frame names are deemed to be semantically related.

For example, the FrameNet ADORNING frame contains 17 verbs: *adorn, blanket, cloak, coat, cover, deck, decorate, dot, encircle, envelop, festoon, fill, film, line, pave, stud, and wreath*. The SemFrame ORNAMENTATION frame contains 12 verbs: *adorn, caparison, decorate, embellish, embroider, garland, garnish, gild, grace, hang, incrust, and ornament*.⁶¹ Only two of the verbs—*adorn* and *decorate*—are shared. But the frame names are semantically related through WordNet verb synset 01323008, which includes *decorate, adorn* (which CatVar relates to ADORNING), *grace, ornament* (which CatVar relates to ORNAMENTATION), *embellish*, and *beautify*. The two frames are designated as

⁶¹The definitions for the specific LDOCE verb senses designated in SemFrame are:

To adorn1.1 is to add beauty or ornament to.

To adorn1.2 is to add importance or attractiveness to.

To caparison1.1 is to put a CAPARISON on (a horse), as was done in former times.

To decorate1.1 is to provide with something ornamental, for a special occasion.

To decorate1.3 is to be or serve as an ornamentation to.

To embellish1.1 is to make more beautiful, by adding ornaments.

To embellish1.2 is to add details, perhaps untrue, to (a statement or story).

To embroider1.1 is to do ornamental needlework on (cloth).

To garland1.1 is to put one or more garlands on (someone).

To garnish1.1 is to add to (something) as an ornament and, in the case of food, as an improvement to its taste.

To gild1.2 is to make bright as if with gold.

To grace1.1 is to ornament; give pleasure to, by one's / its presence.

To hang1.5 is to fix (wallpaper) on a wall.

To incrust1.1 is ENCRUST.

To ornament1.1 is to add ornament to.

corresponding frames by meeting the Minimal_Verb_Overlap and Framename_Semrel criteria.

This procedure implements no word sense disambiguation, so frame names could be found semantically related when relationships exist only between some word sense other than the one(s) appropriate to the frame name. That the comparison is only done where one or more of the verbs in the respective frames match each other reduces this possibility.

In comparing FrameNet and SemFrame verbs for semantic relatedness, a somewhat similar procedure, Verb_Semrel, is executed. In this case, there is no need to search WordNet's noun network, nor is any expansion for verb components or morphological variants performed. Instead, each FrameNet and SemFrame verb is searched for all WordNet verb synsets it belongs to, and a record is made of all the synsets to which each of these synsets is directly related.⁶² If the sets of synsets that correspond to two verbs share one or more synsets, the two verbs are deemed to be semantically related.⁶³ Once a verb associated with a SemFrame frame is found through this procedure to be semantically related to a verb associated with a FrameNet frame, the SemFrame verb data is deleted from the SemFrame side of the comparison and added to the FrameNet side of the comparison, as if it had been listed originally as a lexical unit evoking the FrameNet frame. One further round of comparison ensues, such that a SemFrame verb that is found to be semantically related to a SemFrame verb that has already been found to be semantically

⁶²All relationships that exist in WordNet 1.7.1 are employed.

⁶³Since the comparisons are done between WordNet synsets, it is really semantic relationships between verb senses that are being processed.

related to a FrameNet verb will also be established as a frame-evoking verb. If half or more of the verbs listed for a SemFrame frame are established as evoking the same frame as the list of WordNet verbs, then the FrameNet and SemFrame frames are hypothesized to correspond.

This procedure for detecting correspondence is needed for circumstances where FrameNet's list of frame-evoking verbs is much smaller than SemFrame's. This may result from low recall on the FrameNet side, low precision on the SemFrame side, and/or more general semantic scope on the SemFrame side. Establishing that half or more of the verbs on the SemFrame side of the comparison are semantically related to the fully precise set of verbs on the FrameNet side bolsters the likelihood that the SemFrame frame corresponds to the FrameNet frame against which the comparison is being made more than to any other Frame Net frame (excepting other FrameNet frames that are related hierarchically or compositionally).

For example, FrameNet ABUNDANCE frame includes 4 verbs: *crawl*, *swarm*, *teem*, and *throng*. The SemFrame FLOW frame likewise includes 4 verbs: *pour*, *teem*, *stream*, and *pullulate*.⁶⁴ Only 1 verb—*teem*—is shared, so the Verb_Majority criterion is not met, nor is the Framename_Semrel criterion met, as the frame names are not semantically related. The Verb_Semrel criterion, however, is met, most directly through the WordNet verb

⁶⁴The definitions for the specific LDOCE verb senses designated in SemFrame are:
To pour1.2 is (of people) to rush together in large numbers.
To pullulate1.1 is to breed or multiply quickly and in great numbers.
To stream1.1 is to flow fast and strong; pour out.
To teem1.1 is to be present in large numbers; ABOUND.

synset 01596222, which includes *pour*, *swarm*, *stream*, *teem*, and *pullulate*. All 3 of the SemFrame verbs not shared with FrameNet are in the same synset with 2 of the FrameNet verbs. Having met the Minimal_Verb_Overlap criterion, the two frames are deemed to correspond by the Verb_Semrel criterion.

Of the 197 FrameNet frames that include at least one LDOCE verb, 175 have been found to have a corresponding SemFrame frame through Figure 15's overall Correspondence function. But this 89% recall level should be balanced against the precision ratio of SemFrame verb framesets. After all, we could get 100% recall by listing all verbs in every SemFrame frame.

The Verb_Semrel function of Figure 15 computes a precision ratio for each pair of FrameNet and SemFrame frames being compared. At the 89% recall level, the overall precision is ca. 55%. By setting various minimum precision thresholds, the balance between recall and precision can be investigated. The effect of varying this threshold is shown in Tables 14 (stage 1 clustering threshold of 1.5) and 15 (stage 1 clustering threshold of 2.0). The harmonic mean of the frame recall and verb precision figures is also computed; the best balance (bolded in each of the tables) occurs for the SemFrame version based on a stage 1 clustering threshold of 2.0 and a minimum precision threshold of 0.4, which yields a recall of 83.2% and overall precision of 73.8%.

To interpret these results meaningfully, one would like to know if SemFrame achieves more FrameNet-like results than other available verb category data. A lower bound for the harmonic mean of frame recall and verb precision can be set, for example, by treating branches within the WordNet verb hierarchy as verb framesets. There are

Minimum precision	Number of frames that correspond (max: 197)	Recall (frame-based)	Total number of SemFrame verbs	Number of frame-evoking verbs	Precision (verb-based)	Harmonic mean of recall and precision
0.00	175	0.888325	5184	2884	0.556327	0.684178
0.05	175	0.888325	5046	2879	0.570551	0.694829
0.10	173	0.878173	4639	2848	0.613925	0.722650
0.15	169	0.857868	4348	2811	0.646504	0.737338
0.20	167	0.847716	4167	2780	0.667147	0.746669
0.25	166	0.842640	4093	2764	0.675299	0.749745
0.30	165	0.837563	3839	2696	0.702266	0.763971
0.35	161	0.817259	3655	2637	0.721477	0.766387
0.40	159	0.807107	3558	2600	0.730748	0.767031
0.45	152	0.771574	3285	2486	0.756773	0.764102
0.50	152	0.771574	3228	2460	0.762082	0.766798
0.55	141	0.715736	2772	2226	0.803030	0.756874
0.60	136	0.690355	2594	2124	0.818813	0.749117
0.65	129	0.654822	2260	1918	0.848673	0.739251
0.70	114	0.578680	1895	1674	0.883377	0.699279
0.75	112	0.568528	1738	1562	0.898734	0.696475
0.80	107	0.543147	1503	1384	0.920825	0.683269
0.85	95	0.482234	1044	1009	0.966475	0.643424
0.90	90	0.456853	840	831	0.989286	0.625055
0.95	89	0.451777	742	742	1.000000	0.622378
1.00	89	0.451777	742	742	1.000000	0.622378

Table 14. FrameNet-based Evaluation of SemFrame, Cluster Threshold = 1.5

Minimum precision	Number of frames that correspond (max: 197)	Recall (frame-based)	Total number of SemFrame verbs	Number of frame-evoking verbs	Precision (verb-based)	Harmonic mean of recall and precision
0.00	175	0.888325	5126	2929	0.571401	0.695459
0.05	175	0.888325	4984	2924	0.586677	0.706657
0.10	173	0.878173	4670	2901	0.621199	0.727665
0.15	169	0.857868	4381	2864	0.653732	0.742016
0.20	167	0.847716	4196	2833	0.675167	0.751666
0.25	167	0.847716	4117	2816	0.683993	0.757104
0.30	166	0.842640	3858	2746	0.711768	0.771694
0.35	165	0.837563	3711	2698	0.727028	0.778391
0.40	164	0.832487	3600	2656	0.737778	0.782276
0.45	159	0.807107	3406	2576	0.756312	0.780884
0.50	159	0.807107	3336	2544	0.762590	0.784217
0.55	148	0.751269	2869	2305	0.803416	0.776468
0.60	143	0.725888	2650	2180	0.822642	0.771242
0.65	137	0.695431	2307	1969	0.853489	0.766396
0.70	121	0.614213	1979	1750	0.884285	0.724912
0.75	118	0.598985	1832	1645	0.897926	0.718605
0.80	112	0.568528	1527	1415	0.926654	0.704702
0.85	100	0.507614	1108	1073	0.968412	0.666085
0.90	96	0.487310	923	912	0.988082	0.652711
0.95	95	0.482234	798	798	1.000000	0.650685
1.00	95	0.482234	798	798	1.000000	0.650685

Table 15. FrameNet-based Evaluation of SemFrame, Cluster Threshold = 2.0

357 root-level verb synsets (i.e., with no hypernym synset) with at least one descendant synset in the WordNet 1.7.1 verb network. All descendant verb synsets of a root-level verb synset can be assumed (naively) to evoke the same frame. These WordNet verb framesets have been compared with FrameNet's verb framesets in essentially the same way that the comparison with SemFrame was conducted. For these purposes, the name of a WordNet-based frame is taken from the words for the root-level synset. Results of this comparison are set forth in Table 16. The best balance between recall and precision can be seen to occur with a minimum precision threshold of 0.15, which yields a recall of 52.8% and overall precision of 46.6%. The SemFrame results, at 83.2% recall and 73.8% precision are markedly better than the WordNet-based lower bound. (FrameNet itself constitutes the corresponding upper bound of 100% recall and 100% precision.)

