Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions

Thomas Paul O'Hara

August 31, 2005

Abstract

This thesis discusses the automatic acquisition of conceptual distinctions using empirical methods, with an emphasis on semantic relations. The goal is to improve semantic lexicons for computational linguistics, but the work can be applied to general-purpose knowledge bases as well.

The approach is to analyze dictionary definitions to extract the distinguishing information (i.e., *differentia*) for concepts relative to their sibling concepts. A two-step process is employed to decouple the definition parsing from the disambiguation of the syntactic relations into the underlying semantic ones. Previous approaches tend to combine these steps through pattern matching geared to particular types of relations. In contrast, here a broad-coverage parser is first used to determine the syntactic relationships, and then statistical classification techniques are used to disambiguate the relationships into their underlying semantics.

There are several contributions of this thesis. First, it introduces an empirical methodology for the extraction and disambiguation of semantic relations from dictionary definitions. Second, it introduces a statistical representation for these semantic relations using Bayesian networks, which are popular in artificial intelligence for representing probabilistic dependencies. Third, it shows how improvements in word-sense disambiguation can be achieved by augmenting a standard statistical classifier approach with a probabilistic spreading-activation system using the semantic information extracted using this process.

Contents

CHAPTER 1
INTRODUCTION

Words are the basic unit of language for conveying meaning (Miller, 1996). Basic word knowledge involves the ability to determine entities that a particular word might refer to (e.g., the class of objects that can be labeled by the word). This aspect is referred to as the *extension* or *denotation* of the word (Lyons, 1977). For example, the word 'dog' denotes members of *Canis familiaris*, whereas the word 'cat' denotes members of *Felis domesticus*. Word knowledge also involves being able to identify salient aspects associated with the underlying entities, especially those distinguishing similar concepts. This aspect is part of the *intension* or *connotation* of the word, which also covers a variety of pragmatic aspects. For example, knowledge about 'dog' and 'cat' includes the recognition that dogs and cats are typical pets for humans and also that dogs make a harsh sound compared to a cat's soft sound. Strictly speaking, this saliency aspect is more concerned with conceptual knowledge rather than word knowledge. Nonetheless, lexicons for natural language processing should include this information as well as basic denotations.

This research addresses how distinguishing properties of concepts underlying word meaning can be acquired from dictionary definitions; that is, the focus is on the automatic acquisition of conceptual distinctions. In particular, this thesis aims to improve semantic lexicons for natural language processing by automatically extracting information from conventional English language definitions. The motivation for the work is that broad-coverage lexicons often do not provide sufficient information to differentiate sibling concepts. Consequently, words mapping to such undifferentiated sibling concepts are effectively treated as synonyms. For instance, WordNet's (Miller, 1990) representations for the concepts *Beagle* and *Wolfhound* are semantically equivalent (i.e., both specializations of *Hound*), although they should be quite distinct, especially with respect to information on typical size.

In the short term, these additional properties for use in word meanings can help word-sense disambiguation,[1] as shown later, by allowing for the determination of more interconnections among word senses. More ambitiously, extensions to the research could be used to help achieve the long-term goal of deep understanding, as part of that involves analyzing definition-like descriptions contained in text. A more intermediate extension would be to improve web searching. Current web search engines do not exploit word meanings when retrieving pages relevant to a particular query and instead mainly rely upon word

[1] The glossary at the end explains some of the technical terms used here.

co-occurrences. Although effective when the words are specific, this lack of understanding can lead to extraneous results that users must sift through. As with word-sense disambiguation, interconnections among the concepts underlying words can be exploited to filter out unlikely cases.

The acquisition of distinguishing information has been addressed by various approaches in the past. However, most of the previous work has relied upon manually derived extraction rules. This has limited the coverage of the associated system as well as introduced a bottleneck in the adaptation to new types of information. Previous approaches also tended to be specific to particular dictionaries, for example, by taking advantage of the lexicographic conventions used by a particular dictionary publisher.

There are several contributions of this thesis. First, it introduces an empirical methodology for the extraction and disambiguation of semantic relations from dictionary definitions. Second, it introduces a statistical representation for these semantic relations using Bayesian networks (Pearl, 1988), which are popular in artificial intelligence for representing probabilistic dependencies. Third, it shows how improvements in word-sense disambiguation can be achieved by augmenting a standard statistical classifier approach with a probabilistic spreading-activation system using the semantic information extracted using this process.

The rest of this chapter is organized as follows. Section 1.1 presents a high-level overview of the research presented in later chapters. Section 1.2 follows with motivation for this work both from within natural language processing (i.e., computational linguistics), as well as from other disciplines, namely psychology and lexicography. Section 1.3 then presents more details on the contributions of this thesis. Lastly, Section 1.4 outlines the rest of the thesis.

1.1 Overview and Example

High-quality lexicons are critical for natural language processing applications. Words are generally defined in terms of other words via *lexical relations*. Computational lexicons are quite effective at specifying the denotative aspect of the meaning of a given word, in particular through type specifications and generalization relations. However, they are not good at conveying distinguishing relationships, such as for words of the same type. For instance, even though WordNet (Miller, 1990) recently introduced over 6,000 domain category and location assertions in Version 2.0, about 38% of the concepts (i.e., synsets) for nouns are still not explicitly distinguished from sibling concepts. In effect, this leads to over 25,000 extraneous synonyms. This problem also exists to a lesser extent with traditional knowledge bases, such as Cyc (Lenat, 1995), of which

computational lexicons are a special case. The main reason for this lack of precision is that encoding differentiating assertions is more time-consuming than adding categorical assertions.

Dictionaries are a prime source of differentiating relations. In fact, dictionaries have evolved from simple word lists into encyclopedic reference works (Landau, 2001), making them repositories of important distinguishing characteristics for words similar in meaning. Several reasons account for the differentiating aspect of definitions. One is that dictionaries are often perceived as authorities on language. Another reason for this is that publishers usually have tight space constraints.

This thesis presents an approach for extracting differentiating information (*differentia*) from dictionary definitions. This information encompasses attributes such as *has-size* as well as functional relations such as *used-for*. Unlike previous approaches, the thesis emphasizes empirical methods, providing for more robust and adaptable extraction. Earlier extraction approaches have relied predominantly on manually derived rules for this process. A drawback is that the inclusion of more relations necessitates additional rule development. Instead, an empirical approach is taken to address this problem by exploiting existing annotations on relation usage, in particular, from the Penn Treebank (Marcus et al., 1994) and Berkeley's FrameNet (Fillmore et al., 2001). This in effect replaces the manual rule construction of knowledge-based approaches with manual text annotation. Thus, there still is a bottleneck before the acquisition of new types of relations can be acquired. However, the use of annotations allows for more flexibility because less technical training is required to prepare them. It also might be possible to approximate relation annotations by paraphrasing assertions from knowledge bases (e.g., Cyc). As discussed later in Section 4.3.4.1, one approach is to use corpora to extract text that most likely refers to the same information as that contained in the assertions (e.g., "dryer is for drying" given the assertion ⟨drying, *is-function-of*, dryer⟩).

Figure 1.1 shows the overall steps involved in enhancing a computational lexicon with differentiating information from dictionary definitions. The process requires three input resources. The *Basic Lexicon* is assumed to contain mappings from words to concepts, along with hierarchical relations for the concepts. WordNet is used here, but other semantic lexicons could be used (e.g., the one in Cyc). The *Dictionary* is an English language dictionary. For simplicity, the definitions from WordNet are used here. The *Corpus* consists of annotations on word-sense distinctions and relation usage in English text. This currently uses the word-sense annotations from Extended WordNet (Harabagiu et al., 1999) and the relation annotations from Treebank and FrameNet.

There are two main steps in the extraction process. During *Differentia Extraction*, the definition text is parsed, and the output converted into a list of

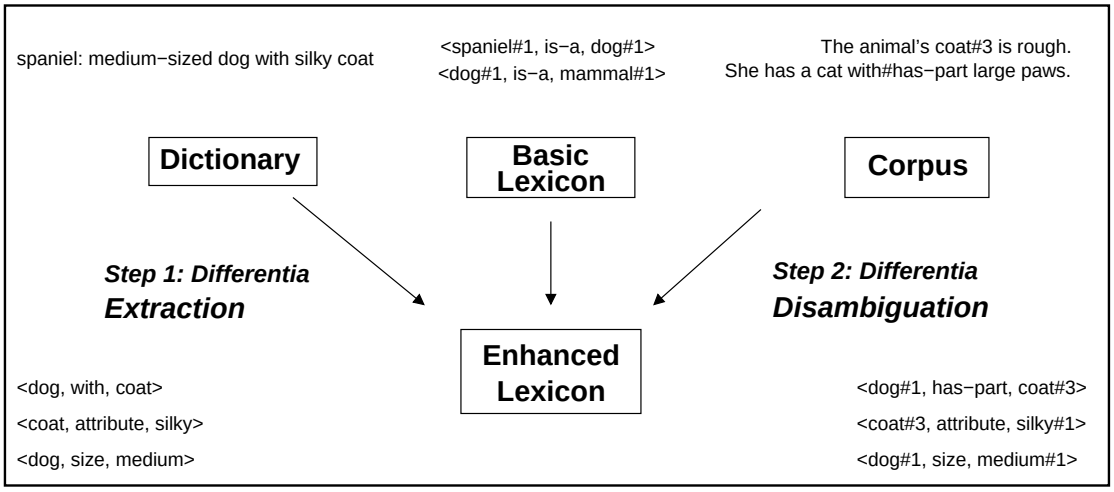


Figure 1.1: ***Overview of differentia extraction and disambiguation.*** Sample data is simplified version of actual input.

relational tuples. The components of the tuples at this stage are just words taken from the definition. During *Differentia Disambiguation*, the words are disambiguated to produce relations involving concepts. The example in Figure 1.1 shows that the content word 'coat' was disambiguated into the sense *coat#3* and also that the preposition 'with' was disambiguated into *has-part*. This second step is the focus of this thesis. In addition, the emphasis is on the disambiguation of prepositions, which has received much less attention than content-word disambiguation.

1.2 Motivation

It almost seems self-evident that differentiating relations are important for inclusion in semantic lexicons (or in knowledge bases in general). However, in practice, this type of information is often overlooked. Therefore, this section presents support for why it is important to incorporate such information.

1.2.1 Differentiating Relations are Important

There are several reasons why differentiating relations (i.e., differentia) are needed for natural language processing. The main motivation is that these are the properties that distinguish similar concepts from one another. Without accounting for them, applications would not be able to recognize the important

characteristics of particular concepts. Although some of these characteristics might emerge from corpus analysis, such analysis would also yield incidental associations not important for categorization. In other words, extracting the properties from definitions provides a more direct means of obtaining this information than other approaches, such as corpus analysis; and, in some cases, this might be the only automated way to obtain the information. Nonetheless, depending on the application, other approaches could be useful in order to maximize the information available. This section discusses in depth why differentiating relations are needed, and it shows why dictionaries are the best resource for extracting them.

1.2.1.1 Support from Lexicography

Dictionary definitions emphasize differentiating relations, because most dictionaries adhere to the *analytic* type of definition (Ayto, 1983, p. 89):

> The basic tool of lexicographic semantic analysis is in fact mirrored on the dictionary page, in the form of the classical 'analytic' definition. This consists of a 'genus' word designating a superordinate class to which that which is defined belongs, and 'differentiae,' which distinguish it from others in the same class.

There are several historical reasons for the predominance of this type of definition, including tradition and the influence of classical logic (Béjoint, 1994). Even so, the format does suit the needs of most users. Consider the main purpose of dictionaries: they are often perceived as 'authorities' on word meaning (Kilgarriff, 1997). Publishers often capitalize on this perception, such as in the following advertisement for *Merriam-Webster's Collegiate Dictionary*:[2]

> This best-selling dictionary is the 'voice of authority' with an in-depth quality that has earned the trust of schools and scholars for several generations.

There are two complementary aspects to the "dictionary as authority." An author might use a dictionary to make sure she is using a particular word in a commonly recognized sense. If her intended usage differs considerably from the senses detailed in the dictionary, she would probably consider rewording the selection. In contrast, a reader might use a dictionary to determine the

[2] From http://stage1.worldbook.com/products/htmla/mw.htm.

meaning of an unknown word or of an unfamiliar sense of a known word. The first aspect (word choice) is addressed by having a dictionary concentrate more on the differences in word meanings rather than commonalities; in contrast, a thesaurus addresses commonalities. The analytic definition clearly fits this need. The second aspect (word understanding) is addressed by having a dictionary use common language in the definitions whenever possible. Because this leads to the use of overly general category terms (i.e., the *genus* terms), more differentiation in the definitions is required for precision (i.e., the *differentiae*), again making the analytic definition suitable. For instance, consider the LDOCE[3] definition of 'money' versus the corresponding one in WordNet.

definitions for 'money':

LDOCE	pieces of metal made into coins, or paper notes with their value printed on them, given and taken in buying and selling
WordNet	the most common medium of exchange; functions as legal tender

The LDOCE definition incorporates more common terms than the WordNet definition, making it easier to grasp. This requires the use of additional differentiating information, for instance, to distinguish money from other "pieces of metal."

Dictionaries typically just *relate* a word sense to an underlying concept and provide enough information to distinguish it from other words related to the same concept. That is, it is assumed that the underlying concept is understood and need not be described; otherwise, an encyclopedia could in principle be consulted. The emphasis again is on what distinguishes particular concepts rather than on describing them in detail.

1.2.1.2 Support from Psychology

Distinguishing features play a prominent role in categorization. For instance, in Tversky's (1977) influential *contrast model*, the similarity comparison incorporates factors that account for features specific to one or the other category, as well as a factor for common features:

$$S(A, B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

[3] *LDOCE is the Longman Dictionary of Contemporary English* (Procter, 1978).

where $f(X)$ is a salience measure over a set of features and θ, α, and β are weighting factors.

By having separate weighting factors for the differences, this accounts for the common intuition that similarity is asymmetric. For instance, consider the differences in the following comparisons:

Butchers are like *surgeons*.
Surgeons are like *butchers*.

In these comparisons, the distinctive features of the italicized terms set the stage for the analogy. In addition, Tversky later conducted experiments (Gati and Tversky, 1984) showing that in certain cases, the distinctive features are given more weight than common ones. Similar results are reported in (Medin et al., 1993).

Rosch's research into the use of family resemblances in categorization highlighted the use of distinctive features as well as common features in categorization, especially with respect to *natural categories*, which are those for which people can readily associate typical members (Rosch, 1973; Rosch and Mervis, 1975). For instance, *chair* is a natural category, but not *furniture*. Natural categories are also referred to as basic-level categories, because they tend to occur at the level in a taxonomy where most of the information resides with respect to attributes (Rosch, 1973; Rosch and Mervis, 1975). One important finding was that natural categories are generally organized to maximize the similarity within a class and to minimize the similarity across classes. In effect, categorization relies on distinctive features through the use of *cue validities*, which refers to the degree to which a feature is associated with a particular category compared to the association with contrasting categories. In probabilistic terms, cue validity is the conditional probability of a class given a feature (Smith and Medin, 1981):

cue validity of feature F_i for class C_j:

$$P(C_j|F_i) = \frac{P(F_i|C_j)}{P(F_i|C_j) + P(F_i|C_k)}$$

where C_k is a concept that contrasts with C_j (assuming just one for simplicity).

Rosch further noted that class prototypes are those members or abstract representations that maximize the total cue validities. Thus, the notion of family resemblances can be cast in terms of cue validities, but she prefers the former term for clarity.

1.2.1.3 Support from Knowledge Representation

Conceptual knowledge is commonly organized into hierarchies called ontologies (Mahesh and Nirenburg, 1995). The concepts in these hierarchies are usually partially ordered via the instance and subset relations (i.e., *is-a* and *is-subset-of*). Each is a *relation of dominance*, which Cruse (1986) considers as the defining aspect of hierarchies. This ontological structure is implicit in dictionaries in the relations among the *genus* terms, each of which corresponds to a concept serving as the general category for the meaning of a word. Cruse (1986) points out that an important part of branching hierarchies is the horizontal differentiation among siblings. (Non-branching hierarchies correspond to a simple linear ordering and thus only require a relation of dominance.) Without the differentiating relations, the information in hierarchical lexicons would only indicate how the concepts represented by words are ordered without indicating differences among the concepts.

Manually derived lexicons, such as the one for English in the Mikrokosmos system for machine translation (Onyshkevych and Nirenburg, 1995), often contain differentia in the rich case-frame structures associated with the underlying concepts. This contrasts with semi-automatically derived lexicons such as WordNet (Miller, 1990), which emphasize the lexical hierarchy over the underlying semantics. For instance, Mikrokosmos[4] averages about 2.4 properties per concept (including some inverse relations), whereas WordNet[5] averages only 1.3 (including inverses).[6]

This suggests that the reason large-scale lexicons tend not to include such differentiating relations is due more to the difficulty in automatically extracting the information than to the relative worth of the information. This holds for both fully automated and partially automated lexicons. Hirst (1986) goes a step further by advocating the inclusion of case structures to standard dictionaries, in the same manner that learner's dictionaries indicate verbal subcategorization frames. This would provide a common resource for more-detailed language knowledge, useful for humans as well as for computerized processing.

---

[4] 1998 version of Mikrokosmos (crl.nmsu.edu/Research/Projects/mikro/index.html).

[5] Version 1.7 of WordNet (www.cogsci.princeton.edu/~wn).

[6] *Properties* refers to functional relations, attributes and part-whole relations (e.g., *is-member-meronym-of*), excluding just the instance and subset relations. WordNet 1.6 only averages 0.64 properties, so version 1.7 represents a substantial improvement.

1.2.2 Dictionary Definitions are the Best Source of Differentiating Relations

As mentioned earlier, corpus analysis is unlikely to be an effective source of differentiating relations. For example, collocations[7] for a particular word (or word sense) might include words that indicate aspects covered by the differentiating relations. They would include other types of relations as well, ranging from strict categorial relationships through generic relatedness down to mere coincidence. However, as generally used, collocation sets are just "bags of words," so that there is no indication of which words indicate which properties.

As an illustration, consider what information collocations might provide for the words 'city' and 'state,' both taken in the administrative sense. Figure 1.2 illustrates the variety of collocational words obtained when using a standard conditional probability test to select collocations indicative of a particular sense (Wiebe et al., 1997). Note that there is very little overlap among the collocations for the two senses, although they are clearly related. The two words that overlap, 'commissioner' and 'license,' only involve *situational relationships* to both senses (Morris and Hirst, 1991). However, the main point to note is that there is no mention of 'state' in the collocation list for $city_{\text{ADMIN}}$ (or vice versa), except for a few states listed along with other proper names. The closest connection would be through 'municipal,' listed under the collocations for $state_{\text{ADMIN}}$. Thus, the fact that cities are governed by states would not be readily inferable by this type of corpus analysis. (A similar result holds for collocations selected from the *Wall Street Journal* portion of the DSO corpus.)

In summary, distinguishing properties are not likely to emerge from statistical analysis of raw corpora using current techniques. Therefore, at least for the time being, the best source for them is dictionary text. Although definitions indicate differentiation using standard conventions, there is the complication that they are given in natural language, which leads to the usual problems with structural and lexical ambiguities during analysis. The only other viable alternative is manual encoding of lexicons, which is undesirable due to the amount of time required. Atkins (1995) estimates that it would take 100 person-years to properly develop a semantic lexical database comparable in scope to a standard college dictionary.

[7] Collocations are used here in the broad sense of words that co-occur often in context: there are no constraints on word order, etc.

city: an incorporated administrative district established by state charter

administration animal battle both calendar can- candidate care choose close collection commission **commissioner** condition convert coordinate council county critic defence department dependent develop dog ease effectively employes end estimate exception fee fight file finance financial fireman fiscal gladden government hand hear improve interest island lead leader **license** lively local locally majority narcotic organize outspoken pay personnel political possible problem property propose purchase raise range reaction reply rule serious service sewer six tax teamster union valley vary welfare year yesterday

state: the territory occupied by one of the constituent administrative districts of a nation

adjust adjustment air allot allotment along assess assessment audio-visual belong blind boat bond border bridge century chain chapter coast **commissioner** confederacy decry designate divide downward draft eighteenth exceed firm five force forth four fourteen guard head hill identify immigration impressive inhabitant intangible item legislature **license** look mail merchant minimum month *municipal* murphy navy nomination non-residents nothing open openly participation particular peas percentage pick pivot pre-1960 prepare questionnaire race recoup reduction refer respective respectively secede sell snow southeastern stay sum tangible taxation team thereby thirteen twenty-five unadjusted uniformly upward

Figure 1.2: ***Collocations for 'city' and 'state' (in administrative sense).*** These were selected from the Brown portion of the word-sense annotated corpus produced by Ng and Lee (1996) at Singapore's Defense Sciences Organization (DSO).