Similar comparisons have been made with Levin verb classes and Lin and Pantel verb clusters. The verb classes in Levin (1993b) are based on an analysis of the syntactic properties of verbs, especially their ability to be expressed in various diathesis alternations; her classes reflect the assumption that the syntactic behavior of a verb is determined in large part by its meaning. Levin's analysis results in 258 leaf nodes within a minimally organized network structure (including a larger group of subclasses [e.g., 32—Verbs of Desire subsumes 32.1—*Want* Verbs and 32.2—*Long* Verbs] and a much smaller group of undivided top-level classes [e.g., 38—Verbs of Sounds Made by Animals]). Membership in these verb classes totals 10,967 verb tokens (3,108 verb types). For purposes of comparison against FrameNet, the names of the verb classes have been hand-edited to isolate the word that best captures the semantic sense of the

Minimum precision	Number of frames that correspond (max: 197)	Recall (frame-based)	Total number of SemFrame verbs	Number of frame-evoking verbs	Precision (verb-based)	Harmonic mean of recall and precision
0.00	203	1.030457	655426	11773	0.017962	0.035309
0.05	128	0.649746	28302	8549	0.302063	0.412403
0.10	108	0.548223	19734	8003	0.405544	0.466211
0.15	102	0.517766	16256	7578	0.466166	0.490613
0.20	92	0.467005	15062	7375	0.489643	0.478056
0.25	78	0.395939	13074	6907	0.528300	0.452642
0.30	58	0.294416	11772	6553	0.556660	0.385135
0.35	49	0.248731	10741	6222	0.579276	0.348026
0.40	47	0.238579	9773	5866	0.600225	0.341441
0.45	42	0.213198	9528	5762	0.604744	0.315255
0.50	37	0.187817	6918	4562	0.659439	0.292365
0.55	24	0.121827	4904	3515	0.716762	0.208257
0.60	20	0.101523	4252	3139	0.738241	0.178499
0.65	10	0.050761	2460	2041	0.829675	0.095670
0.70	8	0.040609	2436	2025	0.831281	0.077435
0.75	8	0.040609	1720	1521	0.884302	0.077652
0.80	4	0.020305	1663	1478	0.888755	0.039702
0.85	3	0.015228	1657	1473	0.888956	0.029944
0.90	2	0.010152	4	4	1.000000	0.020101
0.95	2	0.010152	4	4	1.000000	0.020101
1.00	2	0.010152	4	4	1.000000	0.020101

Table 16. Comparative Results 1: FrameNet-based Evaluation of WordNet Verb Hierarchies

class. For example, the names of classes 32.1, 32.2, and 38 have become *Want*, *Long*, and *Sounds*. Table 17 gives the results of the comparison, with the best balance between recall and precision occurring with a minimum precision threshold of 0.20, yielding 56.9% recall and 55.0% precision. While these results are better than the WordNet-based lower bound, they fall far short of SemFrame’s 83.2% recall and 73.8% precision.

Unlike the hand-crafted verb categories of WordNet and Levin, the verb clusters based on Pantel and Lin (2002) and Lin and Pantel (2001b)⁶⁵ are computationally-based. These verb groups have been induced using an unsupervised clustering algorithm that operates on a similarity matrix. The similarity matrix derives from a set of collocations in a large text corpus, where collocations are defined as dependency relationships that occur more frequently than would be predicted were the occurrence of two words independent of each other.⁶⁶ The clustering has placed 6,588 verb tokens (5,160 verb types) into 272 verb groups (“concepts”). For purposes of comparison against FrameNet, the first verb in a cluster has been designated as the cluster’s name.⁶⁷ Results

⁶⁵The general program of inducing semantic classes from a text corpus is set out in Lin and Pantel (2001b), while an improved clustering algorithm, whose results are evaluated here, is presented in Pantel and Lin (2002).

⁶⁶Lin and Pantel’s approach shares with SemFrame the view that verbs that tend to share the same arguments are semantically related. A major difference between the approaches lies in the sources of data they have analyzed (SemFrame uses lexical resources while Lin and Pantel operate on corpus data). Another significant difference lies in the many other kinds of information about verb senses that SemFrame avails itself of.

⁶⁷Lin and Pantel have taken a similar approach, “naming” their verb clusters by the first three verbs listed for a cluster, i.e., the three most similar verbs.

Minimum precision	Number of frames that correspond (max: 197)	Recall (frame-based)	Total number of SemFrame verbs	Number of frame-evoking verbs	Precision (verb-based)	Harmonic mean of recall and precision
0.00	156	0.791878	13728	3492	0.254371	0.385053
0.05	138	0.700508	8533	3372	0.395172	0.505295
0.10	123	0.624365	6798	3244	0.477199	0.540952
0.15	117	0.593909	6132	3166	0.516308	0.552396
0.20	112	0.568528	5595	3076	0.549777	0.558995
0.25	102	0.517766	5139	2977	0.579296	0.546806
0.30	96	0.487310	4821	2891	0.599668	0.537682
0.35	91	0.461929	4318	2729	0.632006	0.533746
0.40	87	0.441624	4160	2671	0.642067	0.523309
0.45	77	0.390863	3734	2492	0.667381	0.492995
0.50	74	0.375635	3473	2369	0.682119	0.484475
0.55	68	0.345178	3031	2138	0.705378	0.463527
0.60	58	0.294416	2531	1850	0.730936	0.419757
0.65	50	0.253807	1958	1492	0.762002	0.380783
0.70	46	0.233503	1512	1194	0.789683	0.360429
0.75	40	0.203046	1265	1018	0.804743	0.324274
0.80	33	0.167513	530	459	0.866038	0.280726
0.85	22	0.111675	310	280	0.903226	0.198774
0.90	14	0.071066	102	99	0.970588	0.132435
0.95	11	0.055838	66	66	1.000000	0.105769
1.00	11	0.055838	66	66	1.000000	0.105769

Table 17. Comparative Results 2: FrameNet-based Evaluation of Verb Classes (Levin 1993b)

of the comparison are given in Table 18, where the best balance between recall and precision can be seen to occur with a minimum precision threshold of 0.15, yielding 47.2% recall and 40.7% precision. These results fall far short of SemFrame’s 83.2% recall and 73.8% precision.

7.3 Summary

Section 7.1 discusses three issues that constrain the quality of SemFrame’s results: the sparseness of the input data set, the shallowness of the mapping between WordNet and LDOCE word senses, and some degree of inadequacy in the process that merges smaller groups of verbs into larger groups of verbs that correspond to frames. Antidotes to the first two of these threats are offered; the third threat will require more careful analysis of the merging process.

Section 7.2 compares SemFrame’s inventory of semantic frames and verb framesets with the handcrafted output of FrameNet. SemFrame’s frames cover much of the semantic territory covered by FrameNet, with almost 90% of the FrameNet frames that have one or more LDOCE verbs listed among their lexical units having at least one corresponding frame in SemFrame. The maximum harmonic mean between frame-based-recall and verb-based-precision is achieved with a recall of 83.2% of FrameNet’s frames having counterparts in SemFrame at the same time that 73.8% of the SemFrame verbs evoke the corresponding FrameNet frame. These results compare very favorably with a WordNet-verb-hierarchy-based lower bound of 52.8% recall and 46.6% precision. Verb classes from Levin (1993b) achieve 56.9% recall and 55.0% precision, while verb

Minimum precision	Number of frames that correspond (max: 197)	Recall (frame-based)	Total number of SemFrame verbs	Number of frame-evoking verbs	Precision (verb-based)	Harmonic mean of recall and precision
0.00	142	0.720812	19054	2377	0.124751	0.212691
0.05	116	0.588832	7644	2252	0.294610	0.392727
0.10	104	0.527919	5876	2127	0.361981	0.429479
0.15	93	0.472081	4955	2015	0.406660	0.436935
0.20	85	0.431472	4329	1908	0.440748	0.436061
0.25	77	0.390863	3826	1797	0.469681	0.426663
0.30	68	0.345178	2923	1549	0.529935	0.418053
0.35	62	0.314721	2623	1451	0.553183	0.401192
0.40	59	0.299492	2343	1346	0.574477	0.393724
0.45	54	0.274112	2226	1297	0.582659	0.372827
0.50	49	0.248731	1576	993	0.630076	0.356664
0.55	40	0.203046	1228	814	0.662866	0.310868
0.60	31	0.157360	804	575	0.715174	0.257961
0.65	25	0.126904	554	422	0.761733	0.217562
0.70	22	0.111675	425	336	0.790588	0.195706
0.75	15	0.076142	302	247	0.817881	0.139315
0.80	12	0.060914	165	141	0.854545	0.113721
0.85	6	0.030457	53	50	0.943396	0.059009
0.90	5	0.025381	38	37	0.973684	0.049472
0.95	4	0.020305	28	28	1.000000	0.039801
1.00	4	0.020305	28	28	1.000000	0.039801

Table 18. Comparative Results 3: FrameNet-based Evaluation of Verb Clusters
(Pantel and Lin 2002, Lin and Pantel 2001b)

clusters based on Pantel and Lin (2002) and Lin and Pantel (2001b) achieve 47.2% recall and 40.7% precision.

Chapter 8

Contributions and Future Work

The major contributions of this dissertation are both theoretical and practical. On the theoretical side, both of the general hypotheses that governed the research have been confirmed. On the practical side, two types of resources have been developed.

On the theoretical level, the research has first demonstrated that semantic frames—which heretofore have been jumpstarted from native speaker intuitions—can be induced automatically from available machine-readable lexical resources. Specifically, the work has demonstrated that existing lexical tools can serve successfully as sources from which to induce sets of verbs and nouns that evoke a common semantic frame. The work has also shown that these same data support the identification of the internal participant structure of a semantic frame. In a head-to-head comparison with frames in FrameNet, the frames developed by the SemFrame approach achieve a recall ratio of 83.2% and the verbs listed for frames achieve a precision ratio of 73.8%.

Second, this research has confirmed the potential benefit of semantic frames to knowledge-intensive tasks. This latter contribution was made by generating improved performance on a text segmentation task with a semantic-frame-enhanced version of TextTiling. It is notable that semantic frame enhancement could achieve statistically significant improvement, since attempts to enhance the algorithm with thesaural relations had degraded performance. This direct comparison with thesaural relations further validates the value of frames as a part of the semantic structure of language.

On the practical level, a large number of semantic frames have been identified both extensionally and intensionally. The extensional identification consists in enumerating an average of from 3.1 to 3.3 (clustering thresholds 2.0 and 1.5, respectively) WordNet noun synsets and from 6.6 to 14.1 (clustering thresholds 2.0 and 1.5, respectively) WordNet verb synsets that evoke each frame. The intensional identification consists in enumerating one or more slots or frame elements that define the participant structure of the frame.⁶⁸

Along the way, this work has also made other, smaller, contributions:

- Three new concepts have been developed and/or named:
 - ▶ The notion of a *strong sense* of a word has been defined, namely, a family of senses that share part of speech and domain (as identified by the lexicographic file within WordNet), whose occurrences within SEMCOR dominate over all other usages combined.
 - ▶ The notion of a *strong relationship* between words has been defined, namely, a direct semantic relationship between any two strong word senses.
 - ▶ The term *frameset* has been proposed to lexicalize the concept of the set of words or word senses that evoke a common frame.
- Two new measures have been introduced:

⁶⁸Of the frames that are identified extensionally approximately one-third are not identified intensionally, due to insufficient data. This still leaves over 1000 frames being identified both extensionally and intensionally.

- ▶ A new similarity measure for word senses was developed. This measure does not assume, as other similarity measures do, that the definitions of similar words are of the same length.
- ▶ A modified conceptual density measure, based on the Agirre and Rigau (1995) conceptual density measure, was implemented. This measure takes actual descendant treesize and node weights into account.
- A high precision strategy for mapping between WordNet and LDOCE word senses, in the context of semantic frames, was presented. The strategy proved useful in a variety of contexts and builds on the new similarity measure.
- An unsupervised, general-purpose semantic frame disambiguation procedure was implemented. It was this procedure that permitted enhancement of the TextTiling with semantic frames for the many nouns and verbs that evoke multiple frames.

At the same time that the research undertaken here achieved a certain degree of success on the two fronts noted above—that is, with respect to the induction of semantic frames and the use of semantic frames to enhance knowledge-intensive tasks—it also has laid a dual foundation for continuing research.

First, as noted previously, this research demonstrates that the extensional identification of semantic frames—that is, the identification of sets of nouns and verbs that evoke a common semantic frame—can be induced automatically from available lexical resources. The use of additional resources to support this task are discussed in Section 8.1. The potential contribution of these resources should be explored, with the

goal of generating frame-evoking content word sets (i.e., framesets) with both higher precision and higher recall.

Second, this research shows that the incorporation of semantic frame information into a knowledge intensive task—namely, text segmentation—can improve performance on the task. This result can be expanded in at least three ways:

- If the semantic frame information base can be enhanced through the exploration of further data resources, we can expect that the positive contribution of semantic frames to text segmentation will likewise be magnified.
- Semantic frames admit all levels of specificity. The emphasis in the current research has been on identifying a broad array of semantic frames in general usage. This is an appropriate starting point, because we can expect that the frames that are relevant to more specific contexts will be hierarchically or compositionally related to general usage frames. Section 8.2 briefly discusses how the processes used here to discover general semantic frames would need to be modified to identify the set of semantic frames operating within a specific subject domain.
- Many other knowledge intensive tasks also encounter the paraphrase problem or face other challenges that semantic frames address. Section 8.3 explores how semantic frame information can enhance other applications, including word sense disambiguation, information retrieval, information extraction, question-answering, text summarization, and machine translation. Among these will be found use not only for the extensionally-oriented frame-evoking content word sets used in the text

segmentation task, but also for the intensionally-oriented internal participant structure of semantic frames, whose direct use has not been demonstrated or evaluated here.

8.1 Other Resources

The basic approach used here to identify semantic frames, both extensionally and intensionally, starts by gathering data from numerous sources as to which pairs of verb senses are likely to evoke a common frame. These data are further processed to support the combining, clustering, and merging of verb sense pairs into larger groups.

Information about nouns associated with these verb senses support the delineation of the corresponding frame's internal structure, which in turn provides the basis for identifying noun senses evoking the same frame. Membership of noun and verb senses in the same semantic frame is then used to enhance a text segmentation algorithm. The success achieved by this enhancement is evidence of the soundness of the basic approach. The limited extent of that improvement raises the question whether possible variations on the approach might yield better results.

One arena whose alternatives have not yet been explored involves selecting data sources to mine for semantic relationships between verb senses. A possible option to pursue is the analysis of contributions made by the ten data sources used here to the enhancement of TextTiling. Such an analysis might result in the elimination of some sources or in a re-weighting of the evidence from individual sources. However, inasmuch as the current results fail to provide coverage for over one-third of the verb

senses in LDOCE and well over half of the verb senses in WordNet, the more pressing need is for new sources or better use of the present sources.

In the current research effort, LDOCE plays the dominant role in that WordNet verb sense pairs have been mapped to LDOCE verb sense pairs prior to the combining, clustering, and merging processes that produced verb sense groups. Table 19 summarizes the relevant statistical properties of the final clustering process as applied to the word stems extracted from LDOCE definitions and from WordNet glosses.

The number of verb *senses* in LDOCE and the number of verb *synsets* in WordNet are of equal magnitude. The number of verb *senses* in WordNet 1.7.1 (24,169), however, is roughly double the number of its synsets, demonstrating WordNet's more comprehensive and specific coverage. The difference between LDOCE's restricted vocabulary and WordNet's open-ended vocabulary shows up clearly in the number of stems used in their vocabularies, with the WordNet vocabulary approximately twice the

Property	LDOCE	WordNet
Number of input units	12,663 (verb senses)	13,214 (synsets)
Number of distinct stems	3,292	6,344
Number of clusters generated	2,778	2,267
Number of input units represented in clustering output	8,084	5,092
Average number of members per cluster	2.9	2.25
Number of (LDOCE verb sense) pairs generated from clustering output	13,694	2,772

Table 19. Comparative Final Clustering Results for LDOCE and WordNet

size of the LDOCE vocabulary. As a result of the less disparate input to the clustering process, LDOCE clustering produces a larger number of clusters than WordNet clustering does; the LDOCE clusters are on average larger as well. The proportion of input units retained within the output of the clustering process—64% for LDOCE, 39% for WordNet—also reflects the added power derived from the restricted vocabulary used in LDOCE. The huge difference in pairs generated by the two clusterings is only partially attributable to the larger number of clusters generated for LDOCE, the larger size of those clusters, and the larger number of distinct units (in this case, verb senses) retained in the clustering output. A significant part of this difference stems from repeated failures in the process of mapping WordNet verb senses to LDOCE verb senses to find equivalent senses. Some of these recall failures reflect differences in the defining vocabularies used, but many of them reflect gaps in LDOCE's coverage. There simply are many cases where WordNet verb senses have no counterpart in LDOCE. Thus much of the power associated with WordNet's broader coverage has been lost.

Why was LDOCE given the predominant role to start with? The major reason is that LDOCE's restricted vocabulary for definitions and example sentences was presumed to be beneficial for the discovery of frame-related verb senses. As just seen in Table 19, the restricted vocabulary in LDOCE results in producing a larger number of verb sense groups for LDOCE than for WordNet, even though WordNet is a slightly larger data source; the verb sense groups produced for LDOCE also include more members than those produced for WordNet. Furthermore, the restricted vocabulary results in a considerably larger percentage of LDOCE verb senses being retained in the verb sense groups than WordNet

verb synsets. A second reason for deferring to LDOCE is that the machine-readable version of LDOCE contains subject field codes, some of which correspond to high-level frames. A third reason is that the words in LDOCE are defined using standard lexicographic conventions, i.e., in terms of genus and differentiae; the systematicity of this approach further accommodates the semantic frame discovery process. In contrast, WordNet makes no attempt to standardize the vocabulary of its glosses and example sentences, lacks an equivalent to LDOCE's subject field codes, and provides glosses, whose intent has been more to distinguish between word senses than to give definitions.

The assumed advantages of LDOCE over WordNet have not been fully realized. Specifically, as noted in Section 4.1.1, the restricted vocabulary of LDOCE is more a basic vocabulary than a controlled vocabulary (in which meanings and terms are ideally in a 1:1 correspondence). While its use still increases the likelihood that a concept will be expressed using the same vocabulary across definitions, it may also increase the possibility that unrelated words will use the same vocabulary in their definitions. Second, subject field codes have been assigned to only 30% of the verb senses (EAGLES 1998, "The Longman Dictionary and Thesaurus"), and the assignments have been criticized as inconsistent (Boguraev and Briscoe 1989, p. 17). Moreover, some of the subjects are too general (e.g., science, economics, business) to correspond well to semantic frames. Still, the specific subject field codes underlying the LZ.pairs data set are, by and large, a solid data source.