1.3 Contributions of this Research

There are three main contributions of this thesis work: 1) the methodology for extracting and disambiguating semantic relations from dictionary text; 2) the representation of these properties using Bayesian networks; and 3) the use of these semantic relations to improve word-sense disambiguation. Each of these is discussed briefly in the following subsections.

1.3.1 Empirical Extraction and Disambiguation of Semantic Relations

Previous approaches have mostly used hand-coded pattern matching rules for extracting relations from dictionary definitions (Vanderwende, 1996; Barrière, 1997). This can be considered as a top-down, model-driven approach. This can be very precise, but achieving broad coverage can be difficult. Instead, we employ a bottom-up, data-driven approach. Specifically, a broad-coverage dependency parser is first used to determine the syntactic relations that are present among the constituents in the sentence. Then the syntactic relations between sentential constituents are disambiguated into semantic relations between the underlying concepts. In other words, the surface-level syntactic relationships determined by the parser are disambiguated into the underlying conceptual relationships.

The disambiguation involves each of the three components of the relationships: the *source* and *target* terms being related, and the *relation type*. Standard word-sense disambiguation (WSD) approaches are used to disambiguate the terms being related, but this aspect is not a focus of the research. Statistical classifiers are used to disambiguate the relation types, exploiting both tagged corpora (e.g., semantic role annotations) and knowledge bases for the training data. Relations indicated by prepositions are addressed here, so the classification can be viewed as preposition WSD. Note that these are general-purpose classifiers, not ones tied into dictionary text. (See Appendix A for a primer on statistical classification for machine learning.)

The method of creating these relation type classifiers and integrating them into the differentia extraction process forms the basis for the most important contribution of this thesis (see Chapter 4). Isolating the disambiguation step from the extraction step in this manner allows for greater flexibility over earlier approaches. For example, different parsers can be incorporated without having to rework the disambiguation process. Although quantitative assessments of this flexibility are not provided, the information produced by the process has been evaluated by several different human judges (as discussed later in Section 5.1).

1.3.2 Representation of Semantic Relations using Bayesian Networks

Often the differentiating information in definitions reflects typical properties of the concept being described. For example, most beagles have brown spots. This is modeled by attaching probabilities to each relation that is extracted from the definition. The result is a semantic network in the form of a labeled directed graph, where each link has a probability attached indicating degree of applicability.

To incorporate these probabilities in applications utilizing semantic relatedness, the semantic network representing the concept relationships is converted into a Bayesian network, which is a directed acyclic graph (DAG). (See Appendix B for a primer on Bayesian networks.) The DAG's are not labeled, so the various relation types (e.g., *is-a*, *used-for*), are conflated into a single *related-to* relation. To account for the different degrees to which the various relation types indicate semantic relatedness, the relation strengths from the original semantic network are scaled by a factor representing the degree to which the type of relationship is specific to the concept compared to similar concepts. This models the salience of a relationship for a particular concept. The result is a Bayesian network where the nodes represent concepts and the links, the degree of relatedness between specific concepts. This network can then be used to implement probabilistic spreading activation.

1.3.3 Improvements in Word Sense Disambiguation

The above Bayesian network representation for the differentiating information is utilized to improve a word-sense disambiguation system that uses both statistical classification as well as probabilistic spreading activation (Wiebe et al., 1998b). The original system combined analytical knowledge about the dependencies among word senses in WordNet along with empirical knowledge for the suitability of particular senses of a word in context. Adding the differentiating relations extracted from the WordNet definition glosses leads to improvements that are statistically significant.

1.4 Organization of Thesis

This chapter has presented an overview of the types of semantic relations in computational lexicons, emphasizing the importance of differentiating relations (i.e., *differentia*). Support for this comes from three different areas. Lexicographic practice generally dictates the use of the analytic type of definition where the *genus* terms indicate general categories and the *differentia*

descriptions indicate how terms mapped to the same category differ. Psychological research shows the importance of differentiation in categorization, as in the notion of *natural categories*, which maximize *cue validities* (e.g., conditional probability of features given a class). Additional support for this is from knowledge representation, where manually encoded KB's (e.g., Mikrokosmos) generally have a much higher degree of differentiating relations than semi-automatically ones (e.g., WordNet).

This chapter also illustrates why dictionaries are a prime source for differentia. For instance, corpus analysis is not likely to be sufficiently directed in order to obtain this type of information while minimizing extraneous associations. The methodology for differentia extraction involves statistical disambiguation of the relation types, given output of a broad coverage parser over the definition text. Other contributions of this thesis include a representation supporting probabilistic spreading activation and two new approaches for word-sense disambiguation. The remainder of this section describes the organization of the thesis proper.

Chapter 2 presents related work mainly in lexical acquisition but also covers work in natural language processing that incorporates machine learning and Bayesian networks. Chapter 3 discusses how the surface-level relationships are extracted from dictionary definitions, using a general-purpose dependency parser. This concentrates on differentia (i.e., the distinguishing relations and properties). In addition, the chapter discusses the structure of WordNet and presents an analysis of the types of semantic relations implicit in the definitions, based on manual annotations that were performed over several hundred definitions. Note that these definition annotations are just used to provide insight on the types of relations that occur in definitions. Chapter 4 discusses the disambiguation of the surface-level relationships into conceptual relationships. This briefly covers the disambiguation of the source and target terms from English words to their most likely word senses, a process which has received much attention recently (Edmonds and Kilgarriff, 2002). More emphasis is placed on the disambiguation of relation types from their English specification (e.g., preposition) into the underlying concept for the relation type. The relation disambiguation exploits large corpora of semantic role annotations in general-purpose text.

Chapter 5 discusses the application of the work to lexicon augmentation and word-sense disambiguation, including detailed evaluations for each. This includes the statistical representation for these relationships using Bayesian networks, chosen to facilitate integration with common statistical approaches used in natural language processing (e.g., Bayesian classifiers).

Chapter 6 compares the research to related work in the acquisition of lexical semantics, with an emphasis on previous work on extracting informa-

tion from dictionaries. It also sketches out areas for future research, such as the long-term goal of applying the techniques to general text analysis rather than just dictionary definitions. Chapter 7 summarizes the work and the main contributions of the research.

There is also an appendix providing brief primers on areas of artificial intelligence that might be unfamiliar to readers with general computational linguistic backgrounds. Appendix A explains the general framework for machine learning and discusses the two main types used in this research. Appendix B gives a basic introduction to Bayesian networks, which are popular in artificial intelligence for representing probabilistic relations. Lastly, a glossary is included for important technical terms as well as those used here in specialized senses.

CHAPTER 2
BACKGROUND ON LEXICAL SEMANTICS ACQUISITION

This thesis approaches the task of acquiring conceptual distinctions from a computational linguistics framework (e.g., computational semantics). In computational semantics, as in linguistics and lexicography, the emphasis is on word-sense distinctions rather than conceptual distinctions in general. Word senses can loosely be considered as concepts, albeit ones specialized to different languages. For example, the canine senses of the words 'perro' and 'dog' both refer to the same underlying concept (i.e., *Canis familiaris*), but strictly speaking they are two distinct senses.

Lexical knowledge encompasses all the information that is known about words and the relationships among them. In addition to strictly linguistic knowledge such as phonology, morphology, and grammatical categories, this includes conceptual knowledge (e.g., semantic categories), and pragmatic knowledge, such as conventional usages for certain words. The emphasis here is on semantic knowledge in the sense of conceptual meaning rather than associative or thematic meaning (Leech, 1974). Conceptual meaning corresponds to the basic denotation for words; in contrast, associative meaning covers stylistic and connotative aspects, and thematic meaning refers to emphasis due to word order, etc.

This chapter primarily reviews work in computational semantics related to the acquisition of word-sense distinctions (Sections 2.2 and 2.3). It also covers the representation and utilization of such lexical information, as well as other supporting areas (Sections 2.1 and 2.4).

2.1 Background on Lexical Semantics

2.1.1 Linguistics

Work in linguistics based on generative grammar tends to treat the lexicon as an ancillary resource providing information on features such as grammatical number and occasionally subcategorizations (Heim and Kratzer, 1998; van Riemsdijk and Williams, 1986). Although generative grammar does incorporate the notions of case and thematic roles, the use is generally restricted to describing how the roles are assigned by particular governing categories. In other approaches, case roles are central to the theory. Two examples are Fillmore's (1968; 1977) work on case frames and Jackendoff's (1990) semantic structures which incorporate thematic relations into his framework for general cognition (Jackendoff, 1983). Formal semantics work based on the Montague

Case	Description
agentive	the typically animate perceived instigator
instrumental	inanimate force or object causally involved in the situation
dative	the animate being affected by the situation
factitive	object or being resulting from the situation
locative	location or spatial orientation of the situation
objective	anything representable by a noun whose role in the situation is identified by semantic interpretation of the verb

Table 2.1: ***Case roles identified by Fillmore.*** Situations refer to both events and states to simplify the original descriptions (Fillmore, 1968, pp. 24-25).

tradition (Dowty, 1979) accounts more centrally for the semantics of words, such as through meaning postulates (Chierchia and McConnell-Ginet, 2000); however, the scope tends to be somewhat limited. There is a variety of other work in linguistics that can serve as useful resources in computational semantics; for example, Raskin and Nirenburg (1995) illustrate a methodology for reconciling various theories of the lexical semantics of adjectives when developing the framework for a computational system.

Fillmore (1968) holds that deep structure defines the relevant case relations: surface case relations are often insignificant, because the mapping from deep structure to surface structure is not one-to-one. Fillmore identifies a half dozen or so cases that any case system should include, shown in Table 2.1; but, he feels more would be needed in practice (e.g., benefactive).

Jackendoff (1983) presents a unified framework for representing conceptual knowledge for all aspects of cognition, not just linguistic competence. The representation is an outgrowth of earlier work in artificial intelligence, such as by Schank and Wilks (Schank, 1973; Wilks, 1975b; Wilks, 1978), discussed later in Section 2.1.3. Jackendoff's framework incorporates a few noteworthy innovations. One is that all semantic categories are treated uniformly: in particular, manners and directions have the same status as things and events. Another is the emphasis on thematic relations, showing how prepositions play an integral part in the analysis of several different semantic fields. Later work (Jackendoff, 1990) builds upon this framework to present a detailed analysis of various types of natural language expressions. An important aspect of this work is that the interpretation of adjuncts is given full treatment, a requirement for sentential interpretations. For example, "Paint ran all over the wall" is represented as

[event GO ([thing paint], [path $\text{TO}_{+\text{DIST}}$ [place $\text{ON}_{+\text{DIST}}$ [wall]]])].

Relation	Description
hyponymy	z in X $\Rightarrow$ z in Y (i.e, subset relation)
taxonomy	X is a kind/type of Y
meronymy	X is part of Y; also called *partonymy*
cognitive synonymy	X exactly equivalent to Y
plesionymy	X is similar in meaning to Y
antonymy	X is opposite to Y
paronymy	X is derived from Y (of different syntactic category)

Table 2.2: ***Basic lexical relations defined by Cruse.*** Descriptions are based on (Cruse, 1986).

Here, the distributive interpretation of the location is indicated by *+DIST*.

Jackendoff's conceptual representations of words, called *Lexical Conceptual Structures* (LCS), tend to be at a coarse level. For instance, distinctions in meaning of perceptual objects and motion verbs are to be captured elsewhere using geometric representations (i.e., via schematics rather than logical descriptions). Dorr and others at the University of Maryland (Dorr, 1997; Dorr et al., 1998) have created a large lexical database for machine translation based on Jackendoff's LCS. This concentrates on verb structure and incorporates information from Levin's (1993) verb classes. Of particular note is the inclusion of lexical entries for prepositions as this information is often omitted in computational lexicons. Over 150 prepositions are included with over 500 distinct LCS structures.

Cruse details the important types of lexical relations with emphasis on paradigmatic rather than syntagmatic relations. (Paradigmatic relations hold between elements that can be substituted for one another in the same syntactic context, whereas syntagmatic relations hold between elements that can occur together in the same sentential context.) Table 2.2 shows a representative sample of the basic relations.

The relations delineated by Cruse tend to be at an abstract level. Certain types of relations useful for representing conceptual distinctions are only implicitly addressed. For example, accounting for Fillmore's *instrumental* relation would require a configuration within which both the instrument and the facilitated action are considered as parts. Work in formal semantics tends not to cover such *functional relations* much, although there are some notable exceptions. Pustejovsky's *Generative Lexicon* theory accounts for them in his *qualia* structure (Pustejovsky, 1995). This encapsulates aspects of lexical meaning separate from argument valency structure, decomposition (e.g., subevents), and type inheritance. Mel'čuk's *Meaning Text Theory* (Mel'čuk and Polguere, 1987) accounts for functional relations via lexical functions in his *Explanatory*

Combinatorial Dictionary (ECD). For a given headword, the lexical functions indicate lexemes that serve in a variety of syntactic (e.g., typical object) and semantic relationships (e.g., opposition). Heylen (1995) discusses the connection between the two theories and shows how most of the qualia components of the Generative Lexicon can be derived from the ECD.

2.1.2 Lexicography

Work in lexicography provides insight into lexical semantics, especially in regard to word-sense distinctions. Kilgarriff (1997) criticizes the assumption of a distinct set of word senses independent of usage in particular applications, which is often implicit in word-sense disambiguation (WSD) research. One problem is that lexicographers might distill disparate usages from citation files into the same sense in a definition. Another is that some senses will not be covered in dictionaries, as not all the citations can be addressed.

Landau (2001) stresses that most dictionaries represent written language, since the citations are predominantly from written sources. Therefore, it is not the ultimate authority on language, just an account of language usage as determined from written text. In addition, a key constraint on dictionary definitions is lack of space, which implies that one should treat definitions as potentially being incomplete or vague about important details needed for fully understanding a concept.

McCawley (1986) offers some suggestions on how lexicography can be improved. For example, it would be helpful if dictionaries explicitly indicate that relational nouns (e.g., ‘husband’) generally involve syntagmatic relations in context. Therefore, grammatical tags analogous to transitivity for verbs are desirable. This is related to the problem that definitions tend to emphasize the referent of the word rather than the specifics of the word itself. Moreover, dictionaries often do not clearly indicate that such encyclopedic information is added mainly for the sake of illustration rather than being a critical part of the word’s meaning.

2.1.3 Computational Semantics

2.1.3.1 Semantic Networks

Quillian’s (1968) work on semantic memory is significant for several reasons. The main contribution is the introduction of semantic networks for knowledge representation, and it was one of the first computational attempts to emphasize semantics over syntax. His work centered on encoding entire dictionary definitions. Each word sense is represented by a graph with nodes for

the defining words and links for the relations between the words, based on the definition.

Schank (1973) popularized the notion of semantic-based analysis in the use of conceptual dependencies to represent meaning. The motivations for the conceptual dependency representation are to facilitate paraphrases, to support inference, and to model human memory. To this end, a small set of semantic primitives was developed, with which all expressions were encoded. Conceptual categories serve as the basic unit of the representation. Relations among these conceptual categories are called *dependencies*. This approach has an advantage over Quillian's in facilitating inferences over the encoded meaning representation. However, subtle distinctions in meaning might be lost in the conversion process

Wilks' (1975b; 1978) work was similar in spirit to Schank's, but he emphasized the resolution of lexical ambiguity. Moreover, his representation clearly distinguishes criterial aspects of word meaning from optional (or preferred) aspects. In his basic mode of analysis, interpretation is performed by finding the set of word-sense formulas maximizing the density of satisfied preferences, which mainly cover selectional restrictions. However, procedural knowledge was used in the heuristics for the selection of competing interpretations. Later extensions (Wilks, 1978) organized the vocabulary through a thesaural hierarchy.

Both Schank and Wilks emphasized the use of case relations in their representations. Bruce (1975) provides a survey of early uses of cases systems in natural language processing, as well as providing background on surface cases versus deep cases. Several criteria for selecting deep cases are discussed, such as the need for distinguishing word senses, for specifying events uniquely, and for modeling relevant domain aspects. His definition of 'case' is thus quite generic (Bruce, 1975, p. 336):

> A case is a relation which is "important" for an event in the context in which it is described.

2.1.3.2 Word Experts/Agents

Small popularized the idea of *word experts* (Small and Rieger, 1982), which are autonomous agents encapsulating the various aspects of knowledge regarding a word (or stem, affix, etc.). In his model, the control mechanism is modeled after the Unix-style processes and demons. Specifically, the word experts become active only for as long as they can perform useful work, such as refining a concept based on long-term memory (e.g., world-knowledge).

When they no longer can do productive work, they suspend themselves until type-specific interrupts or signals occur. Hirst (1988) developed a declarative system for representing lexical knowledge, using a conventional frame-based representational language. This work was influenced by psychological research into negative priming. To model priming, spreading activation is implemented via marker passing among nodes in the knowledge base. Hirst proposed the notion of self-developing objects, called Polaroid Words, for modeling the incremental development of lexical knowledge during comprehension. If the objects are not fully resolved after the sentence is processed, then several fallback (procedural) mechanisms are applied, such as selecting preferred senses, and relaxing the marker passing constraints. Note that Hirst and Small both rely on word-specific *agents* to encapsulate lexical knowledge and world knowledge; however, Small's approach is predominantly procedural, whereas Hirst's is mostly declarative.

2.1.3.3 Ontological Semantics

Onyshkevych and Nirenburg (1995) advocate the ontological approach to lexicon development, where language-dependent information is kept separate from general world knowledge that is organized in a taxonomy called the *ontology*. A rich frame structure is used for both the ontology and the lexicon. Concepts are defined in terms of other concepts using a variety of semantic relations (e.g., *is-a*, *member-of*, and *has-part*). For each lexical entry, the connection between syntax and semantics is established by specifying the correspondence between the grammatical arguments and concepts in the ontology. This is accomplished via variable linkages between argument placeholders in the syntactic structure (SYN-STRUC) and concept placeholders in the semantic structure (SEM-STRUC). See Figure 2.1 for an illustration. Simple lexical mappings are specified directly in terms of a single concept with the lexicon entry mainly providing syntactic information relevant to the word. Complex lexical mappings can override defaults associated with the concepts and provide selectional restrictions associated with the word (e.g., verbal arguments).

The Cyc knowledge base (KB) is a large-scale repository of commonsense knowledge that has been in development for about 20 years (Lenat, 1995), containing over 120,000 concepts and a million assertions. Cyc was initially developed using a frame-based representation, but it now uses first-order predicate calculus with a few minor extensions (e.g., some second-order features for efficient indexing). Natural language lexicons are integrated directly into the Cyc KB (Burns and Davis, 1999). There are several natural language lexicons in the KB, kept separate via microtheories, but the English lexicon is the only full-scale one. The mapping from phrases to concepts is done through

```
(book
  (book-N1
    (cat n)                    ;; category
      (morph)                  ;; morphology
      (anno                    ;; annotations
        (def “a copy of a written work or composition that has been published”)
        (ex “I just read a good book on economics”)
        (cross-ref))           ;; other lexemes referencing this entry
      (syn)                    ;; syntactic features
      (syn-struc               ;; syntactic structure
        (1 ((root $var0)
            (cat n)) ))
      (sem-struc               ;; semantic structure
        (lex-map               ;; lexical mapping
          (1 (book)) ))        ;;      to concept book
      (lex-rules)              ;; lexical rules
      (pragm)                  ;; pragmatics
      (styl)))                 ;; stylistics
```

Figure 2.1: ***Mikrokosmos lexical representation for ‘book’.*** Some of the zones are left unspecified (e.g., *pragm*). Descriptions based on (Onyshkevych and Nirenburg, 1995).

a variety of lexical assertions. Proper name assertions map strings to individuals in the KB. A denotational assertion maps a phrase into a concept, usually a collection. The phrase is specified via a lexical word unit (i.e., lexeme concept) with optional string modifiers. In addition, complex subcategorization assertions are used for mapping the arguments of verbs and other predicates into the underlying semantics.