8.1.1 WordNet 2.0

Some of the advantages which favored LDOCE over WordNet have been set on their head by the recent release of WordNet 2.0. Two particular sets of changes between versions 1.7.1 and 2.0 portend more effective use of WordNet in the future expansion of this research. The first change is that derivational morphology links between nouns and verbs have been added. The second change is that an indication of topical classification has been added to some synsets.

Derivational Morphology Links

Derivational morphology links relate specific senses of nouns and verbs to each other; there are no within part-of-speech derivational morphology links. Of 79,689 noun synsets, 11,709 of them contain one or more links to morphologically-related verbs. Of 13,508 verb synsets, 8,906 of them contain one or more links to morphologically-related nouns. Derivational morphology links are examples of what WordNet refers to as “lexical links”; this means that WordNet provides derivational morphology links between specific word senses, not just between synsets. Linkage on this level is, of course, appropriate: Derivational morphology links are word-specific. In WordNet 2.0, the links are further delineated; they are specific to particular word senses.

For example, WordNet includes 6 senses of the noun *pace*, 2 senses of the noun *pacemaker*, 2 senses of the noun *pacing*, and 4 senses of the verb *pace*, as shown in Table 20. Altogether, 16 derivational morphology links exist between noun senses based on *pace* and verb senses based on *pace*; 40 such links (between 10 noun senses and 4 verb senses)

POS	WordNet 2.0 synsets for <i>pace</i> , <i>pacer</i> , and <i>pacing</i>	With derivational morphology link to:
N	Pace 1. (00273018) pace, stride, tread a step in walking or running	(Vs) Pace 1, 3, 4
N	Pace 2. (04785229) pace, rate the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated"	(Vs) Pace 2, 3
N	Pace 3. (12879290) yard, pace a unit of length equal to 3 feet; defined as 91.44 centimeters; originally taken to be the average length of a stride	(V) Pace 1
N	Pace 4. (12980794) footstep, pace, step, stride the distance covered by a step; "he stepped off ten paces from the old tree and began to dig"	(Vs) Pace 1, 3, 4
N	Pace 5. (14416248) pace, gait the rate of moving (especially walking or running)	(Vs) Pace 1, 3, 4
N	Pace 6. (14418352) tempo, pace the rate of some repeating event	(V) Pace 2
N	Pacer 1. (02300941) pacer a horse trained to a special gait in which both feet on one side leave the ground together	(V) Pace 4
N	Pacer 2. (02301101) pacer, pacemaker, pacesetter a horse used to set the pace in racing	(V) Pace 2
N	Pacing 1. (00276632) pacing walking with slow regular strides	(V) Pace 3
N	Pacing 2. (14400938) tempo, pacing (music) the speed at which a composition is to be played	(None)
V	Pace 1. (00477372) pace, step measure (distances) by pacing; "step off ten yards"	(N) Pace 1, 3, 4, 5
V	Pace 2. (00679319) pace regulate or set the pace of; "Pace your efforts"	(N) Pace 2, 6 (N) Pacer 2
V	Pace 3. (01873407) pace walk with slow or fast paces; "He paced up and down the hall"	(N) Pace 1, 2, 4, 5 (N) Pacing 1
V	Pace 4. (02031682) go at a pace; "The horse paced"	(N) Pace 1, 4, 5 (N) Pacer 1

Table 20. Derivational Morphology Links among *Pace* Words

would be generated if morphology alone were taken into account, without recourse to semantics. The derivational morphology links in WordNet 2.0 thus promise to be both a rich source of relationships between verbs and corresponding nouns, but also a finely nuanced one. In addition, semantic relationships between morphologically-related verbs (or nouns) can also be drawn from this data. For instance, the semantic relatedness of senses 1, 3, and 4 of the verb *pace* are evident in the linking of these three verb senses to senses 1, 4, and 5 of the noun *pace* (and vice versa).

The derivational morphology links in WordNet 2.0 can substitute for two aspects of the current research. First, the LM.pairs data source related LDOCE verb senses that were morphologically related to one another. As just noted, a secondary analysis of derivational morphology links between noun and verb senses in WordNet 2.0 can also reveal semantic relatedness between verb senses. Using WordNet data will promote far greater precision in the linking of verb senses based on morphology than was possible in the generation of the LM.pairs data source.

Second, Section 5.1.2 noted that certain arguments of a verb sense, specifically those corresponding to the grammatical subject and direct object, are often only implied in its definition. Sometimes this gap is offset by the existence of nouns that are morphologically related to the verb sense—as *payer*, *payee*, and *payment* are related to *pay*; often such nouns are good indicators of the semantic type of the very arguments that are missing from the definition. Given this phenomenon, this research effort included a process to discover such nouns prior to and as input to the computing of conceptual density; the results of this computation led to the identification of both a name for the overall frame

and the internal participant structure of the frame. This process was comprised of appending nominal suffixes to verb stems to form possible nouns; these word forms were verified against WordNet to see if they corresponded to actual nouns. If so, the LDOCE-WordNet mapping algorithm was used to identify corresponding WordNet noun senses. This process generated, for any given LDOCE verb sense, a set of WordNet noun synsets both morphologically and semantically related to it. However, because of the paraphrase problem, which also plagues the mapping algorithm, the results enjoyed much higher precision than recall: Many relevant noun-verb relationships are assumed to have been missed. Access to derivational morphology links in WordNet 2.0 should vastly improve this part of the process, with recall and precision being limited only by the comprehensiveness and accuracy of the derivational morphology links themselves.

Topical Classification

Another significant addition to WordNet 2.0 is the inclusion of domains, of which three kinds are recognized: category, region, and usage. In all three cases, the domain is associated with a specific noun synset; members of the domain, regardless of their part of speech, point to the noun synset. A category domain collocates words pertaining to a subject (e.g., the verb *overbid* pertains to the *auction*, *auction_sale*, *vendue* domain); a region domain collocates words applicable to specific geographic areas (e.g., the noun *bairn* comes from *Scotland*); a usage domain collocates words with certain linguistic properties (e.g., the noun *motel* is an example of a *blend*, *portmanteau_word*,

portmanteau). Of these, only category domains are semantic in basis and have relevance for the identification of semantic frames.

Category domain links in WordNet 2.0 perform the same function as LDOCE's subject field codes. LDOCE recognizes 100 main subject fields and 246 subdivisions. These codes are assigned to 30% of LDOCE's verb senses (ca. 4,750) and 59% of its noun senses (ca. 22,000) (EAGLES report). Many noun senses and most verb senses are thus left outside the subject field coding process. At the same time, it should be noted that some of the subject field codes are very general and oft used; the ten most commonly assigned subject field codes (those corresponding to medicine, politics, food, military, religion, law, business, science, clothing, and music) account for approximately 35% of all assignments.

WordNet 2.0 includes 422 noun synsets that are treated as category domains; of these, 45 further indicate that they are members of other domains.⁶⁹ Significant overlap exists between LDOCE's subject fields and WordNet's category domains. Indeed, the largest difference at present lies in the much smaller proportion of entries in the WordNet database that participate in its subject classification scheme: Fewer than 3,650 noun senses and 1,200 verb senses are members of category domains.⁷⁰

⁶⁹It is unclear at present whether a systematic hierarchical structuring of subject domains is intended or whether the presence of noun synsets that indicate both that they have domain members and that they are domain members of another domain reflects the vagaries of an initial release.

⁷⁰Continuation of the current research effort will assist in bringing further category domains to light and/or in supporting the assignment of WordNet synsets to category domains.

Category domain links in WordNet 2.0 thus have the potential of substituting for the LT.pairs and LZ.pairs data sources within the current research. Although the number of word senses linked to subject classes in WordNet 2.0 is much smaller than in LDOCE, the overall effect of using WordNet 2.0 data in lieu of LDOCE data might not be so very different. Because the weighting scheme used with LT.pairs and LZ.pairs penalized subject field codes that were assigned frequently, most of the verb sense pairs in especially the LT.pairs data source did not survive the imposition of the cumulative weight threshold.

Additional Advantages

Two new features in WordNet 2.0—derivational morphology links and category domain links—have been shown to correspond to the same kinds of semantic relationships underlying three of the six LDOCE data sources, namely, LM. pairs, LT.pairs, and LZ.pairs. Two features at the heart of WordNet, in any version—the collocation of synonyms in synsets and the hierarchical organization of synsets—further correspond to two other LDOCE data sources, namely, LS.pairs and LI.pairs, respectively. Meanwhile, the WC.pairs data source corresponds to the LC.pairs data source.

That all six of the data sources used from LDOCE have correspondences in WordNet 2.0 does not mean that equivalent verb sense pairs would be generated within WordNet 2.0 as were generated within LDOCE. After all, corresponding WordNet data sources exist for half of the LDOCE data sources in the current research effort, and only 10% of the verb sense pairs were supported by both LDOCE and WordNet. Moreover, the

restricted vocabulary used in LDOCE resulted in twice as many verb pairs being generated for LC.pairs in comparison to WC.pairs. It is thus not the case that the use of WordNet 2.0 alone can fully substitute for the use of LDOCE. However, the larger coverage of WordNet 2.0 (WordNet 2.0 has more than twice as many noun synsets as LDOCE has noun senses) and especially the elimination of the need to map back and forth across LDOCE and WordNet word senses makes the sole use of WordNet 2.0 appealing.

A further advantage comes with WordNet 2.0 that has only been mentioned in passing. In previous versions, relationships operated only inside the network for a particular part of speech; there were no semantic links between nouns and verbs, for example. With the addition of derivational morphology links and category domain links, WordNet 2.0 is becoming a single, integrated network rather than a set of separate networks. This integration brings with it implied relationships, which can be further mined. For example, and as suggested above, the set of verb senses with derivational morphology links to the same set of nouns are themselves both morphologically and semantically related.

As a resource under continuing development, WordNet will continue to be improved. One of the near-term enhancements envisioned is the sense-tagging of content words in WordNet glosses in Senseval-3 (Litkowski 2004b).

While the focus of this discussion has been on how information in WordNet 2.0 might prove useful in the effort to induce semantic frames, it should be noted as well that results from the current research effort can be useful in turn to the future development of the topical classification of WordNet.

8.1.2 FrameNet

Some of the products of the FrameNet project, described in Section 2.3, could also contribute to an expansion of the current research effort. The first of these is the list of lexical items (of various parts of speech) that evoke each semantic frame analyzed in FrameNet. Few, if any, of the lists will include all relevant lexical items, but the items that constitute each frameset are assumed correct. Wherever a frame discovered inductively corresponds closely to a FrameNet frame, the addition of lexical items from FrameNet would benefit the extensional identification effort here. It should be noted that, although the FrameNet investigators had initially planned to map such lexical items to WordNet senses, they found too high a degree of mismatch between how WordNet distinguishes among word senses and their own work. It is not important for this research that lexical items borrowed from FrameNet be able to be put into one-to-one correspondence with WordNet senses. What is important is merely that there be enough information available for the words taken from FrameNet to identify one or more senses of the words that evoke the same semantic frame.

A second product from FrameNet that could prove useful to the current effort is its emerging hierarchical and compositional structuring of frames. At the same time, if the extensional and intensional identification of frames can be implemented sufficiently well in the continuation of this research, methods for automatically discovering such organizational structures should emerge.

As noted with respect to WordNet, it is entirely conceivable that the results of this research effort will feed back into FrameNet, as well as benefit from it. On the one hand,

where frames developed in FrameNet and SemFrame correspond with each other, SemFrame may identify additional lexical items that evoke the frame beyond those identified in FrameNet.⁷¹ On the other hand, a frame generated by SemFrame may have no corresponding frame in FrameNet. Its identification may help expand FrameNet's coverage.

8.1.3 Corpus Data

A specific goal of the current research effort has been to determine whether the data in available lexical resources are sufficiently rich to support the identification of semantic frames. While the answer is affirmative, it is only mildly so. Enhancement of this research effort with corpus data is likely to be needed to achieve results that are comprehensive at the general level. Such enhancement is also likely to be needed in frame identification on a more specific level. As rich a resource as WordNet 2.0 is and continues to become, it is still challenged by data sparseness in the endeavor to induce semantic frames.

The challenges faced by using corpus data for this undertaking are altogether different. Where WordNet recognizes word sense distinctions at many turns, corpus data are not disambiguated. Where definitions and glosses tend to refer to semantic arguments with semantic-type-like words, the semantic arguments in corpus data are often realized by

⁷¹To the extent that FrameNet and SemFrame frames agree, the direction of research undertaken here may help further the line of research reported in Gildea and Jurafsky (2002). In their work, given an input sentence and a frame, syntactic constituents are labeled with their semantic roles. It is likely that appropriate frames can be automatically identified using an approach like that described in Section 6.3.

more specific words. Where WordNet contains only a limited amount of data for any specific word sense, a corpus may contain a great deal of data about a specific word sense. Where WordNet contains approximately the same amount of information for all word senses, a corpus is likely to have widely disparate amounts of information for different word senses.

Corpus data can be used to address various specific aspects of the current research effort, namely, identifying verbs that are likely to evoke the same frame, identifying nouns associated with specific verbs, and identifying the semantic types of the arguments of classes of verbs. Although these are distinct aspects in the current research, they would tend to be handled interdependently in corpus data analysis, as seen below.

What characteristics of word behavior in open text can help identify verb senses that evoke the same frame? First, we expect verb senses that evoke a common semantic frame to occur within the same discourse segment significantly more often than would happen by chance. Second, we expect verb senses that evoke a common semantic frame to share at least some of the same argument types. Furthermore, within a discourse segment, some of the shared argument types may well be realized by identical lexical items or by lexical items that are semantically related to each other.

The first of these characterizations addresses the identification of the membership of a verb frameset as a task independent of the arguments of the verbs. If we assume the availability of a part-of-speech tagged corpus,⁷² then the computation of mutual

⁷²The American National Corpus (<http://americannationalcorpus.org/>) is such a corpus. The first part (10 million words) has just recently been released. The full corpus of 100 million words is slated to be available in fall 2005.

information scores for pairs of verbs co-occurring within a paragraph or within a small number of sentences could be used to reveal verb pairs that potentially evoke the same frame.

The second of these characterizations treats as interdependent operations tasks that are distinct in the current research. If we assume the availability of a parsed corpus, then something like verb/dependency-relationship/noun triples can be extracted from the corpus. These data can be analyzed, again using mutual information, to identify verbs and nouns that tend to co-occur with each other. Further analysis with respect to a lexical resource like WordNet could address the semantic classes of the arguments of specific verbs (Resnik 1998; Clark and Weir 2001) or of classes of verbs (Agirre and Martinez 2001, 2002).

8.2 Other Contexts

A question one might ask is: How easily can the SemFrame approach be migrated to other contexts (for example, to other languages or to specific subject domains)? Since the current approach relies fundamentally on the availability of lexical resources with a certain set of characteristics, the answer depends almost wholly on whether such lexical resources exist for the new context.

Minimally, the lexical resources should have the following three properties, each of which will be explored in turn:

- *The various semantic senses of lexical items should be distinguished in the resource.*
-

- *The lexical resource should encode and/or facilitate the recognition of some set of semantic relationships among its lexical entries.*
- *The lexical resource should facilitate discovery of the semantic types of a predicate's (semantic) arguments.*

We begin with the first of these properties: *The various semantic senses of lexical items should be distinguished in the resource.* Sense distinctions can be made at different levels of granularity. A lexical resource making only gross distinctions is likely to fare better than one than making relatively fine distinctions, since the senses of a word that evoke the same frame need to be collocated anyway. Thus, the minimum level of sense differentiation needed (and the optimal level desired) is to recognize distinct senses of a word when its different uses consistently evoke different semantic frames. It should follow that when a word consistently evokes the same semantic frame in all its uses, as would typically be the case in a specific subject domain, no differentiation of senses is needed (or desired).

The lexical resource should encode and/or facilitate the recognition of some set of semantic relationships among its lexical entries. Semantic relationships are the basis for the extensional identification of frames in SemFrame. In other words, semantic relationships are exploited in the identification of framesets, sets of words that evoke the same frame. For this purpose, the encoding of a large number of relationship types among word senses would be the optimal situation. Of the three facets of this desideratum—(1) encoding relationships vs. facilitating recognition of relationships, (2) a large number of relationship types vs. a smaller number of relationship types, (3) relationships among word

senses vs. relationships among words—the number and nature of the relationship types that are accessible through the resource is the most important. The semantic relationships among words that evoke the same frame are varied, and in many cases the only definable semantic relationship between two such words is expressly that they evoke the same frame. At the same time, such regular semantic relationships as synonymy, antonymy, and hyponymy play a large role in identifying words within a common frame. The larger the number of semantic relationships that can be employed in the discovery process, the more likely it is that framesets with satisfactory recall levels can be identified. The explicit encoding of those relationships among word *senses* helps to maintain high levels of precision simultaneously. It should be noted, however, that some of the data sets used in this research effort are based on other than explicitly encoded semantic relationships (e.g., LC.pairs, LI.pairs, WC.pairs). Furthermore, many of the relationships underlying the LS.pairs data set, a data source of high precision, link words and not word senses. Thus, the explicit encoding of semantic relationships is not necessarily absolutely required, nor do the relationships expressed in the resource have to be at the level of word senses, but may exist only at the level of words (assuming that sufficient data exist to do the necessary word sense disambiguation in a separate step).

The lexical resource should facilitate discovery of the semantic types of a predicate's (semantic) arguments. This desideratum includes two aspects. First, the resource should facilitate discovery of the semantic arguments of the predicate around which semantically related word senses are grouped. Second, the resource should facilitate the semantic typing of those arguments. Definitions have been shown to facilitate the identification of a

subset of the semantic arguments of the predicate, as well as to facilitate the semantic typing of those arguments. However, it is often not clear which semantic arguments have not been retrieved from definitional data, nor which of these arguments are not accorded appropriate semantic typing. Hand-crafted example sentences and corpus data are better sources for identifying the full range of semantic arguments than definitions, but generally are not as good for semantic typing, since arguments may be instantiated at very specific levels in natural language use. All of these resources are profitably supplemented by the generation of morphologically-related nouns. There is no reason to suppose that all of the semantic predicate's arguments have corresponding morphologically-related nouns, but where this is the case, this data source—which exists independently of specific lexical resources—is often key to the semantic typing of those arguments. It should be noted that the efficacy of this data source depends on the nature of the morphological processes commonly operative within a specific language.

Those languages for which WordNets⁷³ have been or are being developed—Basque, Bulgarian, Catalan, Czech, Dutch, Estonian, French, German, Greek, Italian, Portuguese, Romanian, Serbian, Spanish, Swedish, Turkish—would be the most natural languages to migrate the SemFrame approach to.⁷⁴ Most non-English WordNets are modeled after the Princeton-developed WordNet and map their concepts to nodes in WordNet 1.5. Such

⁷³Such development is fostered and tracked by the Global WordNet Association (<http://www.globalwordnet.org>).