WordNet (Miller et al., 1990) combines aspects of a traditional dictionary and thesaurus. It is structured around groups of synonymous words called *synsets* (for synonym sets). WordNet also provides definitions and usage examples; but, more importantly, it provides explicit relationships among the synsets, emphasizing taxonomic (*is-a*) and part-whole (*has-a*) relationships. Thus, WordNet represents a basic ontology. The main drawback to the WordNet ontology is that it is particular to English. Therefore, separate ontologies would be needed for other languages. The EuroWordNet project (Vossen et al., 1997) is seeking to tie together separate "wordnets" that are being developed for several different languages. It is addressing the problem of having separate ontologies for each language by specifying high-level correspondences.

Hirst (1995) proposed a variation of the ontological approach to lexicon semantics to account for subtle word-sense distinctions dealing with near synonyms (called *plesionyms*). The approach sketched out is to represent the differences among the plesionyms as objects in order that they can be manipulated directly. He suggests a two-level knowledge representation scheme, modeled after proposals common in the literature. Course-grained conceptual knowledge would be stored in a taxonomy, whereas fine-grained language-specific knowledge is stored in the lexicon. For plesionyms that represent distinct concepts, differences can be determined by comparing attributes, including those inherited from ancestors leading to a common ancestor.

Edmonds (1999) follows up in this line of research by showing how the differences among near synonyms can be represented using conventional ontologies augmented with non-denotational relations to account for the stylistic differences (Edmonds and Hirst, 2002). Specifically, the plesionyms would have traditional denotations to common concepts (e.g., 'mistake' and 'error' to *generic-error*). In addition, there will be additional relations to account for the pragmatic information associated with words. These would not provide necessary and sufficient conditions as with the denotations but rather preferences typical of the words. For example, 'blunder' would imply a high degree of perjorativeness. Inkpen and Hirst (2001) discuss how to automate the acquisition of such fine-grained distinctions by analyzing specialized synonymy dictionaries. Decision lists of indicative keywords are learned for the broad types of pragmatic distinctions, and these are then manually split into decision lists for the particular values of each distinction.

2.2 Manual Acquisition

Manual acquisition has been most commonly used when the lexicon quality is critical (Onyshkevych and Nirenburg, 1995). For example, most of the Mikrokosmos ontology was manually created, as was the core of the lexicons. The Mikrokosmos project has also investigated ways of capitalizing on the existing lexical knowledge in the ontology to help automate the creation of new lexical entries (Viegas et al., 1996). They stress that even with well-constrained rules, manual review is inevitable, and thus this cost needs to be accounted for during semi-automatic lexicon acquisition.

The core of the Cyc knowledge base has been carefully constructed by knowledge engineers, many of whom have formal backgrounds in logic and philosophy. Similarly, most of the knowledge in the Cyc Lexicon was manually entered by knowledge engineers with backgrounds in computational linguistics or philosophy of language. Some of the lexical information was provided by knowledge engineers without backgrounds in linguistics, but most of this has been reviewed by the computational lexicographers at Cycorp.[1] Recently, there has been work on providing interfaces for non-technical users to enter both general and lexical knowledge into the system (Witbrock et al., 2003), but this is still in the experimental stages.

WordNet was originally motivated by psycholinguistic principles of meaning representation (Miller, 1996). However, it has become very useful for general research in computational linguistics. Princeton's cognitive science group (Miller et al., 1993) manually created WordNet, using *Collins English Dictionary* as a starting point for the senses. Initially, the definitions were simply used for clarification rather than for defining word meaning as in traditional dictionaries. For example, if the combination of words in a synonym set clearly indicated the intended meaning, then the definition might be omitted. Later, more emphasis was placed on the definitions, due both to increased ambiguity as WordNet expanded and to requests from users who expected fuller definitions. Recently, there has been work on making some of the information in the WordNet definitions more explicit, using semi-automated techniques as discussed later in Section 2.3.3.

2.3 Automated Acquisition

Given the cost involved in manual acquisition, it is desirable to automate the process as much as possible. Complete automation is often not feasible. It

[1] This information is based on personal experience from when working at Cycorp.

might even be undesirable, unless the acquired information is similar in quality as that for manual entry. Otherwise, there is liable to be a considerable amount of post-editing, depending on the level of detail. This section concentrates on the acquisition of semantics. There has been much work on acquiring syntactic information, such as part-of-speech and subcategorization frames (Boguraev and Briscoe, 1989; Wilks et al., 1996; Litkowski, 1997).

2.3.1 Corpus Analysis

2.3.1.1 Word Classes

Word clustering is commonly used in order to infer classes from untagged corpora. For example, Pereira et al. (1993) determine word class for nouns based on how similar the distributions are with respect to co-occurrence with specific verbs, using *relative entropy* (or *divergence*) as their similarity measure:[2]

$$D(p_{n1} \mid\mid p_{n2}) = \sum_{v} p_{n1}(v) \log(\frac{p_{n1}(v)}{p_{n2}(v)})$$

where $p_n(v) = f_{vn} / \sum_{v} f_{vn}$. Lin (1998) provides for thesaurus-like classes by checking for a wide variety of syntactic contexts rather than just direct objects. A broad-coverage parser is first used to extract dependency tuples of the form $\langle$word1, grammatical-relation, word2$\rangle$. He measures word similarity based on frequency of the tuples and their constituents using *mutual information* (Manning and Schütze, 1999), which measures the difference in the joint occurrence of two events versus the occurrence expected by chance (i.e., $-\log(\frac{p(xy)}{p(x)p(y)})$). The mutual information (MI) for the co-occurrences of two words in a particular grammatical relationship is defined as follows:

$$MI(w1, r, w2) = -\log(\frac{(P(r)P(w1|r)P(w2|r))}{P(w1, r, w2))}$$

The similarity of two words is then calculated by the ratio of the summed MI scores for common words that both are related to versus the summed MI scores for all the words related to either of them.

[2]Entropy measures the uniformity of a distribution: $\sum_x -p(x)\log(p(x))$. See Appendix A for more details.

As seen later, the *Longman Dictionary of Contemporary English* (LDOCE) is often used in dictionary analysis (Procter, 1978). Slator et al. (1990) apply clustering to preposition descriptions derived from LDOCE to infer semantic classes based on usage. The prepositions are manually annotated as a vector with features for aspects of the LDOCE definition and for the semantic codes of the complements used in the examples. For instance, one component is the set of subject codes for the object of the preposition. A distance metric is defined and then Pathfinder (Schvaneveldt et al., 1988) is used to reduce the network of pairwise distances into one in which each link is maintained only if the transitive closure does not produce a shorter path. The resulting clusters then represent the classes for the prepositions.

2.3.1.2 Lexical Associations and Selectional Restrictions

Lexical associations derived from corpus analysis have been shown to be useful for structural disambiguation and other tasks. Hindle and Rooth (1993) were the first to demonstrate the basic technique. They show how to induce lexical associations from simple syntactic relationships (e.g., verb/object) extracted using shallow parsing in combination with a few heuristics for resolving ambiguous relationships. These associations can be considered as conditional probabilities that a particular preposition is attached to the noun or verb, given that the latter is present. Attachment is resolved by selecting the case with the higher association. To train the system, they first applied a part-of-speech (POS) tagger and a shallow parser to a large newswire corpus and then extracted tuples of the form $\langle$verb, noun, prep$\rangle$ from the parses, where either the verb or the noun might be empty. Next, heuristics were applied to associate the preposition with the verb or noun, and the results were tabulated to produce $\langle$verb, prep$\rangle$ and $\langle$noun, prep$\rangle$ bigram frequency counts.

To decide on the attachment for test data, the POS tagging and parsing are performed as above, along with the extraction of the tuples. Then, instead of using the heuristics on each ambiguous tuple (i.e., those with both verb and noun non-empty), the bigram frequencies are used in a log-likelihood ratio test:

$$\log_2 \frac{\mathrm{P}(\text{verb-attach prep} \mid \mathrm{v}, \mathrm{n})}{\mathrm{P}(\text{noun-attach prep} \mid \mathrm{v}, \mathrm{n})}$$

where $\mathrm{P}(\text{verb-attach prep} \mid \mathrm{v}, \mathrm{n})$ is estimated by $\mathrm{freq}(\text{verb}, \text{prep})/\mathrm{TotalFreq}$ and likewise for the noun attachment probability.

Basili et al. (1996b) show how the same type of disambiguation can be achieved using selectional restrictions that are semi-automatically acquired from corpus statistics. These are relational tuples of the form $\langle word_1$, relation,

word$_2\rangle$. They define *semantic expectation* as the probability that a pair of concepts occurs in a given relationship, based on the selectional restrictions for words mapping into the concepts. Manual effort is first required to assign the high-level concepts to the entries in the lexicon. However, once this has been done, the rest of the process is automatic (i.e., the determination of the selection restrictions for particular words). An experiment in deciding prepositional attachment shows how this method improves upon an extension to Hindle and Rooth's (1993) technique.

Building upon this basic framework for determining verb subcategorizations, Basili et al. (1996a) show how verbs can be hierarchically clustered into classes. The classification is based on maximizing the extent to which categories are associated with different attributes, which is similar to the notion of cue validities discussed in the previous chapter (see Section 1.2.1.2):

$$\sum_{k=1}^{K} P(C_k) \sum_{ij} P(attr_i = val_j \mid C_k)^2$$

This can be seen as minimizing the mean entropy of the distribution of the likelihood for the attribute values. The attributes are based on the pairings of thematic roles and conceptual types derived from the relational tuples. The main advantage of this clustering approach is that the thematic roles can serve in the semantic description of the classes.

Resnik (1993) has done some influential work on combining statistical approaches with more traditional knowledge-based approaches. For instance, he defines a measure based on information content for the semantic similarity of nouns that uses the WordNet hierarchy along with frequency statistics for each synset. His technique relies on the use of WordNet synsets to define the classes over which frequency statistics are maintained. This avoids the data sparseness problem associated with statistical inference at the word level; the classes also provide an abstraction that facilitates comparison. For instance, he defines selectional preference profiles for verbs by tabulating the distribution of the classes for the verbal subjects and objects. The degree to which verbs select for their arguments can be summarized by a measure called the *selectional preference strength*. This is the relative entropy of the distribution for the conditional probability of the classes given the verb compared to the distribution of the prior probabilities for the classes:

$$S(v) = D(P(C|v) \mid\mid P(C)) = \sum_{c \in C} P(c|v) \log\left(\frac{P(c|v)}{P(v)}\right)$$

To find out the preference for a particular class, the *selectional association* measure is defined as follows:

$$A(v,c) = \frac{1}{S(v)} P(c|v) \log\left(\frac{P(c|v)}{P(c)}\right)$$

This is the relative contribution that the class makes to the selectional preference strength.

There has been considerable work on domain-specific case frame acquisition, especially in the context of information extraction. Much of this has relied upon manually derived extraction rules, such as in the work by Lehnert et al. (1992) at the University of Massachusetts. They later (Lehnert et al., 1993) implemented steps to partly automate this process, such as in the use of semantic dictionaries inferred from the training data. Riloff and Schmelzenbach (1998) further automate this process by learning selectional restrictions from corpora. In follow-up work, Phillips and Riloff (2002) show how to learn semantic categories for words using highly constrained syntactic patterns (e.g., appositive with proper noun followed by common noun).

2.3.1.3 Translation Lexicons

In addition to analyzing large corpora of the same language, there have been several projects that have used bilingual corpora of the same text in different languages, for example, transcripts of the Canadian parliament (Hansards) in French and English (Brown et al., 1990). Once the sentences have been aligned, fairly accurate lexical associations can be made between synonymous words in the two languages (Gale et al., 1993). This has the advantage of producing a quick and dirty translation lexicon tuned to a particular corpus. It has also been found to be useful in lexical ambiguity resolution, since an ambiguous word might be consistently associated with different unambiguous words in the other language (Dagan et al., 1991).

Fung and Church (1994) present a simple approach for inducing a translation lexicon given two parallel texts. Both texts are divided in fixed blocks of a given size. For each word in the text, a vector of the block size is produced indicating if the word occurs in each of the blocks. Given the occurrence vectors, contingency matrices are produced and used to derive mutual information statistics. More sophisticated models for word alignment were developed specifically for machine translation (MT). The models originally developed at IBM are now available in the publicly available statistical MT package GIZA (Brown et al., 1993; Al-Onaizan et al., 1999). Melamed (2000) discusses lexicon induction in depth and presents a formal statistical model for the process. He improves upon earlier approaches via his *Competitive Linking* algorithm, which does not allow word linkages to be considered twice when inducing the translation lexicon.

2.3.2 Lexical Rules

In computational semantics, most of the work involved in exploiting existing manually encoded knowledge deals with lexical rules. There has been much work on the coercion of count nouns into mass nouns (and vice versa), such as the 'grinding rule' (Briscoe et al., 1995), a special case of which covers animal terms becoming mass nouns when referring to the food (e.g., "Let's have chicken tonight."). Gillon (1999) generalizes this and similar cases to a rule that converts a count noun usage for any object to a mass noun usage referring to an aggregate part of the object (e.g., meat in the case of the animal grinding rule).

As mentioned earlier, Viegas et al. (1996) use lexical rules to extend the Mikrokosmos lexicons. For instance, they use lexical rules to infer morphologically related entries for Spanish verbs, using online dictionaries and corpora to guard against overgeneration. For example, the entry for 'comprador' (buyer) would be derived from the one for 'comprar' (to buy). They also point out difficulties associated with lexical rules, such as for English adjectives. For example, '-able' is a very productive affix for converting a verb into an adjective, but it is not applicable to all senses of the verb or involves a restricted interpretation (e.g., 'perishable' does not apply to humans).

Briscoe et al. (1995) present a formal account of how to model defaults in the lexicon while still allowing the defaults to be overridden. At issue is how to allow for blocking of lexical rules (for inheritance networks) in certain situations, such as when another lexical item is equivalent. For instance, the animal grinding is normally blocked for 'cow' since another word, 'beef', already accounts for it.

Pustejovsky's Generative Lexicon (1995) can be seen as formalizing the use of lexical rules. The main goal of the Generative Lexicon is to minimize the need for enumerating different senses of a word by providing operations for deriving most senses for a word from a core sense. This contrasts with the standard approach based on traditional lexicography ("sense-enumeration"), in which numerous distinct senses are listed for particular words. To reduce redundancy, senses are derived in context: with type coercion, the semantic type of an object is changed to suit the predicate (e.g., event interpretations of static objects when used with certain verbs); in contrast, with co-compositionality, the interpretation of predicates adapts to that of the arguments (e.g., creation-event interpretation of verbs when used with certain objects).

2.3.3 Analysis of Dictionary Definitions

Around 1980, a trend began towards building more realistic applications for natural language processing. Earlier work, in addition to being restricted to specialized domains, generally dealt with limited lexicons. Therefore, the analysis of machine-readable dictionaries (MRD's) became a popular way to overcome this limitation. The initial approaches concentrated on using the information explicitly provided, such as grammatical codes, with the exception that definitions were analyzed to establish the *is-a* hierarchies that were implicitly specified for the terms defined. Much work was done with LDOCE, partly because of favorable research licensing but mainly due to its *controlled vocabulary* of defining terms and its use of explicit grammatical codes. For a good survey of early LDOCE-related research, see the collection of papers in (Boguraev and Briscoe, 1989). This illustrates some of the difficulties commonly encountered, such as format errors in the typesetting and inconsistencies in the definitions. See (Wilks et al., 1996) for a comprehensive survey of work on MRD-related research, including good discussions on its historical development and on the philosophical issues involved.

The main contribution of Amsler's (1980) thesis is the development of procedures for the extraction of genus hierarchies from MRD's. This is a manually intensive process because the genus terms must be disambiguated by human informants. The noun and verb hierarchies extracted from the *Merriam-Webster Pocket Dictionary* were analyzed in depth. In addition, this work contains useful information on other aspects of analyzing dictionary definitions: 1) the analysis of the differentia descriptions commonly used in motion verbs, 2) suggestions for parsing dictionary definitions, 3) indications of what might be expected from a deep analysis (e.g., unraveling morphological relations), and 4) a technique for disambiguating dictionary definitions based on word overlap. A similar disambiguation technique was popularized by Lesk (1986); this is discussed later in Section 2.4.3. Note that these analyses establish a practical limit for what might be expected from automated analysis of dictionary definitions.

Alshawi (1989) discusses how to extract semantic information from dictionary definitions. This represents one of the first attempts to extract information outside of the genus terms. Pattern matching rules are applied to the definitions; an example follows, given here as extended regular expressions:

genus-identification:
    N .* (DET)? .* (ADJ)* (NOUN)?
predication-extraction:
    N (DET)? (ADJ)* (NOUN)* NOUN THAT-WHICH ⟨VERB-PRED⟩

Markowitz et al. (1986) illustrate several common patterns used in dictionary definitions. Some are specifically used to resolve the genus relationships given uninformative genus headwords such as 'any.' This is the empty-head problem noted by Bruce and Guthrie (1991). One common pattern for these is *Any-NP* or *Any-of-NP*, in which the genus is given by the NP element. There is a special case of this pattern for definitions of terms in biology:

> x: any of / [*modifier*] *taxon* (*formal name*) / of [*modifier*] *superordinate* / *attributes*
>
> grass: any of a large family (Gramineae) of monocotyledonous mostly herbaceous plants ...

Jensen and Binot (1987) also use pattern matching over definitions to determine whether certain relations hold (e.g., *instrument* and *part-of*); however, they perform matching over the output from a parser rather than just using string matching over words or parts of speech. This extraction is in support of a system for resolving prepositional phrase attachment. When making attachment decisions, they first check whether the definition of the preposition's complement has a pattern indicative of one of the relations considered. They then check for linkages to the headword of the governing constituent for the attachment, using the hierarchy of the definition genus terms.

Wilks et al. (1989) describe three different methods for analyzing LDOCE. All deal with aspects of converting the information in LDOCE into a lexical knowledge base. The first method is based on co-occurrence analysis of the controlled vocabulary usage in the definitions and examples. Using Pathfinder, reduced networks are produced showing the connectivity of related terms. Comparisons of the co-occurrence-based semantic relatedness scores versus human ratings show high correlations. The second method is based on bootstrapping a lexicon from a handcrafted lexicon for a subset of the controlled vocabulary. The third method, called the Lexicon-Producer, creates lexical entries based on the explicit information in the online version of LDOCE (e.g., grammar code, semantic restriction "box code", and subject code), as well as from pattern matching over parses of the definition to yield the genus term, basic features (e.g., modifiers), and some functional properties (e.g., *used-for*). Two methods are described for using this information. One is the Lexicon-Consumer, which parses text using the word-sense frames from the Lexicon-Producer. The other is the system of collative semantics, which is designed for producing mappings between sense frames to capture their relatedness.

Slator and Wilks (1987) sketch out an approach for deriving rich lexical entries from the information present in LDOCE, augmenting the Lexicon-Producer. The definitions are parsed to extract information from the differentia.

Cause	Domain
Hypernym	Location
Manner	Material
Means	Part
Possessor	Purpose
Quasi-hypernym	Synonym
Time	TypicalObject
TypicalSubject	User

Table 2.3: ***Relations extracted by Vanderwende's system.*** Adapted from (Richardson, 1997, Table 2.1).

The parse tree is added to the entry as well as case information derived via pattern-matching rules. A preliminary investigation of the patterns in LDOCE suggests that the case usage is fairly uniform. Guo (1995a) later fleshed out the bootstrapping process mentioned above as part of his thesis work. Lexical entries are created by parsing the LDOCE definitions, guided by the manually encoded preference knowledge for a subset of the controlled vocabulary. These consist of thematic-style relations for pairs of word senses. This information is used primarily for word-sense disambiguation of the definition text. Analysis of the definitions and example sentences yields further preference information; and, inductive machine learning techniques are used to generalize these via the genus hierarchy to cover a larger number of cases.

Vanderwende (1996) extracts detailed semantic information from LDOCE, building upon the work of Jensen and Binot (1987), again that uses a general-purpose parser rather than string matching (e.g., over words or parts of speech). Subcategorizations are only used to guide the parse, not to rule out potential parses, which is important because definitions incorporate ellipsis more so than normal text. The following is a typical rule from her system (Vanderwende, 1996, p. 193):

> LOCATION-OF pattern: if the hypernym is post-modified by a relative clause which has as its relativizer *where* or a wh-PP with the preposition *in, on*, or *from*, then create a LOCATION-OF relation with the head of the relative clause as the value.

In addition to extracting thematic roles, similar rules are used to extract functional information, such as the underlying subject and object for embedded verbals. The full set of relations extracted is shown in Figure 2.3.

Barrière's (1997) thesis illustrates how to acquire semantic knowledge from a dictionary written for children, in particular the *American Heritage First Dictionary* (AHFD). There are four basic steps in her process: parsing the definitions using a general grammar; transforming the parses to conceptual graphs; disambiguating the lexical relations in the conceptual graphs; and, combining the conceptual graphs from different definitions. Her parser uses a simple context-free grammar with some customization to dictionary definitions, such as the use of a "meaning verb" category. The conversion into conceptual graphs is a surface-level transformation from the parse tree into the conceptual graph notation. Some rules are more general than required for dictionary definitions to account for the AHFD's typical-usage sentences (e.g., "ash is what is left ...").

Barrière uses semantic relation transformation graphs (SRTG's) to extract relations from the initial conceptual graph representation resulting from the shallow parses for the dictionary entries. Some rules are quite specific and lead to unambiguous semantic relations; others are mainly heuristics about plausible interpretations. Table 2.4 lists the relations extracted by her system. As can be seen, some of these are special purpose (e.g., *home* as in "a hive is a home for bees"). A sample rule from her system follows:

Name: PART-OF
Description: part of an object
CG Representation: (part-of)

Sample definitions:
an arm is a part of the body
pines have needles on their branches

Before:
[something:A]←(agent)←[be]→(object)→[part]→(of)→[something:B]

[something:B]←(agent)←[have]→(object)→[something:A]

SRTG:
[something:B]→(part-of)→[something:A]

Nastase and Szpakowicz (2003) use Longman's dictionary to augment WordNet with noun-verb relatedness relations (e.g., derived from). They take advantage of LDOCE's controlled vocabulary in order to establish connections between the noun and related verb. Word sense disambiguation of the definitions is needed prior to establishing connections with WordNet, and this is done

About	Accompaniment	Act
Agent	As	Attribute
Cause	Content	Direction
During	Event	Experiencer
Frequency	Function	Goal
Home	Instrument	Intention
Like	Location	Manner
Material	Method	Modification
Name	Object	Obligation
Opposite	Path	Possession
Process	Recipient	Result
Sequence	Synonymy	Taxonomy
Transformation		

Table 2.4: ***Relations extracted by Barrière's system.*** Based on transformation rules in Appendix E of (Barrière, 1997).

via a simple word overlap algorithm similar to Lesk's (1986) approach. In addition, the relation type that holds between the noun and verb is inferred using classifiers induced over tagged examples.

The Extended WordNet (XWN) project represents one of the most ambitious attempts at extracting differentia from dictionary definitions (Harabagiu et al., 1999; Moldovan and Rus, 2001). The main goal is to transform the definitions into a logical form representation suitable for drawing inferences, such as for question answering; in addition, the content words in the definitions are being disambiguated with respect to the WordNet sense inventory (i.e., synsets). Given the open-ended nature of the task, they use a logical form that is closer to the surface-level representation than to deep semantics. For example, there will be predicates for each of the content words in the definition, as illustrated for 'supporter':

supporter: a person who backs a politician
$\Rightarrow$ [person:n(subj1) & back:v(e1,subj1,obj2) & politician:n(obj2)]

In addition, there will be separate predicates for prepositions, as well as for some other functional words (e.g., conjunctions). They achieve high precision in the transformation into logical form by concentrating on the commonly occurring grammar rules that occur in their parses (Rus, 2001; Rus, 2002). For

these cases, manually encoded transformation rules are developed, as in the following one for handling past participles:

$$\text{NP} \rightarrow \text{NP VP} \quad \Rightarrow \quad \text{noun(obj2) \& verb(e, subj1, obj2).}$$

Barnbrook's (2002) definition analysis is more in the vein of language exploration, in particular for the genre of definitions for language learners. He analyzes definitions from *Collins Cobuild Student Dictionary* (CCSD), a simplified version of *Collins Cobuild English Language Dictionary*. In the Cobuild dictionaries, definitions use complete sentences incorporating the headword. In addition, information about grammatical function and usage pragmatics are indicated implicitly in the definition rather than explicitly using various typographic conventions (e.g., grammar codes and usage labels). Therefore, his grammar was developed to account for this non-traditional definition style, so that the definition proper can be isolated correctly from the rest of the sentence constituents. As an illustration, about 24% of the definitions in CCSD are defined using initial 'when' clauses containing the headword. For example,

> When a country liberalizes its laws or its attitudes, it makes them less strict and allows more freedom.

Such cases are used mostly for verbs with the 'when' clause setting the background for the action description.

There have been a few papers criticizing work on extracting information from machine-readable dictionaries. Amsler (1995) suggests that dictionaries are perceived by computational linguists as being more definitive than they actually are. However, there are problems due to lack of uniformity in the quality of different dictionaries. He further notes that the information might not be suitable for a broad range of uses. Ide and Véronis (1993) provide more details on uniformity issues, particularly with respect to the genus hierarchies. They point out how differentiating relations are included haphazardly. For instance, in the definitions of 'abricot' (apricot) and 'péche' (peach) in three different French dictionaries, only peach is described as having a hard pit. These criticisms highlight that it will not be sufficient to rely solely on dictionaries as a basis for a knowledge base.

2.4 Supporting Areas

This thesis covers areas outside of lexical semantics as well. A brief review of the some of the previous work is included here to provide background on some of the material that might be unfamiliar.

2.4.1 Semantic Relatedness

Richardson (1997) discusses how to determine semantic relatedness based on the information extracted from MRD's, incorporating relations extracted using Vanderwende's (1996) techniques. This forms the basis for Microsoft's MindNet system (Richardson et al., 1998), a semantic network with weighted semantic links between arbitrary words and named entities. All the relations extracted from the same definition are first grouped into the same structure, with backward links made explicit in the network. To alleviate problems due to ambiguous words, paths are generally restricted to occur within a single structure. Extended paths are possible, but there is a penalty based on the frequency of the joining word. The highest weighted path between two words can be used to determine the relatedness of the words.

Hirst and St-Onge (1998) determine semantic relatedness based on the WordNet links between words. Strong relatedness is assigned if the words occur in the same synset or if they are in a sibling relationship. For medium relatedness, specific patterns for WordNet have been determined. For example, an upward direction is not allowed following a downward direction segment. In addition, only one change of direction is permitted.

2.4.2 Relation Weighting

Richardson (1997) also developed a novel procedure for weighting semantic relationships, using notions derived from the weighting of terms in information retrieval by the combination of term frequency (TF) and inverse document frequency (IDF), which is referred to as TF*IDF. Specifically, a term's weight is proportional to its overall frequency but inversely proportional to the number of documents it occurs in. This scheme is adapted to weighting relations derived from definitions by considering the set of definitions for the same word as a document. He uses semantic relations in place of terms, so the frequencies are those for the relational tuples. In addition, in place of TF*IDF, he uses a technique, called *averaged vertex probability*, that combines frequency

scaling and probability smoothing. Frequent relations are scaled back whenever the frequency exceeds that of the vertex of a Zipfian hyperbolic function.[3]

2.4.3 Word-sense Disambiguation

Many distinct approaches have been developed for disambiguating words in context (Ide and Véronis, 1998). These differ in the amount of training data that is needed beforehand and in the range of words that are targeted. Supervised approaches are quite precise but they only target a limited number of words, specifically those for which there are sufficient annotations on the senses that occur in text. Unsupervised approaches do not require annotations for the words to be disambiguated, so they can be applied to all words in the text, although with reduced precision. There are also hybrid approaches that use a supervised approach to tag the senses for which there are training data available and then apply heuristics to determine senses of words related to those already tagged.

2.4.3.1 Supervised WSD

The standard approach to statistical WSD is based on example-based learning over word-sense annotations (Ng and Lee, 1996; Bruce and Wiebe, 1999). (For an overview of example-based learning, see Appendix A.) Figure 2.2 shows sample annotations for 'circuit.' Prior to using machine learning to induce classifiers from such annotations, word-sense annotations must be converted into a tabular format with one row per example instance and one column for each distinct feature used to describe the instance, as well as a column for the instance classification (i.e., the word-sense from the annotation).

Figure 2.3 shows features that are commonly used in word-sense disambiguation. The subscripted features are actually a series of related features. For example POS_{-i} indicates i features for parts of speech for the i words preceding the target word, where i is typically 2 or 3. Similarly, $Word_{+i}$ indicates i separate features to represent the i words following the target word. The last group of features ($WordColl_s$) is for collocations, which provide important clues for word-sense disambiguation. Collocational features are typically binary and indicate the presence of a context word that is strongly associated with a particular sense of the word.

[3] Zipf's law states that term frequency is inversely proportional to rank; for example, the third most-common term has one-third the frequency of the first. The curve plotting this relationship can be viewed as the top half of a hyperbola (rotated 45 degrees).

⟨wf sense=5⟩Circuits⟨/wf⟩ are normally flown with climb or take-off flap at eighty knots, reducing to seventy with landing flap on final approach.

This means that there are only half as many samples in the ⟨wf sense=1⟩circuit⟨/wf⟩ as there are delaying stages.

This term is derived from the fact that the way in which these ⟨wf sense=1⟩circuits⟨/wf⟩ operate is roughly analogous to buckets of water being passed along a human chain (as in the old method of fire fighting).

So are the reports that have flourished on the LA gossip ⟨wf sense=4⟩circuit⟨/wf⟩ - Kilmer is going overboard; Kilmer thinks he is Jim Morrison; Kilmer has it written into his contract that everybody has to address him as Jim.

Figure 2.2: ***Sample word-sense annotations for 'circuit' from Senseval II.***

Morph:	morphology of the target word
POS$-$i:	part-of-speech of *i*th word to left
POS$+$i:	part-of-speech of *i*th word to right
Word$-$i:	*i*th word to the left
Word$+$i:	*i*th word to the right
WordColl_s:	occurrence of word collocation for sense *s* in context

Figure 2.3: ***Typical features for supervised word-sense disambiguation.***

Supervised WSD currently is only feasible for a limited number of target words (e.g., the "lexical-sample task" in SENSEVAL). Providing sufficient annotated training data for unrestricted word-sense disambiguation (e.g., the "all-words task" in Senseval), would require a large corpus of sense-tagged data for all content words. No existing corpus meets this requirement. For instance, although a quarter of the one million word Brown corpus was sense-tagged by the WordNet project members (Miller et al., 1994), this only covers about 15% of the senses for the 45,000 word types that were in WordNet (out of 120,000 distinct word types).

Fortunately, this situation might improve in the near future. For instance, the OpenMind project is aiming to produce a large-scale corpus with a broad variety of content words tagged against the WordNet sense inventory (Chklovski and Mihalcea, 2002). This is an all-volunteer effort, in order to circumvent the traditional high cost of producing annotations. Multiple taggings for the same word occurrence are used as a way to ensure better quality. OpenMind has annotated about 25,000 distinct sense occurrences per year. These annotations, along with the annotations produced for the biannual Senseval conferences, will likely make broad-covered supervised WSD viable in about ten years.

Supervised approaches to WSD rely mainly on collocations that co-occur significantly with the sense in the training data, because clue words that only occur once are considered unreliable. Attempts at using dictionary definitions to augment these clue words have run into complications. Although some of the words occurring in dictionary definitions are often quite indicative of a sense, there is no straightforward way to filter out the unrelated defining words that inevitably occur. For example, Veenstra et al. (2000) use definition clue words for supervised WSD just to supplement standard word collocations. The approach described in this thesis using related-word collocations illustrates one way to address the insufficient data problem dealing with definition clue words. By applying conditional probability tests, only those definitional words related to words that co-occur frequently with the sense are considered. See Section 5.2.1.2.

2.4.3.2 Unsupervised WSD

Given the limitations of supervised WSD, unsupervised approaches are more suitable as a general mechanism for WSD. A simple but effective method is the definition word-overlap approach developed by Lesk (1986). The sense selected is the one whose definition has the most overlap of content words with the sentential context for the word to be disambiguated.

Cowie et al. (1992) extend the idea by using simulated annealing to optimize a configuration of word senses simultaneously in terms of degree of word

overlap. Véronis and Ide (1990) develop a neural network model to overcome another limitation of word-overlap approaches, which only address pairwise dependencies. Using dictionary definitions, they construct a network where there is a link from a word node to nodes for each of its senses and links from each of the sense nodes to the words used in the definition. By activation through the neural network, longer-distance dependencies are addressed. Their model introduces noise by utilizing links from senses to words, and meaningful lexical relations are not distinguished from incidental ones. The probabilistic spreading activation approach discussed in Chapter 5 improves upon word overlap in several respects: for example, it provides differential weighting of the words based on semantic relatedness; and, by using Bayesian networks (Pearl, 1988), evidence is combined in a sound manner. Rosenzweig developed a recent version of the Lesk algorithm incorporating TF-IDF weighting for the overlap terms. This performed quite well in SENSEVAL I (Kilgarriff and Rosenzweig, 2000).

Nastase and Szpakowicz (2001) present a variation on word overlap that exploits the structure of WordNet in order to disambiguate entries in Roget's Thesaurus. In particular, they include word overlap among definitions for hypernym and hypernym synsets, as well as other related synsets. Word-overlap is useful for a variety of other tasks as well, although in general it is more suitable as a fallback rather than the main approach. For example, O'Hara et al. (1998) use word-overlap heuristics to augment their main structural heuristics used in aligning the Mikrokosmos ontology with WordNet.

Sussna (1993) minimizes pairwise distance among senses in a semantic network based on WordNet, using a weighting scheme that accounts for both fan-out and depth in the hierarchy. Of the approaches discussed here, his is most similar to the use of an analytical component in the hybrid empirical/analytical approach discussed in the applications chapter (Chapter 5), which is based on Wiebe et al. (1998b). However, he uses a symmetric weighting scheme to model similarity among senses, and he bases symmetry on the shortest available path. Wiebe et al. support asymmetric weighting and incorporate all paths in the similarity measure.

A drawback to the word-overlap approach is that it only accounts for words used in the definitions or examples associated with particular word senses. Yarowsky (1992) developed an approach that also incorporates word collocations that are associated with particular thesaural categories. He uses corpus analysis to see which words are generally indicative of each of the 1,000+ categories in Roget's Thesaurus. A simple Bayesian classifier is used to select the category that receives the highest collocational support given the words in the sentential context for the word to be disambiguated. To integrate this with WSD using the WordNet distinctions, the resulting category can be mapped

into WordNet and the closest synset for the target word to the category can be chosen as the sense. Figure 2.4 shows the revised algorithm.

1. Train Roget classifier over corpus

   N = total number of words
   N_{cat} = number of words associated with Roget category *cat*
   $\mathrm{freq}(\mathrm{word})$ = number of times *word* co-occurs in corpus
   $\mathrm{freq}_{cat}(\mathrm{word})$ = number of times *word* co-occurs with any word in *cat*

   $$P(cat) \simeq \frac{N_{cat}}{N}$$

   $$P(word) \simeq \frac{\mathrm{freq}(word)}{N}$$

   $$P(word|cat) \simeq \frac{\mathrm{freq}_{cat}(word)}{N_{cat}}$$

2. Disambiguate words by finding category with most support

   $$P(cat|context) = \sum_{w \in context} \log(\frac{P(w|cat) \times P(cat)}{P(w)})$$

3. Map best category into WordNet and find closest synset for target word.

Figure 2.4: ***Word-sense disambiguation using Roget-based classifier.*** Steps 1 and 2 are based on (Yarowsky, 1992). Step 3 is an extension for WSD using WordNet distinctions.

2.4.3.3 Semi-supervised WSD

Since supervised systems do achieve the best performance when there is training data available, it makes sense to incorporate them when possible. One simple scheme would be to use a hierarchy of taggers, using the supervised classifiers if there is sufficient data available and then falling back to unsupervised classifiers if not. Alternatively, hybrid systems could be developed that exploit the training data used by supervised systems while retaining flexibility for handling other words. For example, Mihalcea and Moldovan (2001) used this to achieve the highest performing WSD system in the all-words task for Senseval II. Figure 2.5 shows the heuristics used by their system.

2.4.4 Class-based Collocations

Class-based features are often used to address sparse data problems in training data. A simple type of class-based feature uses part-of-speech labels

1. Apply named-entity tagging.

   For example, a person's name implies $person_{\text{BEING}}$.

2. Tag monosemous words.

3. Assign contextual positional bigrams the same sense in SemCor.

   If all occurrences of $W_{-1}W$ and WW_{+1} from the text have the same sense for W and occur more than N times then tag as that sense.

4. Check overlap of noun-contexts for each noun sense with current text.

   A noun-context is the set of hypernym words for a given sense along with the words within ten words of the tagged sense in SemCor.

5. Tag synonyms for words already disambiguated: words at semantic distance 0 (i.e., in same synset).

6. Tag words at semantic distance of 1 from disambiguated words.

7. Tag synonyms among non-disambiguated words with sense for synset.

8. Tag non-disambiguated words at semantic distance of 1 from synset.

Figure 2.5: ***Heuristics for semi-supervised WSD using bootstrapping.*** Adapted from (Mihalcea and Moldovan, 2001). This forms the basis for one of the systems used in the preparation of Extended WordNet (Novischi, 2002). *SemCor* refers to the word-sense annotations (semantic concordance) that comes with WordNet.

(Charniak, 1993), where words are replaced by their grammatical class. With hypernym collocations, semantic classes are used instead. Scott and Matwin (1998) use WordNet hypernyms for classification, in particular topic detection. They include a numeric density feature for each synset that subsumes words appearing in the document, potentially yielding hundreds of features. Mihalcea (2002) shows how hypernym information can be useful in deriving clues for unsupervised WSD. Patterns for co-occurring words of a given sense are induced from sense-tagged corpora. Each pattern specifies templates for the co-occurring words in the immediate context window of the target word:

$$\langle \text{word-stem, part-of-speech, synset-ID, hypernym-synset-ID} \rangle$$

where any of the components in the pattern can be unspecified. As an example,

⟨*, noun, *, *room*$_{\text{AREA}}$⟩ ⟨'door', noun, *door*$_{\text{BARRIER}}$, *⟩

would match "kitchen door" and "bedroom door."

Other work shows how to use WordNet in deriving traditional collocations. For example, Pearce (2001) combines WordNet synonym information with BNC corpus analysis when extracting collocations.

2.4.5 Relation Disambiguation

Until recently, preposition classification has received little attention, especially with respect to broad coverage rather than just special purpose usages. Halliday (1956) did some early work on this in the context of machine translation. Later work in that area addressed the classification indirectly during translation. In some cases, the issue is avoided by translating the preposition into a corresponding foreign function word without regard to the preposition's underlying meaning (i.e., direct transfer). Other times an internal representation is helpful. Trujillo (1995) discusses these issues in depth. He favors a transfer approach at the level of an internal representation for lexemes, rather than at a surface level as traditionally done. Japkowicz and Wiebe (1991) illustrate the deep meaning approach in using conceptual structures to account for the differences in how prepositions are used to conceptualize objects. In story understanding work, preposition classification often is implicitly handled in the conversion of text to case structures (Schank, 1973), as is also the case for text extraction (Lehnert et al., 1992). Taylor (1993) discusses general strategies for preposition disambiguation using a cognitive linguistics framework and illustrates them for 'over.' There has been quite a bit of work in this area but mainly for spatial prepositions (Zelinsky-Wibbelt, 1993).