⁷⁴Farreres, Rigau, and Rodríguez (1998) discuss the development of multilingual WordNets based on (English) WordNet and existing lexical resources.

lexical resources will recognize word senses, explicitly encode some number of semantic relationships among word senses, and facilitate to some degree the identification of the (semantic) types of semantic arguments of the predicate corresponding to generated frames. Thus languages with WordNets like the proto-WordNet for American English would be optimal candidates for application of the SemFrame approach to semantic frame discovery.

As the languages with WordNets tend to lexicalize the same overall set of concepts, we would expect that the set of frames recognized across them should correspond closely on the conceptual level.⁷⁵ (Conversely, where convergence around a conceptual frame exists, we can claim the validity of the frame with greater assurance.) The framesets from multiple languages that are associated with a frame that is shared across languages on the conceptual level need not have corresponding lexical items.

Of the future work prospects discussed in this chapter, extension of this research effort to specific subject domains may be the furthest removed from current capabilities. While some portion of the general semantic frame inventory developed for the general context would be applicable as well to a specific subject domain, only the tip of the subject domain iceberg is likely to be included. Thus many frames specific to the subject domain would

⁷⁵The equivalence of semantic frames across languages is assumed, for instance, in the work of Riloff, Schafer, and Yarowsky (2002) to induce information extraction systems by cross-language projection. On a more concrete level, initial data from Erk, Kowalski, Padó, and Pinkal (2003) indicate that semantic frames developed for English “can be reused for German, with minor changes and extensions.”

need to be identified, including both the words that evoke the same frame and the internal participant structure of the frames.

This further development effort would need access to considerable information about the sublanguage of the subject domain. This is to be found in two kinds of sources. One such source is manually developed lexical resources, as characterized above. Here lack of available subject-specific machine-readable dictionaries is likely to present an obstacle. Even where such tools exist, they may present words and definitions only rather than also encode semantic relationships. Other knowledge-based resources, for example, a thesaurus or an ontology, may also be useful for revealing the conceptual organization of the subject domain. The issue here will be whether the tool also sets forth the language of the domain and how it relates to the concepts of the domain. The Unified Medical Language System® (UMLS®; <http://www.nlm.nih.gov/research/umls/>) of the National Library of Medicine is probably one of the few tools that approximates these goals.

The other source of information on the sublanguage of a specific subject domain is a corpus of subject-specific texts. Consistent with the use of corpus data for the development of semantic frames within a general inventory, analysis of subject specific corpora should likewise permit the identification of at least some of the semantically related verbs and nouns of the domain (Habert and Fabre 1999). This venture will also require the identification of subject-specific texts, but in the current environment, this will generally not be problematic.

It is not immediately clear whether the analysis of corpus data would more readily contribute to the expansion of the general semantic frame inventory or to the development of a subject specific semantic frame inventory. On the one hand, subject

specificity should promote the identification of semantic frames more easily than the lack thereof. On the other hand, the use of corpus data will typically be better balanced by the availability of lexical tools in the general context than in the specific subject context; as indicated above, medicine constitutes a rare exception.

8.3 Other Applications

Word sense disambiguation, information retrieval, information extraction, question answering, text summarization, and machine translation are representative of knowledge-based activities that could benefit from semantic frame information. In some cases, task performance could be improved by having access to framesets. In other cases, access to automatically generated semantic frame structures would benefit task performance.

8.3.1 Word Sense Disambiguation (WSD)

“You shall know a word by the company it keeps” (Firth 1957). This observation, oft repeated in connection with word sense disambiguation efforts, reminds us that knowing the other words that a specific word co-occurs with often allows us to determine the intended sense of that word. Framesets constitute a particular way of structuring lexical neighborhoods. To know what other words evoke the same semantic frame(s) may be as good a means of determining the intended sense of a word as to know its verbal head, if it is a noun, or to know the identities or the semantic classes of its arguments, if it is a verb.

The SemFrame1 and SemFrame2 enhancements of TextTiling include an implicit word sense disambiguation stage. Although text words have been WordNet-synset-disambiguated in SEMCOR, these sense indications were disregarded in the text segmentation evaluation reported in Chapter 6. A given word may be associated with various semantic frames; through hierarchical and compositional relationships between semantic frames, even a specific word sense may be associated with multiple frames. Section 6.3 explains how frameset information about sets of words that evoke the same frame was used to determine, for each occurrence of a noun or verb represented in the results of stages 1 and 2, which was the most likely semantic frame that it evoked. Since specific senses of words evoke frames, the semantic frame disambiguation implemented there was also identifying a specific sense, or a closely related family of senses, for each noun or verb. The same process could be applied to any text that has been pre-processed by part-of-speech tagging.

Senseval-3 will include a task to disambiguate the senses of words in WordNet 2.0 glosses (Litkowski 2004b). Here we lack the rolling text environment required for WSD with semantic frames just described. Instead characteristics specific to definitions and to WordNet 2.0 will need to be taken into account, specifically taking into account WordNet's network of semantic relationships (Harabagiu, Miller, and Moldovan 1999). Because the data available within any one synset is relatively small, although rich, being able to enhance that data with semantic frame information may hold the key to accurate disambiguation of content words within the WordNet glosses.

8.3.2 Information Retrieval (IR)

Information retrieval seeks to identify documents that are relevant to an information need. All too often information retrieval is treated as a problem of predicting which words or phrases would be included in a relevant text. But what makes a text relevant is not the specific words it uses to convey its conceptual content, but the conceptual content itself. Given the paraphrase problem, that content can be expressed in many ways, which makes the problem of predicting all the ways it might be expressed difficult.⁷⁶

The challenge that the paraphrase problem poses for information retrieval has been addressed in two quite different ways. In the one approach, often referred to as a bag-of-words approach, the burden is placed on the end user or searcher to predict the words or phrases that would occur in a relevant document; at the grossest level, retrieval depends on the same linguistic expressions being found in a relevant document as are in the search statement. In the other approach, a controlled vocabulary—playing the same role as an interlingua—mediates between concepts and linguistic expressions; retrieval

⁷⁶One might counter that information retrieval need not address the full force of the paraphrase problem. After all, few users really want to see all the material relevant to their need. Instead, they often place a higher value on precision than on recall, desiring to see only relevant documents. If this is so, it might not seem necessary to predict all the ways that relevant conceptual content might be expressed, but only some of the ways it might be expressed. While the observation that users place a higher premium on precision than on recall is often the case, that does not mean the user has no preference as to which of the relevant documents are retrieved: The user wants to see the best of the relevant documents. But the expression of relevant conceptual content in the best of the relevant documents may vary as much as in all the relevant documents. It is better to conceive of retrieval as a two-phase process in which as many relevant documents as possible are first identified, followed by a filtering process in which non-relevant documents are winnowed out. In the context of this model, information retrieval does face the full brunt of the paraphrase problem.

depends on both indexer and searcher representing information needs with the same descriptor (a.k.a. index term).

On the surface, the bag-of-words approach does not seem to address the paraphrase problem at all; its power instead relies on being able to “see” all the words in a document. The breadth of this approach permits retrieval by more access points than are available under the controlled vocabulary approach. Thus, the bag-of-words approach is able to retrieve documents that are relevant even when only small segments of the document address the user need in the user’s terms; such documents would typically be missed by the controlled vocabulary approach, where the vastly smaller number of access points assigned precludes representing many document segments in the indexing.⁷⁷

However, the bag-of-words approach can be enhanced in ways that do directly address the paraphrase problem. In a typical enhancement, search statements are processed through a stemmer. The search statement is thus modified to account for paraphrases related through morphological derivation (stemmers do not, however, cover the full gamut of morphological variation) or through a comparative vs. superlative alternation. Another possible enhancement is the automatic expansion of the search statement to include synonyms of terms supplied by the user, as identified in a thesaurus. This permits the bag-of-words approach to account for paraphrases of a lexical

⁷⁷The limitation on access points is not inherent to the controlled vocabulary approach *per se*. The reasons for such limitations are instead practical: (1) Typically it is human indexers who assign terms from controlled vocabularies. Assigning more terms than is now done would make this kind of indexing prohibitively expensive. (2) In the past and even to some extent in the present, systems built on controlled vocabularies have been paper-based. Again, assigning more terms drives up the cost of the system.

synonymy nature. A third type of enhancement is based on relevance feedback. After the user identifies one or more especially relevant documents, the search statement is automatically modified to model relevant documents better. Such documents are likely to include paraphrases of various types. Through relevance feedback the full array of paraphrase types may (in theory) be taken into account.

Framesets from SemFrame could be incorporated into the bag-of-words approach to information retrieval by replacing thesaural expansion of the search statement, in either non-interactive or interactive mode. In interactive mode, if the search statement contains ambiguous terms, various framesets would be displayed to the user, who could select which framesets most closely approximate the concepts of interest. In either mode, the search statement would then be modified to include at least those frameset terms with the highest potential for discriminating between relevant and non-relevant documents. In interactive mode, the user could specify terms from the frameset to include in the search statement, regardless of tf-idf weights.

But regardless of the number of enhancements made to it, the basic bag-of-words approach will never be able to surmount a barrier inherent to the approach. The allegiance of this approach to at most shallow processing of language is a restriction that will constrain it from dealing with semantic information in any meaningful way.

It is sometimes claimed, however, that information retrieval does not need deep natural language processing (NLP)—that empirical studies have shown that approaches that treat documents as bags-of-words, e.g., keyword searching, are as effective as more sophisticated retrieval techniques. What studies actually show is that different retrieval

techniques complement each other, implying that no known technique is without room for improvement. Indeed, in light of typical precision and recall ratios, most techniques leave considerable room for improvement. Strzalkowski (1999, p. xiv) notes that the “limitations [of standard IR approaches] have led to recurring interest in NLP approaches, as the latter [have become] more efficient, more robust, and can handle large amounts of data.” He hints (p. xvii) that we should “[aim] at a semantically-motivated concept-based representation of text,” for which even deeper NLP will be needed instead of the more shallow syntax-oriented approaches that have been common in the past. Such a development would make information retrieval a true natural language understanding task.