There is currently more interest in this type of classification. Litkowski (2002) presents manually derived rules for disambiguating prepositions, in particular for 'of.' Srihari et al. (2001) present manually derived rules for disambiguating prepositions used in named entities, but the disambiguation is more oriented to delineating the constituents of the prepositional phrase rather than determining the type of relation.

Gildea and Jurafsky (2002) classify semantic role assignments using the annotations from FrameNet, for example, covering all types of verbal arguments. They use several features derived from the output of a parser, such as the constituent type of the phrase (e.g., NP) and the grammatical function (e.g., subject). They include lexical features for the headword of the phrase and the predicating word for the entire annotated frame. They report an accuracy of 76.9% with a baseline of 40.6% over the FrameNet semantic roles.

Blaheta and Charniak (2000) classify semantic role assignments using the annotations from Treebank. They use a few parser-derived features, such as the constituent labels for nearby nodes and part-of-speech for parent and grandparent nodes. They also include lexical features for the head and alternative head (since prepositions are considered as the head by their parser). They report an accuracy of 77.6% over the form/function tags from the Penn Treebank with a baseline of 37.8%.[4] Van den Bosch and Bucholz (2002) also use the Treebank data to address the task of assigning function tags to arbitrary phrases. For features, they use parts of speech, words, and morphological clues. Chunking is done along with the tagging, but they only present results for the evaluation of both tasks taken together; their best approach achieves 78.9% accuracy (at 79.1% recall).

Nastase and Szpakowicz (2003) assign relation types to the noun-verb relationships inferred in WordNet. Separate binary classifiers are used for each of 20 different relations, but no performance results are given. Their features are based just on the words that occur in the definitions; as they note, this doesn't support generalizations. They use a limited training set with less than 300 total examples; and, the use of binary classifiers leaves open the problem of conflict resolution, such as if more than one of the classifiers returns a positive result.

Liu and Soo (1993) present a heuristic approach for relation disambiguation relying upon syntactic clues as well as occurrence of specific prepositions. They assign roles to constituents of a sentence from corpus data provided that sufficient instances are available. Otherwise, a human trainer is used to answer questions needed by the system for the assignment. They report an 86% accuracy rate for the assignment of roles to verbal arguments in about 5,000 processed sentences. Most recently, there have been two recent workshops featuring competitions for semantic role tagging (Litkowski, 2004; Carreras and Màrquez, 2004).

There has been more work in prepositional phrase interpretation dealing with structural disambiguation for prepositional phrase attachment (Dalgren and McDowell, 1986; Hindle and Rooth, 1993; Kayaalp et al., 1997). In a knowledge-based approach, Dalgren and McDowell (1986) develop heuristics for resolving prepositional phrase attachment. These heuristics incorporate taxonomic information of the prepositional objects, from a manually encoded knowledge base. An example of one of their rules follows:

[4]They target all of the Treebank function tags but give performance figures broken down by the groupings defined in the Treebank tagging guidelines. The baseline figure shown above is their recall figure for the 'baseline 2' performance.

```
at-rule:
if abstract(Object) or place(Object) then
    s_attach(PP)
else
    np_attach(PP)
```

Section 4.3.2.1 later illustrates that ‘at’ is used in a temporal sense in an abstract context. Temporal interpretations are more likely to apply to the sentence as a whole rather than just the modified object. Thus, this approach automatically acquires some of the knowledge implicitly assumed by these rules. It will be interesting to see whether such attachment rules can be automatically acquired, such as combining the corpus-based structural disambiguation approach of Hindle and Rooth (1993) with the relation classification approach discussed in Chapter 4.

2.5 Conclusion

This chapter has reviewed some of the literature related to lexical acquisition with an emphasis on lexical semantics. Linguistic work on case roles (Fillmore, 1968; Jackendoff, 1990) provides practical approaches for representing important semantic relations. Other work in linguistics serves to clarify the types of semantic relations to be found in lexicons (Cruse, 1986; Mel’čuk and Polguere, 1987). Semantic role inventories are an integral part of this work, as discussed in Section 4.2. Lexicography also provides insights into what needs to be represented and what can be expected of traditional dictionaries (Kilgarriff, 1997; Landau, 2001). Although not directly addressed here, these issues might affect extensions to the research. Similarly, the knowledge representation issues addressed in computational semantics (Wilks, 1975b) and ontological semantics (Nirenburg and Raskin, 2004) will be important for adapting this thesis work for more general tasks.

A variety of automated acquisition approaches was presented. Corpus analysis is often used for this, in particular via lexical associations (Hindle and Rooth, 1993). Lexical rules are commonly used for acquisition involving highly productive types of patterns, such as for the count/mass distinction and derivational morphology (Briscoe et al., 1995; Viegas et al., 1996). Analysis of dictionary definitions complements this in addressing the idiosyncratic information associated with particular words. Pattern matching is usually applied to the definition using rules tailored for specific semantic relationships (Vanderwende, 1994; Barrière, 1997). Chapters 3 and 4 build upon such dictionary analysis work by making it more corpus-driven, particularly in the relation disambiguation process.

CHAPTER 3
DIFFERENTIA EXTRACTION

This thesis is motivated by the desire to have more conceptual distinctions in semantic lexicons and in knowledge bases in general. The first step in this process involves extracting the distinguishing relations (*differentia*) indicated in dictionary definitions. WordNet is used both as the source of definitions and as the semantic lexicon to be augmented. This first step involves determining the important surface-level relations present in the definitions and is implemented via a broad-coverage dependency parser. Dependency parsers stress the connections between words rather than the structural configuration of syntactic categories, which is typical of traditional phrase-structure parsers (Sleator and Temperley, 1993; Jurafsky and Martin, 2000).

Definitions taken from a dictionary are first preprocessed to make them more suitable for parsing (e.g., conversion to complete sentences). Then the sentences are parsed to produce a list of low-level syntactic relations among the words. After parsing, the relations are postprocessed to convert them into more traditional grammatical relations. The postprocessing also makes the relations centered on function words, such as prepositions. Certain relations will be more salient for the term being defined than other relations extracted from the definition. A way to quantify this is included during an optional relation-weighting step. The relations are also converted into a format easier for humans to read.

This chapter is organized as follows. Section 3.1 presents an overview of WordNet. It also discusses manual annotations of a subset of the definitions, prepared during the preliminary stage of this work but not used by the extraction system. Section 3.2 discusses the definition parsing. This uses the Link Grammar parser (Sleator and Temperley, 1993), along with postprocessing support developed specifically for the extraction system. Section 3.3 discusses the relation extraction, including the use of cue validities for estimating relation salience. Section 3.4 closes with a summary of the extraction algorithm, along with an example illustrating the processing done at each stage.

3.1 Analysis of Definitions in WordNet

The dictionary used here as the source of definitions is WORDNET (Fellbaum, 1998).[1] WordNet incorporates aspects of a thesaurus as well as a dic-

[1]WordNet version 1.7.1 is used throughout unless otherwise noted. WordNet is freely available from Princeton. The database along with full documentation can be found at www.cogsci.princeton.edu/~wn.

tionary. Words are grouped into synonym sets called *synsets*, which serve as the underlying concepts referred to by words in the lexicon. For example, for the animal sense of 'dog,' the corresponding synset would be

{dog, domestic dog, Canis familiaris}.

Instead of listing the entire synset or using sense labels (e.g., *dog#1*), subscripted category names (e.g., dog_{CANINE}) are generally used here when referring to synsets with words that are ambiguous.

3.1.1 Structure of WordNet

Figure 3.1 shows some of the information given for the word 'dog.' The distinguishing feature of WordNet compared to traditional dictionaries is the use of explicit links among the synsets; for example, the '$\Rightarrow$' links in the figure are for WordNet's *hypernym* relation (same as *is-a*). Table 3.1 gives descriptions and usage statistics of all the relations in WordNet. These explicit relations form the basis for a knowledge base and rudimentary ontology (Mahesh and Nirenburg, 1995; Sowa, 1999).

As a dictionary, WordNet is somewhat broader in scope than a learner's dictionary such as LDOCE, the *Longman Dictionary of Contemporary English* (Procter, 1978); but, it not as comprehensive as a college dictionary such as *Merriam Webster's Collegiate Dictionary* (Mish, 1996). WordNet covers the core lexicon of English, but also includes some scientific and technical terms. Table 3.2 shows some statistics on the number of entries in WordNet. The *Entries* column lists the number of words or phrases with distinct entries in the dictionary; this is the number dictionary publishers often highlight to indicate the size. *Senses* column refers to the total number of sense distinctions for all the entries (e.g., six for 'dog'). As with traditional dictionaries, the senses are numbered, but there are no further subdivisions (e.g., '1a'). Unlike traditional dictionaries, definitions are not given for the word senses but instead for the synsets (i.e., the synonym sets). Thus, the *Synsets* column refers to the number of underlying concepts, the targets for the senses. The sense versus synset distinction is not apparent in Figure 3.1 alone, but it can be seen when also considering the entry for 'cad' shown in Figure 3.2. Both entries incorporate the following synset:

{cad, bounder, blackguard, dog, hound, heel}

That is, sense 4 of 'dog' and sense 1 of 'cad' map into the same synset.

3.1.2 WordNet Definition Annotations

Manual annotations are commonly used in computational linguistics to provide insight into the genre of text being studied. They also are used to de-

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun dog
6 senses of dog
Sense 1
dog#1, domestic dog#1, Canis familiaris#1 – (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; “the dog barked all night”)
⇒ canine#2, canid#1 – (any of various fissiped mammals with nonretractile claws and typically long muzzles)

Sense 2
frump#1, dog#2 – (a dull unattractive unpleasant girl or woman; “she got a reputation as a frump”; “she’s a real dog”)
⇒ unpleasant woman#1, disagreeable woman#1 – (a woman who is an unpleasant person)

Sense 3
dog#3 – (informal term for a man; “you lucky dog”)
⇒ chap#1, fellow#1, feller#2, lad#1, gent#2, fella#1, blighter#2, cuss#2 – (a boy or man; “that chap is your host”; “there’s a fellow at the door”; “he’s a likable cuss”)

Sense 4
cad#1, bounder#1, blackguard#1, dog#4, hound#2, heel#3 – (someone who is morally reprehensible; “you dirty dog”)
⇒ villain#1, scoundrel#1 – (a wicked or evil person; someone who does evil deliberately)

Sense 5
pawl#1, detent#1, click#4, dog#5 – (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
⇒ catch#6, stop#10 – (a restraint that checks the motion of something; “he used a book as a stop to hold the door open”)

Sense 6
andiron#1, firedog#1, dog#6, dogiron#1 – (metal supports for logs in a fireplace; “the andirons were too hot to touch”)
⇒ support#10 – (any device that bears the weight of another thing; “there was no place to attach supports for a shelf”)

Figure 3.1: ***Definitions for the noun ‘dog’ in WordNet.***

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun cad

2 senses of cad

Sense 1
cad#1, bounder#1, blackguard#1, dog#4, hound#2, heel#3 – (someone who is morally reprehensible; “you dirty dog”)
    ⇒ villain#1, scoundrel#1 – (a wicked or evil person; someone who does evil deliberately)

Sense 2
computer-aided design#1, CAD#2 – (software used in art and architecture and engineering and manufacturing to assist in precision drawing)
    ⇒ software#1, software system#1, software package#1, package#3 – ((computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory; “the market for software is expected to expand”)

Figure 3.2: ***Definitions for the noun ‘cad’ in WordNet.***

Relation	Usage	Description
has-hypernym	88381	superset relation
is-similar-to	22492	similar adjective synset
is-member-meronym-of	12043	constituent member
is-part-meronym-of	8026	constituent part
is-antonym-of	7873	opposing concept
is-pertainym-of	4433	noun that adjective pertains to
also-see	3325	related entry (for adjectives and verbs)
is-derived-from	3174	adjective that adverb is derived from
has-verb-group	1400	verb senses grouped by similarity
has-attribute	1300	related attribute category or value
is-substance-meronym-of	768	constituent substance
entails	426	action entailed by the verb
causes	216	action caused by the verb
has-participle	120	verb participle

Table 3.1: ***Relation usage in WordNet.*** Based on analysis of database files for WordNet 1.7.

POS	Entries	Senses	Synsets
Noun	109195	134716	75804
Verb	11088	24169	13214
Adjective	21460	31184	18576
Adverb	4607	5748	3629
Total	146350	195817	111223

Table 3.2: ***Number of entries in WordNet by part of speech.*** Based on documentation accompanying WordNet 1.7.1 (i.e., WNSTATS manual page). The *Synsets* column gives the number of concepts and indicates the number of distinct definitions. (LDOCE in comparison has about 70,000 definitions).

velop example-based learning systems. Notably, these types of systems currently achieve the highest performance in word-sense disambiguation (Kilgarriff and Palmer, 2000b; Edmonds and Kilgarriff, 2002). One of the largest sets of such annotations was prepared by Singapore's Defense Sciences Organization (DSO) under the direction of Ng and Lee (1996). They tagged the senses for 190 common nouns and verbs occurring in parts of the *Wall Street Journal* and *Brown Corpus*, yielding 190,000 distinct annotations.

As part of this thesis work, the WordNet definitions for the words tagged in the DSO data were annotated to determine the semantic relations implicit in the definitions. Roughly 650 of the 1,500 definitions for these words were annotated, yielding about 2,400 tagged relation instances. Figure 3.3 shows a sample of the annotations. In the first case, the mechanism sense of 'action,' is defined in terms of the part category, which is quite generic. This is specialized via an attribute about being operational and the qualification regarding transmission of power.

Table 3.3 shows the most common semantic relations occurring in these annotations, along with a description of each. The next chapter will discuss other semantic role annotations, mainly those centered on thematic roles (e.g., verb arguments and adjuncts). Thematic roles are also included in these WordNet annotations (e.g., *location* and *source*); however, specialization-type relations are more commonly used here (e.g., *spec*, *qual*, and *concerning*). For example, four of the top six annotated relations in Table 3.3 deal with specialization. This reflects the focus of definitions on static descriptions rather than actions or situations in general. The emphasis on specialization relations in these annotations adds support that definitions are differential in nature. Note that Figure 3.3 and Table 3.3 also serve to illustrate the types of information the system attempts to extract from definitions. The steps in the process are discussed next.

Relation	Freq.	Description
genus	.274	category for concept (definition *genus*)
attr	.199	generic attribute
object	.133	affected object or event
spec	.107	specialization involving given type of participant
manner	.047	manner in which action is done
qual	.036	qualification of the genus category
subject	.026	generic actor for action described
example	.024	example for which the definition applies
alt-genus	.024	secondary category for concept
location	.017	physical or abstract location
purpose	.016	what an object is used for; why an action is done
result	.010	result produced by some action
means	.009	means by which an action or condition is achieved
concerning	.006	objective theme used for specialization
subject-attr	.006	attribute of sentence subject (actor)
action	.006	action in descriptive subordinating clause
agent	.005	agent performing an event
field	.005	domain indicated by usage label
source	.005	physical or abstract starting point
complement	.005	non-objective complement of a verb

Table 3.3: ***Common semantic relations from the definition annotations.*** *Freq.* is the relative frequency out of about 2,400 total instances.

action#4: the operating part that transmits power to a mechanism
    genus part
        attr operating
    qual transmits
        object power
        recipient mechanism

experience#1: the accumulation of knowledge or skill that results from direct participation in events or activities
    genus accumulation
        spec knowledge or skill
    qual results
        source participation
            attr direct
            containment events or activities

add#4: make an addition by combining numbers
    genus make
        object addition
    manner combining
        object numbers

keep#15: maintain by writing regular records
    genus maintain
    means writing
        object records
            attr regular

Figure 3.3: ***Sample of WordNet semantic relation annotations.*** Definitions for the 190 words targeted in DSO corpus were annotated at NMSU. See www.cs.nmsu.edu/~tomohara/wordnet-annotations for all of the annotations.

3.2 Definition Parsing

The approach to differentia extraction is entirely automated. This starts with the use of a broad-coverage parser to determine the syntactic lexical relations that occur in the sentence. Before parsing, however, the definitions must be preprocessed in order to minimize parse failures. For example, definition fragments are replaced with complete sentences. Afterwards, the low-level dependency relations are converted into higher-level grammatical ones.

3.2.1 Definition Preprocessing

Dictionary definitions are often given via sentence fragments that omit the *entry word* (i.e., word being defined). (Some learner's dictionaries now give definitions in complete sentences (Barnbrook, 2002), but this is not common practice.) For example, the definition for $lock_{\text{FASTENER}}$ is "a fastener fitted to a door or drawer to keep it firmly closed." Therefore, prior to running a general-purpose parser, the sentences are converted into complete sentences. Other preprocessing steps include the removal of example sentences and the removal of domain indicators (e.g., "(chemistry) two or more atoms ...").

Definitions for words of a given part of speech are usually given via phrases having the same part of speech. One reason for this is to aid in understanding contextual usages of the word being defined. When the same part of speech is used, the definition could be substituted for the entry word in a given sentence without affecting grammaticality (Landau, 2001).

At least 80% of the noun definitions in WordNet follow this pattern, based on analysis of output from a part-of-speech tagger and a simple phrase chunker. Table 3.4 shows the most common part-of-speech patterns occurring in the noun definitions. A similar analysis for the verb definitions shows that about 70% start with a verb. See Table 3.5 for the top occurring part-of-speech patterns. Note that it is likely that many of the cases starting with NP shown in the table (e.g., 'NP adverb') are due to erroneous part-of-speech assignments. Part-of-speech taggers tend to have trouble labeling words in fragments because the complete sentential context is not available. In addition, the WordNet definitions have not been through the amount of editing that definitions for commercial dictionaries go through.[2] Therefore, commercial dictionaries undoubtedly will have a higher percentage of category-conforming definitions as a result of better quality control (e.g., more emphasis on ensuring uniformity of definitions).

Table 3.6 shows the patterns that are used for forming definitional sentences, based on the grammatical type of the word being defined. The patterns are designed to form complete sentences from the definitional fragments while minimizing the introduction of extraneous semantic content. For nouns, the optional determiner is used whenever the word is a count noun (e.g., 'dog' in contrast to 'sand'). The determiner 'a' is used unless the word being defined starts with a vowel. Thus, for $lock_{\text{FASTENER}}$, the following sentence would be used for the parse: "A lock is *a fastener fitted to a door or drawer to keep it*

[2]The WordNet definitions were initially not included in the lexicon, because it was felt the synset groupings would be sufficient to determine the meaning intended for each sense (Miller, 1990).

Pattern	Freq.
NP	11785
NP verb preposition NP	2700
NP punctuation	1552
NP verb NP	1244
NP punctuation NP	1158
NP preposition verb NP	1005
NP determiner verb NP	758
NP pronoun verb NP	676
NP punctuation number punctuation number punctuation	624
verb NP	360

Table 3.4: ***Top part-of-speech patterns for WordNet noun definitions.*** 76,191 total definitions were analyzed.

Pattern	Freq.
verb preposition NP punctuation	605
verb preposition NP	385
NP punctuation	332
verb NP punctuation	327
NP	290
verb NP	271
verb NP preposition punctuation	265
verb adverb punctuation	225
verb adjective punctuation	154
NP adverb punctuation	132

Table 3.5: ***Top part-of-speech patterns for WordNet verb definitions.*** 13,406 total definitions were analyzed.

Part of speech	Sentence-completion template
noun	⟨optional-determiner⟩ ⟨word⟩ is ⟨definition⟩
verb	To ⟨word⟩ is to ⟨definition⟩
adjective	⟨word⟩ things are ⟨definition⟩
adverb	It occurs ⟨definition⟩

Table 3.6: ***Templates for definitional sentences.*** This only includes content words, as function words are not represented in WordNet.

Pattern	Freq.
verb preposition NP punctuation	773
verb NP punctuation	750
verb NP	449
verb preposition NP	341
preposition conjunction verb to-preposition NP	314
preposition conjunction verb to-preposition NP punctuation	298
preposition NP punctuation	296
adverb adjective punctuation	274
adjective preposition NP punctuation	272
adverb adjective	171

Table 3.7: ***Top part-of-speech patterns for WordNet adjective definitions.*** 18,700 total definitions were analyzed.

firmly closed." In contrast, for *anemia*$_{\text{DISEASE}}$, the result would be "Anemia is *a deficiency of red blood cells*."

For verbs, the definition is first changed if necessary to an infinitive phrase. The definition is then used as a verbal complement with the subject formed via an infinitive phrase with the entry word. For *delay*$_{\text{WAIT}}$ this yields the definitional sentence "To delay is to *act later than planned, scheduled, or required*." A better pattern might be just to convert the definition into the past tense and supply a dummy subject of 'it' (e.g., "It *act*ed *later than planned, scheduled, or required*"). This would make it easier to discard the parts of the parse structure that correspond to the sentential template and not the definition proper. However, this would require morphological support for recognizing or producing the past tense, which has numerous special cases. Future work will investigate producing better patterns, possibly incorporating different patterns based on the form of the definition fragment.

Adjectives and adverbs present more of a problem in the conversion to complete sentences. Tables 3.7 and 3.8 show the common part-of-speech patterns from the WordNet definitions. The adjective definitions often start with verbs in the past or present participle, which also serve as modifiers. The adverb definitions show a preponderance of prepositional phrases, with a special case being 'in a ⟨adjective⟩ manner' (779 cases). Although most modifiers use simple defining phrases, occasionally these are defined through a series of defining phrases indicating various aspects of the corresponding concepts. In WordNet, roughly 10% of the modifier definitions incorporate semicolons to specify additional aspects of the meaning. For example, 'matte' is defined as "not reflecting light; not glossy." Such constructions are turned into disjunc-

Pattern	Freq.
preposition NP punctuation	1592
preposition NP	198
preposition NP punctuation preposition NP punctuation	103
preposition determiner adjective conjunction NP punctuation	97
to-preposition NP punctuation	78
preposition determiner adverb NP punctuation	71
preposition determiner verb NP punctuation	65
adverb punctuation	41
verb NP punctuation	36
preposition NP conjunction to-preposition NP punctuation	21

Table 3.8: ***Top part-of-speech patterns for WordNet adverb definitions.*** 3,636 total definitions were analyzed.

tions. For adjectives, the definition is used as a predicate complement of a subject phrase with a dummy word (e.g., 'thing') modified by the adjective in question. For 'incredible' this yields "Incredible things are *beyond belief or understanding*." For adverbs, the definition is used as a post modifier in a generic "It occurs ..." sentence. For 'worthily' this yields "It occurs *in a worthy manner* or *with worthiness*."

Instead of converting the definitions into sentences, an alternative approach would be to customize the grammar of the parser to accommodate the definition fragments. However, this would involve a substantial amount of work, much of which might be specific to the parser being used. The definitions would still likely require some form of reformatting (e.g., for isolating punctuation). Thus, it is better to put more emphasis on preprocessing, which is less dependent on the parser implementation.

3.2.2 Dependency Parsing

Traditional parsers are based on phrase-structure rules that indicate how constituents of a sentence are decomposed (e.g., $S \rightarrow NP\ VP$). In contrast, a *dependency parser* stresses the connections among the words in the sentence (Manning and Schütze, 1999; Jurafsky and Martin, 2000). For example, the transitive verb 'open' would require connections to a subject and object, which are its *dependents* in the analysis. By emphasizing word-level relations over constituent-level ones, dependency parsers make it easier to determine the syntactic relationships involving the phrase headwords in a sentence.

The *Link Grammar Parser* (Sleator and Temperley, 1993) is used here for parsing the definitions. As an illustration, Figure 3.4 shows the parse for the definition of 'wine' in the alcohol sense: "fermented juice (of grapes especially)." This shows that the Link Parser outputs syntactic dependencies among punctuation and sentence boundary elements, as well as among words. The parse output includes word offsets. This is a minor customization to the parser added to facilitate relation extraction.

3.2.3 Parse Postprocessing

After parsing, a series of postprocessing steps is performed prior to the extraction of the lexical relations. For the Link Parser, this mainly involves conversion of the binary dependencies into relational tuples and the realignment of the tuples around function words. In addition, the part of speech specification is normalized in terms of a prefix rather than suffix. Note that the first parse produced by the parser is the one used for analysis: this is the simplest approach for resolving structural ambiguity. Alternatively, the parses could be analyzed to see which makes most sense in terms of lexical associations as sketched out later in the future work chapter.

The Link Parser uses quite specialized syntactic relations, so these are converted into general ones prior to the extraction of the relational tuples. The mapping for performing this conversion was created as part of this research. For example, the relation *A*, which is used for pre-noun adjectives, is converted into *modifies*. Table 3.9 shows the conversions for some of the common relations encountered.

The syntactic relationships are first converted into relational tuples using the following format:

⟨source-word, relation-word, target-word⟩

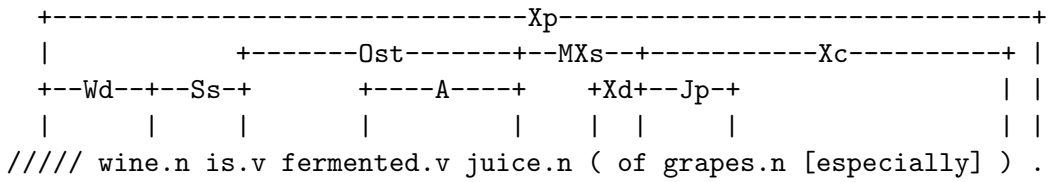
This conversion is performed by following the dependencies involving the content words, ignoring cases involving sentence boundary elements or punctuation. The first tuple extracted from the parse in Figure 3.4 would thus be ⟨n:wine, v:is, n:juice⟩. More specifically, the conversion checks for pairs of tuples involving the same function word (or linking verb) first as a target term and subsequently as a source term. These are collapsed into a single tuple that uses the function word for the grammatical relation:

(⟨*word1*, relation1, *function-word*⟩ and ⟨*function-word*, relation2, *word2*⟩)
$\Rightarrow$ ⟨*word1*, *function-word*, *word2*⟩

Definition sentence:

Wine is fermented juice (of grapes especially).

Parser output:



	/////	Xp	<---Xp---->	Xp	.
(m)	/////	Wd	<---Wd---->	Wd	wine.n
(m)	wine.n	Ss	<---Ss---->	Ss	is.v
(m)	is.v	O*t	<---Ost--->	Os	juice.n
(m)	fermented.v	A	<---A----->	A	juice.n
(m)	juice.n	MXs	<---MXs--->	MX	of
(m)	(	Xd	<---Xd---->	Xd	of
(m)	of	Xc	<---Xc---->	Xc	)
(m)	of	J	<---Jp---->	Jp	grapes.n
	.	RW	<---RW---->	RW	/////

Syntactic relationships:

Relationship	*Relation type*
⟨/////, Wd, 1. n:wine⟩	*connection to left boundary*
⟨/////, Xp, 10. .⟩	*sentence-ending period*
⟨1. n:wine, Ss, 2. v:is⟩	*singular subject*
⟨10. ., RW, 11. /////⟩	*connection to right boundary*
⟨2. v:is, Ost, 4. n:juice⟩	*object of verb*
⟨3. v:fermented, A, 4. n:juice⟩	*pre-noun modification*
⟨4. n:juice, MXs, 6. of⟩	*post modification*
⟨5. (, Xd, 6. of⟩	*preceding separator-punctuation*
⟨6. of, Jp, 7. n:grapes⟩	*preposition with plural object*
⟨6. of, Xc, 9.)⟩	*following separator-punctuation*

Figure 3.4: ***Link Grammar parse for*** **wine$_{\text{ALCOHOL}}$.** The *Relationship* column shows the tuple-based representation based on the default output of the Link Parser over the input (shown above). The syntactic relationships include word offsets, produced via a simple extension to the parser.

Link Rel.	General Rel.	Description
A	modifies	pre-noun adjectives to following nouns
AN	modifies	noun-modifiers to following nouns
Am	comparative	used with comparatives
B_	subject-of	noun to verb [relative clause]
CC	joined-with	clauses to following coordinating conjunctions
C_	subject-of	subject of clause
D	determiner-of	determiners to nouns
E	modifies	verb-modifying adverbs which precede the verb
I	modal-verb	verbs with infinitives
ID_	idiom	words of idiomatic expressions
If	modal	verbs with infinitives
J	prep-obj	prepositions to their objects
K	particle	verbs with particles
MVa	modified-by	verbs and adjectives to modifying post-phrases
MX_	modifies	modifying phrases to preceding noun
O	has-object	transitive verbs to their objects
O_	has-object	other types of grammatical objects
R	modified-by	nouns to relative clauses
RS	subject-of	relative pronoun to the verb
RW	to-wall	right-hand wall to the left-hand wall
S_	subject-of	subject nouns to finite verbs
TO_	modal	verbs and adjectives to the word ‘to’
W_	to-wall	marks beginning or end of sentence (the *wall*)
X_	is-punctuation	used with punctuation
YS	is-possessive	nouns to the possessive suffix

Table 3.9: ***Mapping from Link Parser relation types into general ones.*** *Link Rel.* are Link Parser relations, and *General Rel.* are the general relations resulting from the mapping. Relation types ending with underscores (e.g., *MX_*) stand for a series of relations starting with that prefix (e.g., *MX*, *MXs*, and *MXp*, where ‘s’ and ‘p’ indicate modification of a singular and plural noun, respectively). Detailed documentation on the relation types can be found at http://hyper.link.cs.cmu.edu/link/dict/summarize-links.html.

⟨1. n:wine, 2. v:is, 4. n:juice⟩
⟨3. v:fermented, *modifies-3-4*, 4. n:juice⟩
⟨4. n:juice, 6. of, 7. n:grapes⟩

Figure 3.5: ***Initial lexical relations for*** **wine$_{\text{ALCOHOL}}$.**

Certain types of dependencies are preserved by converting the syntactic relationships directly into a relational tuple involving a special relation-indicating word (e.g., 'modifies'). For the wine example, this yields the tuple ⟨v:fermented, *modifies*, n:juice⟩. The result of the initial conversion for the wine example is shown in Figure 3.5. Offsets are also incorporated into the relation name (e.g., *modifies-3-4*), as shown in the figure. These offsets are used to ensure that the placeholder relation words are unique (for the purposes of relation disambiguation).

3.3 Deriving Lexical Relations from the Parses

The relational tuples shown in Figure 3.5 form the basis for the lexical relations extracted from the definition parse. That is, the final part of the parser output post-processing determines the initial set of relations extracted from the definition. The remaining steps in the process account for structural ambiguity in the parses and for assigning weights to the relations that are extracted.

3.3.1 Attachment Resolution

Certain parsing problems arise during the extraction process. An important one is how to handle phrase attachment, in particular for prepositional phrases. This is not an emphasis of the research, so this is currently handled by only considering the first parse produced by the parser. Two alternatives have been considered. The first handles structural ambiguity resolution by having the parser return multiple parses and selecting the attachments that occur most often. In this case, the relations are weighted by the percentage of the times they occur in all of the parses. This might lead to incompatible relations (e.g., crossing dependencies in the parse), so it is left for future work. The other alternative uses class-based lexical associations and is sketched in the future work chapter.

3.3.2 Assigning Relation Weights using Cue Validities

When using the extracted relations in applications, it is desirable to have a measure of how relevant the relations are to the associated concepts (e.g.,

Cue validity of feature *F* for concept *C*:

$$P(C|F) = \frac{P(F|C)}{\sum P(F|C_i)} \simeq \frac{f(F,C)/f(C)}{\sum f(F,C_i)/f(C_i)}$$

where C_i is a concept that contrasts with *C*

Figure 3.6: ***Calculation of cue validities.*** Formula based on (Smith and Medin, 1981). The features can be considered as attributes or relations associated with a concept (e.g., $\langle$_, size, small$\rangle$).

synset for definition being parsed). One such measure would be the degree to which the relation applies specifically to a given concept, with respect to the occurrence of the relation with contrasting concepts. To account for this, *cue validities* are used. As discussed in Section 1.2.1.2, these can be interpreted as probabilities indicating the degree to which features apply to a given concept versus similar concepts (e.g., sibling concepts).

Cue validities (CVs) are estimated by calculating the percentage of times that features are associated with a concept versus the total associations for contrasting concepts, as shown in Figure 3.6. As an example, 39 of the 180 WordNet definitions for types of dogs mention the size small. Thus smallness is not discriminating with respect to dogs, and it should be assigned a low weight:

$P(\text{small-dog} \mid \text{small}) \simeq 1/39$

In contrast, only eight of the definitions mention the color red. Redness is much more discriminating and should be given a relatively high cue validity weight:

$P(\text{red-dog} \mid \text{red}) \simeq 1/8$

This illustrates that cue validities are inversely proportional to feature frequency.

There are several issues in using cue validities for weighting semantic relations. The main issue concerns what are considered as features: is it just the relational target term, the relation type, or both of these? Note that the entire relationship is not considered as this would most likely be unique, leading to all CV's near 1. Just using the target term does not account for the type of relationship and is thus undesirable; and, if just the relation type is used, informative relations such as *is-a* might be penalized due to high frequency. Therefore, the relation type and target term are used together as the feature, as in $\langle$_, *is-a*, *mammal*$\rangle$. This works well for relations that have comparable frequencies. For example, the range for *is-a* covers a large percentage of the entire set of con-

cepts. Therefore, relationships using it are generally weighted high unless the target term is commonly used (e.g., *animal*). In contrast, low weights are usually assigned to relations having a restricted domain (e.g. *has-attribute* maps just into characteristics). Exceptions would deal with target terms rarely used in such relations (e.g., an uncommon attribute). One aspect this approach does not account for is the informativeness of relations. For instance, a large range of target terms can occur with a generic semantic-relatedness relation (e.g., *related-to*). Such relations might be weighted as high as *is-a*, even though it is much less informative. This is left for future work.

The calculation of cue validities requires a means of determining the set of contrasting concepts for a given concept. The simplest way to do this would be just to select the set of sibling concepts (e.g., synsets sharing a common parent in WordNet). However, due to the idiosyncratic way concepts are specialized in knowledge bases, this likely would not include concepts intuitively considered as contrasting. For example, in WordNet *geometry teacher* and *piano teacher* have different immediate parents, *math teacher* and *music teacher*, respectively, although they do have the common grandparent *teacher*. The related concept $lecturer_{\text{EDUCATOR}}$ specializes the more general *educator* concept, which is the parent of *teacher*. Thus, a lecturer would not be considered as contrasting with geometry and piano teachers in a simple scheme based on sibling or cousin terms.

To alleviate this problem the *most-informative ancestor* is used instead of the parent, and all of its descendants used for the set of contrasting concepts. The most-informative ancestor is determined by selecting the ancestor that best balances frequency of occurrence with specificity. This is similar to Resnik's (1995) notion of most-informative subsumer for a pair of concepts. In his approach, estimated frequencies for synsets are percolated up the hierarchy, so that the frequency increases as one progresses up the hierarchy. Therefore, the first common ancestor for a pair is the most-informative subsumer (i.e., has the most information content). Here attested frequencies from SemCor (Miller et al., 1994) are used, so all ancestors are considered. Specificity is accounted for by applying a scaling factor to the frequencies that decreases as one proceeds up the hierarchy. Thus, 'informative' is used more in an intuitive sense rather than a technical one.

The cue validities for all of the relations are calculated at the same time in a two-step process. Whenever the knowledge base changes, the cue validities might need to be revised, since they are a global measure. (Such updates will typically be restricted to portions of the knowledge base (e.g., mammals), minimizing overhead.) First, for each concept associated with relations (e.g., the synset for the word being defined), its most-informative ancestor (MIA) is determined. Associations are then updated for each of the features present in

the relations (i.e, $f(F, C)$ in Figure 3.6), with features determined by the combination of relation type and target term (i.e., $\langle$_, relation, target$\rangle$). In the second step, the cue validities (i.e., $P(C|F)$) are determined by the ratio of this frequency to the sum of the frequencies for all concepts that are descendants of the MIA (i.e., $\sum f(F, C_i)$).

Cue validities can be viewed as a type of summary statistic, characterizing the state of the knowledge base without requiring external training data (as with lexical associations). In the actual system, they are applied after the disambiguation step discussed in the next chapter. That way, the cue validities are calculated using conceptual features (e.g., $\langle$_, *category*, $grape_{FRUIT}\rangle$) rather than word features (e.g., $\langle$_, of, grape$\rangle$).

3.3.3 Converting into Nested Relation Format

The unstructured relation listing resulting from the parse postprocessing can become unreadable when the parses are complex. To alleviate this problem, the relations are converted into a format that incorporates nesting to account for subordinate relationships. The nesting also dispenses with the need for word offsets. This format makes it easier for humans to evaluate the relation extraction and facilitates revisions in case humans will be doing post-editing of the extracted relations prior to incorporation in a lexicon. Figure 3.7 shows an example of this conversion.

3.4 Differentia Extraction Algorithm

To summarize the extraction process, Figure 3.8 presents a high-level description of the differentia extraction algorithm. This includes the disambiguation step to be discussed in the next chapter, as that needs to be run prior to the relation weighting. The code has been implemented using Perl. It is available for download at www.cs.nmsu.edu/~tomohara/differentia-extraction.

.

Definition sentence:

An artifact is a man-made object taken as a whole.

Extracted relations in tuple format:

⟨2. n:artifact, 3. v:is, 6. n:object⟩
⟨5. man-made, *modifies*, 6. n:object⟩
⟨6. n:object, *modified-by*, 7. v:taken⟩
⟨7. v:taken, 8. p:as, 10. n:whole⟩

Extracted relations in nested format:

noun:artifact
    verb:is noun:object
        modifies man-made
        modified-by verb:taken
            as noun:whole

Figure 3.7: ***Sample conversion from relation tuples into nested format.*** Source terms from the relationships are omitted and instead indicated implicitly by indentation.

Input Definitions for lexicon entries

Output List of word-level relationships:
⟨source-word, relation-function-word, target-word⟩

Example Extracting relations from definition of 'kennel':
"an outbuilding that serves as a shelter for a dog"
⇒
⟨4. noun:outbuilding, 5. pronoun:that, 6. verb:serves⟩
⟨6. verb:serves, 10. prep:for, 12. noun:dog⟩
⟨6. verb:serves, 7. prep:as, 9. noun:shelter⟩

Steps For each definition:

1. Preprocess definition text, isolating punctuation, removing domain indicators and example sentences.
2. Convert definition fragments into complete sentences.
3. Parse definitions using Link Grammar parser producing syntactic relations.
4. Convert into format with relation types based either on high-level grammatical relations or on function words.
5. Disambiguate the parses (see Chapter 4).
6. Weight the relations based on cue validities ($P(C|F)$ with features interpreted as ⟨_, relation, target⟩).
7. Convert from flat relational tuples into nested relation format.

Figure 3.8: ***Differentia extraction algorithm.*** Steps 1 through 4 form the basic differentia extraction process as discussed in Section 3.2; the example just illustrates these steps. Step 5 is discussed in the next chapter. Steps 6 and 7 are discussed in Section 3.3.

CHAPTER 4
DIFFERENTIA DISAMBIGUATION

After the differentiating properties have been extracted from a definition (as discussed in the last chapter), the words for the relation source and object terms should be disambiguated in order to reduce vagueness in the relationships. In addition, the relation types should be disambiguated from surface-level relations or relation-indicating words (e.g., prepositions) into the underlying semantic relationship. For example, the relationship ⟨dog, with, coat⟩ would get transformed into $\langle dog_{\text{CANINE}}, \textit{has-part}, coat_{\text{HAIR}} \rangle$.

Both aspects of this disambiguation are discussed in this chapter with emphasis on relation disambiguation, since word-sense disambiguation (WSD) has received more attention in computational linguistics (Kilgarriff and Palmer, 2000b; Edmonds and Kilgarriff, 2002). The approach to disambiguation uses statistical classifiers learned from training data annotated with semantic relations, so information about these inventories is included as well.