The relevance of semantic frames to “a semantically-motivated concept-based representation of text” is obvious. If the semantic content of both queries and documents were translated into semantic frame structures (say, using information extraction techniques), searches could achieve higher recall than if the end user had to predict all the ways in which relevant predicate types might be conveyed. At the same time, query results could be made more precise by permitting specification of a predicate’s arguments (Green 1989).⁷⁸

⁷⁸Semantic frames have also been used to remind indexers of subject aspects that need to be considered in creating a document surrogate for searching (Humphrey and Miller 1987).

8.3.3 Information Extraction and Related Tasks

Information extraction, another natural language understanding application, seeks to draw specific types of information out of texts. For instance, texts about terrorist activities might be mined to ascertain the identity of the terrorists, their targets, the nature of the terrorist activities they have engaged in, the number of victims of those activities, and so forth. A major challenge faced by information extraction is that any one of those pieces of information may be expressed in numerous ways, thanks to the paraphrase problem.

Information extraction tasks depend critically on the pre-identification of data structures known as “templates.” These templates correspond very closely to the semantic frame representations under discussion. Information can be drawn from texts, based on a knowledge of the underlying structures for specific domains; values for specific slots in a template may be extracted, for example, on the basis of specific phrases and/or syntactic patterns. Instantiated templates or frames output by information extraction techniques can, in turn, be used to support such other tasks as information retrieval, question answering, or text summarization.

For example, the aim of text summarization is to produce document summaries that convey the most important information of the original text in a considerably more compact form. “Abstractive” summarization (that is, summarization that goes beyond mere extraction; see Radev, Hovy, and McKeown 2002) prototypically includes both natural language understanding and natural language generation, in varying proportions. The natural language understanding phase of text summarization may simply be an

information extraction implementation, which contends with the paraphrase problem in needing to recognize a variety of linguistic expressions as having similar meanings; this is equally true of single-document summarization and multi-document summarization. The natural language generation phase must choose among alternative modes of expression, thus again facing the paraphrase problem head-on.

While the work in information extraction is concerned with mapping from texts to templates, not with constructing templates or frames in the first place, clearly the templates must come from somewhere. The semantic frames developed here are one possible source.

8.3.4 Machine Translation

In one sense, the paraphrase problem is what machine translation is all about. The expression of shared semantic content in two different languages is essentially the same phenomenon as the expression of shared semantic content in multiple ways in a single language. In translating from one language to another, it is sometimes the case that there is an exact or nearly exact equivalent in the target language for a word or phrase in the source language. This case parallels lexical synonymy within a single language. In other circumstances, divergences across languages will typically result in translations that parallel other paraphrase processes found within a language (e.g., different argument realizations, overlapping meanings, syntactic variation, noun-noun phrases, and head switching).

Rather than try to translate directly from one language to another, it is typically considered more effective to capture the semantic content of the source language text in an interlingual representation and then choose the lexical items in the target language that best express that semantic content (e.g., Hutchins 1986; Dorr 1993). Thus machine translation, like text summarization, involves both natural language understanding and natural language generation—and faces the paraphrase problem in both stages.

Semantic frames are suitable candidates for this interlingual representation. Moreover, the focus on English in SemFrame mirrors the tack being taken in an NSF-funded (ITR) collaborative research project⁷⁹ to develop a semantic representation for multilingual texts. The definition of the interlingua will be based on annotating bilingual text corpora in which one of the parallel texts is always in English.

Having briefly considered the use of semantic frames to face the paraphrase problem in several natural language understanding tasks, we turn to consider their use with respect to a natural language generation task. In order to make appropriate lexical choices in the generation stage of machine translation, the system needs to know (1) which verbs have the potential of communicating the intended semantic content and (2) what the properties of each verb are vis-à-vis the semantic predication. For example, does the verb imply a semantic argument (as *fly* implies the means of Conveyance is a plane)? If so, and if its

⁷⁹Collaborative Research: Interlingual Annotation of Multilingual Text Corpora <http://crl.nmsu.edu/NewResearch/collab_research.html> will bring together researchers from New Mexico State University (CRL), University of Southern California (ISI), University of Maryland (UMIACS), Carnegie Mellon University, Columbia University, and the MITRE Corporation.

implied argument value does not match the value given in the semantic representation, the verb should not be selected. Does the verb not permit specification of a semantic argument (as *cost* in *The blouse cost Mary a days wages* does not permit specification of a Seller in a COMMERCIAL TRANSACTION predication)? If so, and if the semantic representation includes a value for that argument, the verb would generally not be a good choice. Conversely, verbs that permit specification of all values in the semantic representation should be among the candidates for selection; preference should be given to those verbs that meet this requirement and leave the fewest syntactic argument positions unfilled.

If the interlingual representation is given in terms of semantic/conceptual structures, translation can focus on the correspondence of the lexical choices within individual languages to this underlying conceptual representation, thus avoiding the toil and effort that would be required if machine translation were approached relative to the vastly larger number of pairs of source and target languages and their correspondences with each other.⁸⁰ The kinds of semantic frames generated within task 2 of this research are candidates for use as the conceptual structures for a machine translation interlingua.

⁸⁰The benefits of a single, meaning-based interlingual representation is all the more striking in, for example, cross-language information retrieval tasks, where more than two languages may be of simultaneous concern.

8.4 Summary

At the same time that SemFrame has achieved a modicum of success within the scope of its basic aspirations—to induce semantic frames from available lexical resources well enough to improve performance in a knowledge-intensive task—reasons abound to infer that the results presented here can be improved upon. This chapter has discussed the potential usefulness especially of the recently released version of WordNet (2.0) to the semantic frame venture. The mining of text corpora for semantic relationships is also recommended, for the induction both of general semantic frames and of subject specific semantic frames.

The broad applicability of semantic frame information to knowledge-based tasks (information retrieval, information extraction, text summarization, machine translation) was also reviewed. In order to achieve the full benefit of semantic frames in such applications, additional work at the syntax-semantics interface (such as that represented by FrameNet and Gildea and Jurafsky 2002) will be required. However, significant uses can be foreseen for verb and noun sense framesets and semantic frame participant structures with all of these applications, even without having the more detailed data needed for more full and accurate natural language understanding and generation.

References

- Agirre, Eneko and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning* (ConLL-2001), 15-22.
- Agirre, Eneko and David Martinez. 2002. Integrating selectional preferences in WordNet. In *Proceedings of the First International WordNet Conference*. Available: <<http://arXiv.org/pdf/cs.CL/0204027>> [2004 February 5].
- Agirre, Eneko and German Rigau. 1995. A proposal for word sense disambiguation using conceptual distance. In *1st International Conference on Recent Advances in NLP*. Available: <<http://xxx.lanl.gov/abs/cmp-lg/9510003>> [2004 February 5].
- Alshawī, Hiyan, Bran Boguraev, and David Carter. 1989. Placing the dictionary on-line. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, 41-63. London and New York: Longman; New York: Wiley.
- Alshawī, Hiyan. 1989. Analysing the dictionary definitions. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, 153-169. London and New York: Longman; New York: Wiley.
- Atkins, B. T. Sue. 1993. Building a lexicon: The contribution of lexicography. In M. Bates and R. M. Weischedel (Eds.), *Challenges in Natural Language Processing*, 37-75. Cambridge: Cambridge University Press.
- Atkins, Beryl T., Judy Kegl, and Beth Levin. 1986. *Explicit and Implicit Information in Dictionaries* (Lexicon Project Working Papers 12). Cambridge, MA: Center for Cognitive Science, MIT.
- Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, 86-90.
- Barsalou, Lawrence W. 1992. Frames, concepts, and conceptual fields. In E. Kittay and A. Lehrer (Eds.), *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, 21-74. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT/NAACL 2003*, 16-23.
- Boguraev, Bran and Ted Briscoe. 1989. Introduction. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, 1-40. London: Longman.

- Brun, Caroline and Caroline Hagège. 2003. Normalization and paraphrasing using symbolic methods. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 41-48.
- Budanitsky, Alexander and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application -oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, 29-34.
- Busa, Federica, Danièle Dubois, Christiane Fellbaum, Patrick Saint-Dizier, and Evelyn Viegas. 1999. A comparison of different lexical semantics approaches for transfer verbs with a particular emphasis on Buy/Sell. In P. Saint-Dizier (Ed.), *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, 53-91. Dordrecht: Kluwer Academic.
- Clark, Stephen and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 187-206.
- Cowie, A. P. 1999. *English Dictionaries for Foreign Learners: A History*. Oxford: Clarendon Press.
- Croft, William. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press.
- Dorr, Bonnie J. 1993. *Machine Translation: A View from the Lexicon*. Cambridge, MA: The MIT Press.
- EAGLES Lexicon Interest Group. 1998. *EAGLES Preliminary Recommendations on Semantic Encoding: Interim Report*. Available: <<http://www.ilc.cnr.it/EAGLES96/rep2/rep2.html>> [2004 February 5].
- Edmonds, Philip and Graeme Hirst. 2002. Near synonymy and lexical choice. *Computational Linguistics* 28/2: 105-144.
- Erk, Katrin, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 537-544.
- Farreres, Xavier, German Rigau, and Horacio Rodríguez. 1998. Using WordNet for building WordNets. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, 65-72.