Note that the two disambiguation processes, namely WSD of the source and target words (e.g., 'dog' and 'coat') and WSD of the relation words (e.g., 'with'), are not necessarily sequential. They can be applied in either order or at the same time. As presented here, they are independent, but it might be helpful to interleave the processes. This way the results of disambiguation for some of the source and target terms can influence the disambiguation of the relation types (and vice versa).

This chapter is organized as follows. Section 4.1 briefly discusses term disambiguation and the word-sense annotations used for this purpose. Section 4.2 presents background information on the relation inventories used during classification. Section 4.3 discusses the relation classifiers in depth with results given for three different inventories. Section 4.4 closes with a summary of the disambiguation algorithm, including an example.

4.1 Source and Target Term Disambiguation

The first step in differentia disambiguation is to resolve the relational source and target terms into the underlying concepts (e.g., 'coat' into $coat_{\text{HAIR}}$). Since WordNet serves as the knowledge base being targeted, this involves selecting the most appropriate synset for both the source and target terms. Synsets and word senses are closely related, so word-sense disambiguation serves to resolve the underlying concept at the same time. If another dictionary were being used as the source of the sense inventory for the WSD, there would be an additional step of mapping the word senses into the target knowledge

base (e.g., sense 2b of ‘dog’ in Merriam-Webster’s dictionary into dog_{CHAP}). Some applications might not require disambiguated terms, so this step is optional. An example would be text segmentation where relations among words are used to provide clues for segment cohesiveness (as sketched out later in Section 6.2.3).

Several different approaches to word-sense disambiguation were presented in Section 2.4.3. These fall into two main categories: supervised approaches use examples to induce classifies; and, unsupervised approaches instead use heuristics when deciding on the best word sense (e.g., based on word overlap with dictionary definitions). There have also been a few hybrid approaches.

The Extended WordNet (XWN) project is endeavoring to convert WordNet into a more comprehensive knowledge base by providing logical form representations of the definitions (Harabagiu et al., 1999; Rus, 2002). As part of this, the content words from the definitions are being sense annotated with respect to the WordNet inventory (Novischi, 2002). The disambiguation approach used here is based only on these sense annotations (i.e., table lookup). For other dictionaries, use of traditional word-sense disambiguation algorithms would be required.

Figure 4.1 gives an example of the XWN annotations for the definition of ‘beagle.’ The **wnsn** attribute gives the sense number. For example, ‘breed’ refers to *breed#2*, the animal-group sense, rather than the lineage or caste senses. With respect to word-sense disambiguation, the annotations are treated as follows (with part-of-speech and sense number indicated by subscripts):

$$\text{a small}_{adj1}\ \text{short}_{adj1}\text{-legged}_{adj1}\ \text{smooth}_{adj1}\text{-coated}_{adj1}\ \text{breed}_{n2}\ \text{of hound}_{n1}$$

For Extended WordNet, they have used a semi-automated process to annotate senses. They use two separate systems to sense-tag the definitions: one is tailored to WordNet (see Figure 2.5) and the other is a general word-sense tagger. If the two systems disagree, then the tagging from the system tailored to WordNet is used with a default confidence indicator (i.e., quality=“normal”). If these two systems agree then the selected sense is used and assigned a higher confidence indicator (i.e., quality=“silver”). They have also manually checked some of the annotations (roughly 5% of the data). These are assigned the highest confidence indicator (i.e., quality=“gold”).

Note that the WordNet team is working on an alternative source of sense annotations for the glosses (Langone et al., 2004). These are being manually produced and thus will be more reliable in general than the ones produced for

⟨gloss pos="NOUN" synsetID="02005361"⟩
⟨synonymSet⟩beagle⟨/synonymSet⟩
⟨text⟩
a small short-legged smooth-coated breed of hound
⟨/text⟩

⟨wsd⟩
⟨wf pos="DT"⟩a⟨/wf⟩
⟨wf pos="JJ" lemma="small" quality="normal" **wnsn="1"**⟩small⟨/wf⟩
⟨wf pos="JJ" lemma="short" quality="normal" **wnsn="1"**⟩short⟨/wf⟩
⟨punct⟩-⟨/punct⟩
⟨wf pos="JJ" lemma="legged" quality="silver" **wnsn="1"**⟩legged⟨/wf⟩
⟨wf pos="JJ" lemma="smooth" quality="normal" **wnsn="1"**⟩smooth⟨/wf⟩
⟨punct⟩-⟨/punct⟩
⟨wf pos="JJ" lemma="coated" quality="normal" **wnsn="1"**⟩coated⟨/wf⟩
⟨wf pos="NN" lemma="breed" quality="normal" **wnsn="2"**⟩breed⟨/wf⟩
⟨wf pos="IN" ⟩of⟨/wf⟩
⟨wf pos="NN" lemma="hound" quality="silver" **wnsn="1"**⟩hound⟨/wf⟩
⟨/wsd⟩

⟨parse quality="SILVER"⟩
(TOP (S (NP (NN beagle))
(VP (VBZ is)
(NP (NP (DT a) (JJ small) (JJ short-legged) (JJ smooth-coated)
(NN breed))
(PP (IN of)
(NP (NN hound)))))
(. .)))
⟨/parse⟩

⟨lft quality="GOLD"⟩
beagle:NN(x1) ⇒ small:JJ(x1) short-legged:JJ(x1) smooth-coated:JJ(x1)
breed:NN(x1) of:IN(x1, x2) hound:NN(x2)
⟨/lft⟩
⟨/gloss⟩

Figure 4.1: ***Extended WordNet annotations for 'beagle' definition.*** Based on Extended WordNet version 2.0.1-1. The database is freely available at http://xwn.hlt.utdallas.edu.

Extended WordNet. Future work will incorporate these WSD annotations when available.

4.2 Semantic Relation Inventories

The representation of natural language utterances often incorporates the notion of semantic roles, which are analogous to the slots in a frame-based representation. In particular, there is an emphasis on the analysis of thematic roles, which serve to tie the grammatical constituents of a sentence to the underlying semantic representation. Thematic roles are also called case roles, since in some languages the grammatical constituents are indicated by case inflections (e.g., ablative in Latin).

There is a wide range of variability in the usage of semantic roles in natural language processing. Some systems use just a small number of very general roles, such as *beneficiary*. At the other extreme, some systems use quite specific roles tailored to particular domains, such as *catalyst*.

4.2.1 Background on Semantic Roles

Bruce (1975) presents an early account of case systems in natural language processing. For the most part, the systems had limited case role inventories, along the lines of the cases defined by Fillmore (1968). Palmer (1990) discussed some of the more contentious issues regarding case systems, including adequacy for representation, such as in reliance solely upon case information to determine semantics versus the use of additional inference mechanisms. Barker (1998) provides a comprehensive summary of case inventories in NLP, along with criteria for the qualitative evaluation of case systems: generality, completeness, and uniqueness. Linguistic work on thematic roles tends to stick with a limited number of roles. Frawley (1992) presents a detailed discussion of twelve thematic roles and discusses how they are realized in different languages.

During the shift in emphasis away from systems that work in small, self-contained domains to those that can handle open-ended domains during the past 15 or so years, there has been a trend towards the use of larger sets of semantic primitives (Wilks et al., 1996). These primitives can be seen as a generalization of cases to include properties as well as relations. The WordNet (Miller et al., 1990) lexicon (see Section 3.1) serves as one example of this, where a synset can be defined in terms of any of the 100,000+ synsets rather than using a set of features like [$\pm$ANIMATE]. At the same time, there has been a shift in focus from deep understanding (e.g., story comprehension)

facilitated by specially constructed knowledge bases to shallow surface-level analysis (e.g., text extraction) facilitated by corpus analysis. Thus, issues such as paraphrasability (Schank, 1973) became less critical than representational coverage (Jurafsky and Martin, 2000). Both trends seem to be behind the increase in case inventories in two relatively recent resources, namely FrameNet (Fillmore et al., 2001) and OpenCyc (OpenCyc, 2002), both of which define well over a hundred case roles. It is arguable that once deep understanding becomes back in focus, counter-trends will emerge favoring smaller inventories for tractability. However, provided that the case roles are well-structured in an inheritance hierarchy, both needs can be addressed by the same inventory.

4.2.2 Inventories Developed for Corpus Annotation

With the emphasis on corpus analysis in computational linguistics, there has been shift away from relying on explicitly coded knowledge towards the use of knowledge inferred from naturally occurring text, in particular text that has been annotated by humans to indicate phenomena of interest. For example, rather than manually developing rules for preferring one sense of a word over another based on context, the most successful approaches have automatically learned the rules based on word-sense annotations, as evidenced by the SENSEVAL competitions (Kilgarriff, 1998; Edmonds and Cotton, 2001).

The PENN TREEBANK version II (Marcus et al., 1994) provided the first large-scale set of case annotations for general-purpose text. These are very general roles as with Fillmore's (1968) roles discussed in Chapter 2. The Berkeley FRAMENET (Fillmore et al., 2001) project provides the most recent large-scale annotation of semantic roles. These are at a much finer granularity than those in Treebank, so they should prove quite useful for applications learning semantics from corpora. Relation disambiguation experiments for both of these role inventories are discussed later in this chapter.

4.2.2.1 Penn Treebank

The original TREEBANK (Marcus et al., 1993) provided syntactic annotations in the form of parse trees for text from the *Wall Street Journal*. This resource is very popular for computational linguistics, in particular for inducing part-of-speech taggers and parsers. Treebank II (Marcus et al., 1994) added 20 functional tags, including a few thematic roles such as *beneficiary*, *direction*, and *purpose*. These can be attached to any verb complement but normally occur with clauses, adverbs, and prepositions. For example, here is a simple parse tree with the newer annotation format:

Role	$Freq_{rel}$	Description
temporal	.120	indicates when, how often, or how long
locative	.092	place/setting of the event
direction	.030	starting or ending location (trajectory)
manner	.023	indicates manner, including instrument
purpose	.019	purpose or reason
extent	.012	spatial extent
benefactive	.0003	beneficiary of an action

Table 4.1: ***Frequency of Treebank II semantic role annotations.*** Relative frequencies taken from (Blaheta and Charniak, 2000) and descriptions from (Bies et al., 1995). The syntactic role annotations generally have higher frequencies; for example, the *subject* role occurs 41% of the time.

(S (NP-**TPC**-5 This)	*topic (i.e., discourse focus)*
(NP-**SBJ** every man)	*grammatical subject*
(VP contains	
(NP *T*-5)	*trace element linked to 'this'*
(PP-**LOC** within	*locative*
(NP him))))	

In addition to the usual syntactic constituents such as *NP* and *VP*, function tags are included. For example, the first NP gives the discourse topic. This also shows that the prepositional phrase (PP) is providing the location for the state described by the verb phrase. Frequency information for the semantic role annotations is shown in Table 4.1.

4.2.2.2 FrameNet

FRAMENET (Fillmore et al., 2001) is striving to develop an English lexicon with rich case structure information for the various contexts that words can occur in. Each of these contexts is called a *frame*, and the semantic relations that occur in each frame are called *frame elements*. For example, in the *communications* frame, there are frame elements for *speaker*, *message*, etc. FrameNet annotations occur at the phrase level instead of the grammatical constituent level as in Treebank. An example follows:

⟨S TPOS="56879338"⟩
⟨T TYPE="sense2"⟩⟨/T⟩
It had a sharp, pointed face and

⟨C **FE=“BodP”** PT=“NP” GF=“Ext”⟩ a feathery tail that ⟨/C⟩
⟨C TARGET=“y”⟩arched⟨/C⟩
⟨C **FE=“Path”** PT=“PP” GF=“Comp”⟩ over its back ⟨/C⟩
. ⟨/S⟩

The constituent (C) tags identify the phrases that have been annotated. The frame element (FE) attributes indicate the semantic roles, and the phrase type (PT) attributes indicate the grammatical function of the phrase.

Table 4.2 shows the top 25 semantic roles by frequency of annotation. This illustrates that the semantic roles in FrameNet can be quite specific, as with the roles *cognizer*, *judge*, and *addressee*. In all, there are over 140 roles annotated with over 117,000 tagged instances.

4.2.3 Inventories for Knowledge Representation

The next two case inventories discussed, from Cyc and Conceptual Graphs respectively, are based on the traditional knowledge representation paradigm. With respect to natural language processing, these approaches are more representative of the earlier approaches in which deep understanding is the chief goal. Nonetheless, both are evolving to meet the needs of current applications. Another case inventory is that from Factotum. It is likewise based on the knowledge representation paradigm. However, in a sense it reflects the empirical aspect of the corpus annotation approach, because the annotations were developed to address the relations implicit in Roget's Thesaurus.

Relation disambiguation experiments are only presented for Factotum, given that the others do not readily provide sufficient training data. However, both inventories are discussed because each provides relation types incorporated into the inventory used below for differentia extraction (see Section 4.3.5).

4.2.3.1 Cyc

The CYC system (Lenat, 1995) is the most ambitious knowledge representation project undertaken to date. It has been in development since 1984, originally as part of Microelectronics and Computer Technology Corporation (MCC), but later as a separate company called Cycorp (Lenat and Guha, 1990; Lenat, 1995). The full Cyc KB is proprietary, which has hindered its adaptation in natural language processing. However, to encourage broader usage, portions of the KB have been made freely available to the public. For instance, there is now an open-source version of the system called OPENCYC (www.opencyc.org), which covers the upper part of the KB and also includes the Cyc inference engine, KB browser, and other tools.

Role	$Freq_{rel}$	Description
speaker	.071	person producing the message
message	.060	content which is communicated
self-mover	.058	living being moving under its own power
theme	.054	object in relation to a particular location
agent	.050	entity that acts on another entity
goal	.047	identifies the endpoint of movement
path	.046	trajectory which is neither a source nor a goal
cognizer	.039	person who becomes aware of a phenomenon
manner	.038	property of motion unrelated to the trajectory
source	.032	starting-point of motion
content	.031	entity whose salience is described
experiencer	.030	being who has a physical experience
evaluee	.026	entity about whom/which a judgment is made
judge	.026	evaluator of protagonist's mental state
topic	.026	subject matter of the communicated message
undefined	.022	unspecified frame element
cause	.020	non-agentive cause of the physical experience
addressee	.019	person that receives a message
perceptual source	.019	source of perception (e.g., clatter [of hoofs])
phenomenon	.017	entity the perceiver experiences or senses
reason	.015	reason for the judgment
area	.011	a region in which the motion takes place
degree	.011	*no description*; ex: shook me [to my roots]
body part	.010	the body part in which a sensation is located
protagonist	.009	person to whom a mental property is attributed

Table 4.2: ***Common FrameNet semantic roles.*** The top 25 of 141 roles are shown. Descriptions based on FrameNet 0.75 frame documentation.

Cyc uses a wide range of role types: very general roles (e.g., *beneficiary*); commonly occurring situational roles (e.g., *victim*); and, highly specialized roles (e.g., *catalyst*). Of the 8756 concepts in OpenCyc, 130 are for thematic roles (i.e., instances of *ActorSlots*) with 51 other semantic roles (i.e., other instances of *Role*). Table 4.3 shows the most commonly used thematic roles in the KB. The frequency was determined by using the Cyc's indexing functions that return the assertions associated with a given term.

The Cyc role inventory is not used directly in the experiments discussed later. However, some of the roles are incorporated into the combined role inventory developed for differentia analysis. In addition, the future work chapter

Role	$Freq_{rel}$	Description
doneBy	.178	relates an event to its “doer”
performedBy	.119	doer deliberately does act
objectOfState-Change	.081	object undergoes some kind of intrinsic change of state
objectActedOn	.057	object is altered or affected in event
outputsCreated	.051	object comes into existence sometime during event
transporter	.044	object facilitating conveyance of transportees
transportees	.044	object being moved
toLocation	.041	where the moving object is found when event ends
objectRemoved	.036	object removed from its previous location
inputs	.036	pre-existing event participant destroyed or incorporated into a new entity
products	.035	object is one of the intended outputs of event
inputsDestroyed	.035	object exists before event, is destroyed during event
fromLocation	.034	loc is where some moving-object in the move is found at the beginning
primaryObject-Moving	.033	object is in motion at some point during the event and this movement is focal
seller	.030	agent sells something in the exchange
objectOf-Possession-Transfer	.030	rights to use object transferred from one agent to another
transferredThing	.030	object is being moved, transferred, or exchanged in the event transfer
senderOfInfo	.030	sender is an agent who is the source of information transferred
inputsCommitted	.028	object exists before event and continues to exist afterwards, and as a result of event, object becomes incorporated into something created during event
objectEmitted	.026	object is emitted from the emitter during the emission event

Table 4.3: ***Most common thematic roles in OpenCyc.*** Descriptions based on comments from the OpenCyc knowledge base (version 0.7).

sketches out how annotations for attributes can be inferred from Cyc. This extends the relation marker inference technique discussed below for Factotum.

4.2.3.2 Conceptual Graphs

Conceptual Graphs (CG) are the mechanism introduced by Sowa (1984) for knowledge representation as part of his Conceptual Structures theory. The original text listed two dozen or so thematic relations, such as *destination* and *initiator*. In all, 37 conceptual relations were defined. This inventory formed the basis for most work in conceptual graphs. Recently, Sowa (1999) updated the inventory to allow for better hierarchical structuring and to incorporate the important thematic roles identified by Somers (1987). Four broad categories were used, corresponding roughly to Aristotle's four causes (or *aitia*): initiator, resource, goal, and essence. In addition, six categories of verbs were used: action, process, transfer, spatial, temporal, and ambient. Table 4.4 shows a sample of these roles, along with estimated usage. These roles are generally more abstract than traditional usage. For example, *Duration* can refer to any of a variety of resources used in a temporal process, not just the time of the process.

There are currently no large-scale resources based on CG relations (i.e., neither a knowledge base nor an annotated dataset). Therefore, the role usage is estimated by issuing web searches for the various relation names and abbreviations and seeing when CG-style notation is used with the relation. In effect, this technique uses online academic papers and other CG-related web resources as a corpus. For example, for *patient* a web search would be done on the following query:

> (patient or PTNT) and ("conceptual structure" or CS or "conceptual graph" or CG)

Then the resulting text is analyzed to see how often the relation occurs in CG's linear notation, such as the following:

> [SITUATION: [CAT]←(AGNT)←[EAT]→(PTNT)→[FISH]].

The restriction to parenthesized relation names is important for reducing extraneous hits, because 'CS' and 'CG' are common search terms, dealing with computer science and C.G. Jung, respectively. However, for simplicity, the check for the arrows is omitted, as CG's linear notation makes the arrows optional in certain contexts. Table 4.4 shows the relative frequencies of the roles

Role	$Freq_{rel}$	Description
Agent	.267	entity voluntarily initiating an action
Attribute	.155	entity that is a property of some object
Characteristic	.080	types of properties of entities
Theme	.064	participant involved with but not changed
Patient	.061	participant undergoing structural change
Location	.053	participant of a spatial situation
Possession	.035	entity owned by some animate being
Part	.035	object that is a component of some object
Origin	.035	source of a spatial or ambient situation
Experiencer	.035	animate goal of an experience
Result	.032	inanimate goal of an act
Instrument	.027	resource used but not changed
Recipient	.019	animate goal of an act
Destination	.013	goal of a spatial process
PointInTime	.011	participant of a temporal situation
Path	.011	resource of a spatial or ambient situation
Accompaniment	.011	object participating with another
Effector	.008	source involuntarily initiating an action
Beneficiary	.008	entity benefiting from event completion
Matter	.005	resource that is changed by the event
Manner	.005	entity that is a property of some process
Source	.003	present at beginning of activity
Resource	.003	material necessary for situation
Product	.003	present at end of activity
Medium	.003	resource for transmitting information
Goal	.003	final cause which is purpose or benefit
Duration	.003	resource of a temporal process
Because	.003	situation causing another situation
Amount	.003	a measure of some characteristic

Table 4.4: ***Common semantic roles used in Conceptual Graphs.*** Inventory and descriptions based on (Sowa, 1999, pp. 502-510). The term *situation* is used in place of Sowa's *nexus* (i.e., "fact of togetherness"), which also covers spatial structures. $Freq_{rel}$ gives estimated relative frequency based on web searches.

using this technique for estimation. To expand this beyond the set of relations specified in (Sowa, 1999), the relation name can be omitted from the search and then all of the relation names that occur in $\rightarrow$(relation)$\rightarrow$ or $\leftarrow$(relation)$\leftarrow$ constructions could be tabulated.