- Fellbaum, Christiane. 1998a. Introduction. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 1-17. Cambridge, MA and London: The MIT Press.
- Fellbaum, Christiane. 1998b. A semantic network of English verbs. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 69-104. Cambridge, MA and London: The MIT Press.
- Fillmore, Charles J. 1976. Frame semantics and the nature of language. In S. R. Harnad, H. D. Steklis, and J. Lancaster (Eds.), *Origins and Evolution of Language and Speech*, 20-32. (Annals of the New York Academy of Sciences ; v. 280.)
- Fillmore, Charles J. 1977. The need for a frame semantics within linguistics. *Statistical Methods in Linguistics*, 5-29.
- Fillmore, Charles J. 1982. Frame semantics. In *Linguistics in the Morning Calm*, 111-137. Seoul: Hanshin.
- Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* VI/2, 114-130.
- Fillmore, Charles J. and B. T. S. Atkins. 1992. Towards a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer and E. F. Kittay (Eds.), *Frames, Fields, and Contrasts*, 75-102. Hillsdale, NJ: Erlbaum.
- Fillmore, Charles J. and Collin F. Baker. 2001. Frame semantics for text understanding. In *Workshop on WordNet and Other Lexical Resources*.
- Fillmore, Charles J., Charles Wooters, and Collin Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*.
- Firth, J. R. 1957. *A synopsis of linguistic theory 1930-1955*. London: Longman.
- FrameNet. 2004. Available: <<http://www.icsi.berkeley.edu/~framenet/>> [2004 February 5].
- Francis, W. N. and H. Kucera. 1979. Brown Corpus Manual [Sic]: Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, RI: Department of Linguistics, Brown University. Available: <<http://www.hit.uib.no/icame/brown/bcm.html>> [2004 February 5].

- Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Speech and Natural Language Workshop*, 233-237. San Francisco: Morgan Kaufmann.
- Gildea, Daniel and Daniel Jurasky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28/3: 245-288.
- Green, Rebecca. 1989. *The Expression of Syntagmatic Relationships in Frame-Based Indexing*. Ph.D. dissertation, University of Maryland, College Park.
- Gruber, Jeffrey S. 1976. *Lexical Structures in Syntax and Semantics*. Amsterdam: North-Holland.
- Habash, Nizar and Bonnie Dorr. 2003. A categorial variation database for English. In *Proceedings of North American Association for Computational Linguistics*, 96-102.
- Habert, B. and C. Fabre. 1999. Elementary dependency trees for identifying corpus-specific semantic classes. *Computers and the Humanities* 33: 207-219.
- Harabagiu, Sandra M., George Miller, and Dan Moldovan. 1999. WordNet 2 - A morphologically and semantically enhanced resource. In *SIGLEX99: Standardizing Lexical Resources: Proceedings of a Workshop Sponsored by the Special Interest Group on the Lexicon of the Association for Computational Linguistics and the National Science Foundation*, 1-8.
- Harabagiu, Sandra M. and Dan I. Moldovan. 1998. Knowledge processing on an extended WordNet. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 379-405. Cambridge, MA and London: The MIT Press.
- Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23/1: 33-64.
- Hirst, Graeme. 2003. Paraphrasing paraphrased. Keynote address for *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, ACL 2003. Available: <<http://nlp.nagaokaut.ac.jp/IWP2003/pdf/Hirst-slides.pdf>> [2004 February 5].
- Humphrey, Susanne M. and Nancy E. Miller. 1987. Knowledge-based indexing of the medical literature: The Indexing Aid Project. *Journal of the American Society for Information Science* 38: 184-196.
- Hutchins, W. J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood; New York: Halsted Press.

- Ibrahim, Ali, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 57-64.
- Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 9-16.
- Jackendoff, Ray. 1987. The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18/3: 369-411.
- Johnson, Christopher R., Charles J. Fillmore, Miriam R. L. Petruck, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer, and Esther J. Wood. 2002. *FrameNet: Theory and Practice*, version 1.0. Available: <<http://www.icsi.berkeley.edu/~framenet/book/book.html>> [2004 February 5].
- Jurafsky, Daniel and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.
- Kozlowski, Raymond, Kathleen F. McCoy, and K. Vijay-Shanker. 2003. Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 1-8.
- Krovetz, Robert. 1992. Sense-linking in a machine readable dictionary. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 330-332.
- Lakoff, George. 1993. The syntax of metaphorical roles. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*, 27-36. Dordrecht: Kluwer.
- Landau, Sidney I. 1989. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.
- Levin, Beth. 1985. Lexical semantics in review: An introduction. In B. Levin (Ed.), *Lexical Semantics in Review (Lexicon Project Working Papers 1)*, 1-62. Cambridge, MA: Center for Cognitive Science, MIT.
- Levin, Beth. 1993a. Building a lexicon: The contribution of linguistics. In M. Bates and R. M. Weischedel (Eds.), *Challenges in Natural Language Processing*, 76-98. Cambridge: Cambridge University Press.

- Levin, Beth. 1993b. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: University of Chicago Press.
- Levin, Beth. 1995. Approaches to lexical semantic representation. In D. E. Walker, A. Zampolli, and N. Calzolari (Eds.), *Automating the Lexicon: Research and Practice in a Multilingual Environment*, 53-91. Oxford: Oxford University Press.
- Levin, Beth and Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. In B. Levin and S. Pinker (Eds.), *Special Issue on Lexical and Conceptual Semantics, Cognition* 41: 123-151.
- Lewis, C. I. 1946. *An Analysis of Knowledge and Valuation*. La Salle, IL: Open Court Pub. Co.
- Lewis, C. I. 1952. The modes of meaning. In Leonard Linsky (Ed.), *Semantics and the Philosophy of Language: A collection of readings*, 50-63. Urbana: University of Illinois Press.
- Lin, Dekang. 2001. LaTaT: Language and text analysis tools. In *Proceedings of the First International Conference on Human Language Technology Research*, 222-227.
- Lin, Dekang and Patrick Pantel. 2001a. Discovery of inference rules for question-answering. *Natural Language Engineering*, 7: 343-360.
- Lin, Dekang and Patrick Pantel. 2001b. Induction of semantic classes from natural language text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 317-322.
- Litkowski, Ken. 2004a. Senseval-3 task: Automatic labeling of semantic roles. Available: <<http://www.clres.com/SensSemRoles.html>> [2004 February 5].
- Litkowski, Ken. 2004b. Senseval-3 task: Word-sense disambiguation of WordNet glosses. Available: <<http://www.clres.com/SensWNDisamb.html>> [2004 February 5].
- McKeown, Kathleen R., Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference of the American Association for Artificial Intelligence (AAAI-1999)*, 453-460.
- Miller, George A. 1998. Nouns in WordNet. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 23-67. Cambridge, MA and London: The MIT Press.

- Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 71/1: 21-48.
- Nolan, Rita. 1970. *Foundations for an Adequate Criterion for Paraphrase*. The Hague: Mouton.
- Norman, Donald A. and David E. Rumelhart. 1975. The active structural network. In D. E. Rumelhart, Donald A. Norman, and the LNR Research group (Eds.), *Explorations in Cognition*, 35-64. San Francisco: W. H. Freeman.
- Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 613-619.
- Porter, M. F. 1980. An algorithm for suffix stripping. *Program* 14/3: 130-137.
- Procter, Paul (Ed.). 1978. *Longman Dictionary of Contemporary English*. Longman Group Ltd., Essex, UK.
- Radev, Dragomir R., Eduard Hovy, and Kathleen R. McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics* 28/4: 399-408.
- Radev, Dragomir R. and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics* 24/3: 469-500.
- Rasmussen, Edie. 1992. Clustering algorithms. In W. B. Frakes, B. and R. Baeza-Yates, (Eds.), *Information Retrieval: Data Structures and Algorithms*, 419-442. Englewood Cliffs, N.J.: Prentice Hall.
- Ravin, Yael. 1990. *Lexical Semantics without Thematic Roles*. Oxford: Oxford University Press.
- Resnik, Philip. 1998. WordNet and class-based probabilities. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, 239-263. Cambridge, MA: The MIT Press.
- Riemsdijk, Henk van and Edwin Williams. 1986. *Introduction to the Theory of Grammar*. Cambridge, MA: The MIT Press.
- Riloff, Ellen and Jones, Rosie. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, 474-479.

- Riloff, Ellen, Charles Schafer, and David Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. *Proceedings of the 19th International Conference on Computational Linguistics*.
- Rinaldi, Fabio, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. 2003. Exploiting paraphrases in a question answering system. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 25-32.
- Saint-Dizier, Patrick. 1999. An introduction to the lexical semantics of predicative forms: A computational perspective. In P. Saint-Dizier (Ed.), *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, 1-52. Dordrecht: Kluwer Academic.
- Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading, Mass.: Addison-Wesley.
- Somers, H. L. 1987. *Valency and Case in Computational Linguistics*. Edinburgh: Edinburgh University Press.
- Strzalkowski, Tomek (Ed.). 1999. *Natural Language Information Retrieval*. Dordrecht: Kluwer.
- Tomuro, Noriko. 2003. Interrogative reformulation patterns and acquisition of question paraphrases. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 33-40.
- Voorhees, Ellen. 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management* 22/6: 465-476.
- Vossen, Piek, Willem Meijs, and Marianne den Broeder. 1989. Meaning and structure in dictionary definitions. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*, 171-192. London: Longman.
- Wilks, Yorick, Dan Fass, Cheng-Ming Guo, James McDonald, Tony Plate, and Brian Slator. 1993. Providing machine tractable dictionary tools. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*, 341-401. Dordrecht, Boston, and London: Kluwer.
- Willett, Peter. 1988. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management* 24/5: 577-597.
- WordNet. Available: <<http://www.cogsci.princeton.edu/~wn>> [2004 February 5].

Zukerman, Ingrid, Sarah George, and Yingying Wen. 2003. Lexical paraphrasing for document retrieval and node identification. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003)*, ACL 2003, 94-101.