4.2.3.3 Factotum

The FACTOTUM semantic network (Cassidy, 2000) developed by Micra, Inc. makes explicit many of the functional relations in Roget's Thesaurus.[1] Outside of proprietary resources such as Cyc, Factotum is the most comprehensive KB with respect to functional relations. OpenCyc does include definitions of many non-hierarchical relations. However, there are not many instantiations (i.e., relationship assertions), because it concentrates on the higher level of the ontology.

The Factotum knowledge base was based on the 1911 version of Roget's Thesaurus and specifies the relations that hold between the Roget categories and the words listed in each entry. Factotum incorporates information from other resources as well. For instance, the Unified Medical Language System (UMLS) formed the basis for the initial inventory of semantic relations, which was later revised during tagging.

Figure 4.2 shows a sample from Factotum. This illustrates that the basic Roget organization is still used, although additional hierarchical levels have been added. The relations are contained within double braces (e.g., "{{has_subtype}}") and generally apply from the category to each word in the synonym list on the same line. Therefore, the line with "{{result_of}}" indicates that conversion is the result of transforming, as shown in the semantic relation listing that would be extracted.[2] There are over 400 different relations instantiated in the knowledge base, which has over 93,000 assertions. Some of these are quite specialized (e.g., *has-brandname*). In addition, there are quite a few inverse relations, since most of the relations are not symmetrical. Certain features of the knowledge representation are ignored during the relation extraction used later. For example, relation specifications can have qualifier prefixes, such as an ampersand to indicate that the relationship only sometimes holds.

Table 4.5 shows the most common relations in terms of usage in the semantic network, and includes others that are used in the experiments discussed

[1] Factotum is based on the public domain version of Roget's Thesaurus. The latter is freely available via Project Gutenberg (http://promo.net/pg), thanks to Micra, Inc.

[2] For clarity, some of the relations are renamed to make the directionality more explicit, following a suggestion for their interpretation in the Factotum documentation.

Original data:

A6.1.4 CONVERSION (R144)
#144. Conversion.
N. {{has_subtype(change, R140)}} conversion, transformation.
{{has_case: @R7, initial state, final state}}.
{{has_patient: @R3a, object, entity}}.
{{result_of}} {{has_subtype(process, A7.7)}} converting, transforming.
{{has_subtype}} processing.
transition.

Extracted relationships:

⟨change, *has-subtype*, conversion⟩ ⟨change, *has-subtype*, transformation⟩
⟨conversion, *has-case*, initial state⟩ ⟨conversion, *has-case*, final state⟩
⟨conversion, *has-patient*, object⟩ ⟨conversion, *has-patient*, entity⟩
⟨conversion, *is-result-of*, converting⟩ ⟨conversion, *is-result-of*, transforming⟩
⟨process, *has-subtype*, converting⟩ ⟨process, *has-subtype*, transforming⟩
⟨conversion, *has-subtype*, processing⟩

Figure 4.2: ***Sample data from Factotum.*** Based on version 0.56 of Factotum.

later.[3] The relative frequencies just reflect relationships explicitly labeled in the KB data file. For instance, this does not account for implicit *has-subtype* relationships based on the hierarchical organization of the thesaural groups. The functional relations are shown in boldface. This excludes the meronym or part-whole relations (e.g., *is-conceptual-part-of*), in line with their classification by Cruse (1986) as hierarchical relations. The reason for concentrating on the functional relations is that these are more akin to the roles tagged in Treebank and FrameNet.

Table 3.1 from the previous chapter shows the relation usage in WordNet version 1.7. This shows that the majority of the relations are hierarchical (*is-similar-to* can be considered as a hierarchical relation for adjectives). As mentioned earlier, WordNet 1.7 averages only 1.3 non-taxonomic properties per concept (including inverses). OpenCyc provides a much higher average at 3.7 properties per concept, although with an emphasis on argument constraints and other usage restrictions. Factotum compares favorably in this respect, av-

[3]The database files and documentation for the semantic network are available from Micra, Inc. via ftp://micra.com/factotum.

Relation	$Freq_{rel}$	Description
has-subtype	.401	inverse of *is-a* relation
is-property-of	.077	object with given salient character
is-caused-by	.034	force that is the origin of something
has-property	.028	salient property of an object
has-part	.022	a part of a physical object
has-high-intensity	.018	intensifier for property or characteristic
has-high-level	.017	implication of activity (e.g., intelligence)
is-antonym-of	.016	generally used for lexical opposition
is-conceptual-part-of	.015	parts of other entities (in case relations)
has-metaphor	.014	non-literal reference to the word
$causes_{mental}$	.013	motivation (causation in the mental realm)
uses	.012	a tool needing active manipulation
is-performed-by	.012	human actor for the event
$performs_{human}$	.011	human role in performing some activity
is-function-of	.011	artifact passively performing the function
has-result	.010	more specific type of *causes*
has-conceptual-part	.010	generalization of *has-part*
is-used-in	.010	activity or desired effect for the entity
is-part-of	.010	distinguishes part from group membership
causes	.009	inverse of *is-caused-by*
has-method	.009	method used to achieve some goal
is-caused-by$_{mental}$	.009	inverse of $causes_{mental}$
has-consequence	.008	causation due to a natural association
has-commencement	.007	state that commences with the action
is-location-of	.007	absolute location of an object
requires	.004	object or sub-action needed for an action
is-studied-in	.004	inquires into any field of study
is-topic-of	.002	communication dealing with given subject
produces	.002	what an action yields, generates, etc.
is-measured-by	.002	instrument for measuring something
is-job-of	.001	occupation title for a job function
is-patient-of	.001	action that the object participates in
is-facilitated-by	.001	object or sub-action aiding an action
is-biofunction-of	.0003	biological function of parts of living things
was-performed-by	.0002	*is-performed-by* occurring in the past
has-consequence$_{object}$	.0002	consequence for the patient of an action
is-facilitated-by$_{mental}$	.0001	trait that facilitates some human action

Table 4.5: ***Common Factotum semantic roles.*** These account for 80% of the instances. Boldface relations are used in the experiments (Section 4.3.4.2).

eraging 1.8 properties per concept.[4] Therefore, the information in Factotum complements WordNet through the inclusion of more functional relations.

4.3 Relation Disambiguation

Relations indicated by prepositional phrases are the focus of this work. The goal of general relation disambiguation is to determine the underlying semantic role indicated by particular words in a phrase or by word order. For relations indicated directly by function words, the disambiguation can be seen as a special case of word-sense disambiguation. As an example, disambiguating the relationship ⟨'dog', 'with', 'ears'⟩ into ⟨'dog', *has-part*, 'ears'⟩, is equivalent to disambiguating the preposition 'with,' given senses for the different relations it can indicate. For relations that are indicated implicitly (e.g., adjectival modification), other classification techniques would be required, reflecting the more syntactic nature of the task. For example, adjective modification could be approximated by positing an underlying preposition (e.g., 'modifier-of') that occurs as a trace element in the sentence. Providing a general framework for the disambiguation of implicitly indicated relations is an area for future work.

Traditionally, prepositions have numerous senses. For instance, the preposition 'for' has 20 different senses defined in Merriam-Webster's dictionary (10th Edition), as shown in Table 4.6. (WordNet does not include function words, so sense inventories from other dictionaries are discussed here.) The Treebank roles are more general than these: for the preposition 'for,' there are six distinctions (four with low-frequency pruning). The Treebank role disambiguation experiments thus address a coarse form of sense distinction. In contrast, the FrameNet distinctions are quite specific: there are 41 distinctions associated with 'for' (18 with low-frequency pruning). The FrameNet role disambiguation experiments thus address fine-grained sense distinctions.

4.3.1 Overview of Relation Type Disambiguation

Recall that the output of the extraction step is a list of relationship tuples in the following format:

⟨source-word, relation-word, target-word⟩

These need to be disambiguated into the underlying concepts:

[4] These figures are derived by counting the number of relations excluding the instance and subset ones. Cyc's comments and lexical assertions are also excluded, as these are implicit in Factotum and WordNet. The count is then divided by the number of concepts.

1. in place of; instead of [to use blankets for coats]
2. as the representative of; in the interest of [to act for another]
3. in defense of; in favor of [to fight for a cause]
4. in honor of [to give a banquet for someone]
5. with the aim or purpose of [to carry a gun for protection]
6. with the purpose of going to [to leave for home]
7. in order to be, become, get, have, keep, etc. [to walk for exercise]
8. in search of [to look for a lost article]
9. meant to be received by a specified person or thing, or to be used in a specified way [flowers for a girl, money for paying bills]
10. suitable to; appropriate to [a room for sleeping]
11. with regard to; as regards; concerning [a need for improvement]
12. as being [to know for a fact]
13. considering the nature of; as concerns [cool for July]
14. because of; as a result of [to cry for pain]
15. in proportion to; corresponding to [two dollars spent for every dollar earned]
16. to the amount of; equal to [a bill for $50]
17. at the price or payment of [sold for $20,000]
18. to the length, duration, or extent of; throughout; through [to walk for an hour]
19. at (a specified time) [a date for two o'clock]
20. [Obs.] before

Table 4.6: ***Definitions of preposition 'for' in Merriam-Webster's dictionary.*** Taken from (Mish, 1996), also available at www.m-w.com.

⟨source-concept, *relation-type*, target-concept⟩

For the "dog with coat" example, the relation word 'with' would be disambiguated into the relation type *has-part*

Unlike the situation with the relational source and target concepts (e.g., synsets), there is just a limited number of relation types (e.g., case roles). As an illustration, Cyc has the largest number of thematic roles compared to other resources, but the total number is still just a few hundred (Lehmann, 1996). This is quite small compared to the 100,000+ synsets in WordNet, each of which could serve as a source or target concept. Therefore, a supervised learning approach is much more feasible. In this approach, predefined classifications for several examples of each of the possible relation types (or roles) are input, along with feature descriptions of the examples, into a machine learning system that induces classification rules. The resulting rules can either be symbolic (e.g., decision trees) or statistical (e.g., conditional probability tables). To

simplify the following discussion, the term *roles* is used whenever the relation types are restricted to thematic roles (Fillmore, 1968) rather than relation types in general. This convention alleviates a source of ambiguity between 'relation' as relation type versus relation instantiation (often referred to as 'relationship' here).

4.3.1.1 Class-based Collocations via Hypernyms

A straightforward approach for preposition disambiguation would be to use standard WSD features, such as the parts-of-speech of surrounding words and, more importantly, collocations (e.g., lexical associations). Although this can be highly accurate, it tends to overfit the data and to generalize poorly. The latter is of particular concern here as the training data is taken from a different genre. For example, the Treebank data is from general-purpose newspaper text (i.e., *Wall Street Journal*), but the differentiating relations are being extracted from dictionary definitions. To overcome these problems, a class-based approach is used for the collocations, with WordNet high-level synsets as the source of the word classes. Therefore, in addition to using collocations in the form of other words, this uses collocations in the form of semantic categories.

Word collocation features are derived by making two passes over the training data. The first pass tabulates the co-occurrence counts for each of the context words (i.e., those in a window around the target word) paired with the classification value for the given training instance (e.g., the preposition sense from the annotation). These counts are used to derive conditional probability estimates of each class value given co-occurrence of the various potential collocates. The words exceeding a certain threshold are collected into a list associated with the class value, making this a "bag of words" approach. In the experiments discussed below, a potential collocate is selected whenever the conditional probability for the class value exceeds the prior probability by a factor greater than 20%:

$$\frac{P(C|coll) - P(C)}{P(C)} \geq .20$$

That is, the relative difference between the class conditional probability given the potential collocation occurrence ($P(C|coll)$) and the prior probability for the class value ($P(C)$) must be 20% or higher for the potential collocation word ($coll$) to be treated as one of the actual collocation words. The second pass over the training data determines the value for the collocational feature of each classification category by checking whether the current context window has any

of the associated collocation words. Note that for the test data, only the second pass is made, using the collocation lists derived from the training data.

In generalizing this to a class-based approach, the potential collocational words are replaced with each of their hypernym ancestors from WordNet. The adjective hierarchy is relatively shallow, so it is augmented by treating *is-similar-to* as *has-hypernym*. Adverbs would be included, but there is no hierarchy for them. Since the co-occurring words are not sense-tagged, this is done for each synset serving as a different sense of the word. Likewise, in the case of multiple inheritance, each parent synset is used. For example, given the co-occurring word 'money,' the counts would be updated as if each of the following tokens were seen, shown grouped by sense.

1. { medium_of_exchange#1, monetary_system#1, standard#1, criterion#1, measure#2, touchstone#1, reference_point#1, point_of_reference#1, reference#3, indicator#2, signal#1, signaling#1, sign#3, communication#2, social_relation#1, relation#1, abstraction#6 }

2. { wealth#4, property#2, belongings#1, holding#2, material_possession#1, possession#2 }

3. { currency#1, medium_of_exchange#1, monetary_system#1, standard#1, criterion#1, measure#2, touchstone#1, reference_point#1, point_of_reference#1, reference#3, indicator#2, signal#1, signaling#1, sign#3, communication#2, social_relation#1, relation#1, abstraction#6 }

Thus, the word token 'money' is replaced by 41 synset tokens. Then, the same two-pass process described above is performed over the text consisting of the replacement tokens. Although this introduces noise due to ambiguity, the conditional-probability selection scheme (Wiebe et al., 1998a) compensates by selecting hypernym synsets that tend to co-occur with specific categories.

4.3.1.2 Classification Experiments

A supervised approach for word-sense disambiguation is used following Bruce and Wiebe (1999). The results described here were obtained using the settings in Figure 4.3. These are similar to the settings used by O'Hara et al. (2000) in the first SENSEVAL competition, with the exception of the hypernym collocations. This shows that, for the hypernym associations, only those words

Features:	
Prep:	preposition being classified
POS$-$i:	part-of-speech of *i*th word to left
POS$+$i:	part-of-speech of *i*th word to right
WordColl$_r$:	context has word collocation for role *r*
HypernymColl$_r$:	context has hypernym collocation for role *r*
Collocation context:	
Word:	anywhere in the sentence
Hypernym:	within 5 words of target preposition
Collocation selection:	
Frequency:	$f(\text{word}) > 1$
Conditional probability:	$P(C\|\text{coll}) \geq .50$
Relative conditional probability:	$(P(C\|\text{coll}) - P(C))/P(C) \geq .20$
Organization:	per-class-binary
Model selection:	
Overall classifier:	Decision tree
Individual classifiers:	Naive Bayes

Figure 4.3: ***Feature settings used in preposition classification experiments.*** The *per-class-binary* organization uses a separate binary feature per role (Wiebe et al., 1998a).

that occur within five words of the target prepositions are considered (i.e., a five word context window).[5]

As mentioned above, the main difference from that of a standard WSD approach is the use of WordNet hypernyms as class-based collocations. The feature settings in Figure 4.3 are used in three different configurations: word-based collocations alone, hypernym-collocations alone, and a combination of the two types of collocations. This combination generally produces the best results. This balances the specific clues provided by the word collocations with the generalized clues provided by the hypernym collocations.

4.3.2 Penn Treebank

When deriving training data from Treebank via the parse tree annotations, the functional tags associated with prepositional phrases are converted

[5]This window size was chosen after estimating that on average the prepositional objects occur within 2.35 ± 1.26 words of the preposition and that the average attachment site is within 3.0 ± 2.98 words. These figures were produced by analyzing the parse trees for the semantic role annotations in the Penn Treebank.

Tag	Role	$Freq_{rel}$
LOC	locative	.472
TMP	temporal	.290
DIR	direction	.149
MNR	manner	.050
PRP	purpose	.030
EXT	extent	.008
BNF	benefactive	.001

Table 4.7: ***Treebank semantic roles for PP's.*** *Tag* is the label for the role used in the annotations, whereas *Role* is the full name. $Freq_{rel}$ is the relative frequency of the role occurrence (36,476 total instances).

into preposition sense tags. Consider the sample annotation for Treebank shown earlier:

(S (NP-TPC-5 This)	*topic (i.e., discourse focus)*
(NP-SBJ every man)	*grammatical subject*
(VP contains	
(NP *T*-5)	*trace element linked to 'this'*
(*PP*-**LOC** within	*locative*
(NP him))))	

Treating *locative* as the preposition sense would yield the following annotation:

This every man contains $within_{LOC}$ him.

The relative frequencies of the roles in the Treebank prepositional phrase annotations are shown in Table 4.7.

The frequencies for the most frequent prepositions that have occurred in the prepositional phrase annotations are shown later in Table 4.11. The table is ordered by entropy, which measures the inherent ambiguity in the classes as given by the annotations. Note that the *Baseline* column is the probability of the most frequent sense, which is a common estimate of the lower bound for classification experiments.

4.3.2.1 Illustration with 'at'

As an illustration of the probabilities associated with class-based collocations, consider the differences in the prior versus class-based conditional

Relation	P(R)	Example
locative	.732	workers *at* a factory
temporal	.239	expired *at* midnight Tuesday
manner	.020	has grown *at* a sluggish pace
direction	.006	CDs aimed *at* individual investors

Table 4.8: ***Prior probabilities of roles for 'at' in Treebank.*** *P(R)* is the relative frequency. *Example* usages are taken from the corpus.

Category	Relation	P(R|C)
ENTITY#1	locative	0.86
ENTITY#1	temporal	0.12
ENTITY#1	other	0.02
ABSTRACTION#6	locative	0.51
ABSTRACTION#6	temporal	0.46
ABSTRACTION#6	other	0.03

Table 4.9: ***Sample conditional probabilities of roles for 'at' in Treebank.*** *Category* is WordNet synset defining the category. $\mathrm{P}(\mathrm{R}|\mathrm{C})$ is probability of the relation given that the synset category occurs in the context.

probabilities for the semantic roles of the preposition 'at' in the Penn Treebank (version II). Table 4.8 shows the global probabilities for the roles assigned to 'at.' Table 4.9 shows the conditional probabilities for these roles given that certain high-level WordNet categories occur in the context. In a context with a concrete concept (*entity#1*), the difference in the probability distributions,

$$\mathrm{P}(\mathrm{R} = \mathrm{locative}|\mathrm{C} = \mathrm{entity}\#1) - \mathrm{P}(\mathrm{R} = \mathrm{locative}) = 0.13$$
$$\mathrm{P}(\mathrm{R} = \mathrm{temporal}|\mathrm{C} = \mathrm{entity}\#1) - \mathrm{P}(\mathrm{R} = \mathrm{temporal}) = -0.12$$
$$\mathrm{P}(\mathrm{R} = \mathrm{other}|\mathrm{C} = \mathrm{entity}\#1) - \mathrm{P}(\mathrm{R} = \mathrm{other}) = -0.22,$$

shows that the *locative* interpretation becomes even more likely. In contrast, in a context with an abstract concept (*abstraction#6*), the difference in the probability distributions,

$$\mathrm{P}(\mathrm{R} = \mathrm{locative}|\mathrm{C} = \mathrm{abstraction}\#6) - \mathrm{P}(\mathrm{R} = \mathrm{locative}) = -0.22$$
$$\mathrm{P}(\mathrm{R} = \mathrm{temporal}|\mathrm{C} = \mathrm{abstraction}\#6) - \mathrm{P}(\mathrm{R} = \mathrm{temporal}) = 0.22$$
$$\mathrm{P}(\mathrm{R} = \mathrm{other}|\mathrm{C} = \mathrm{abstraction}\#6) - \mathrm{P}(\mathrm{R} = \mathrm{other}) = 0.001,$$

Experiment	Accuracy	STDEV
Word Only	81.1	.996
Hypernym	85.9	.702
Combined	*86.1*	.491

# Instances:	26616
# Classes:	7
Entropy:	1.917
Baseline:	48.0

Table 4.10: ***Overall preposition disambiguation results over Treebank roles.*** A single classifier is used for all the prepositions. # *Instances* is the number of role annotations. # *Classes* is the number of distinct roles. *Entropy* measures non-uniformity of the role distributions. *Baseline* is estimated by the most-frequent role. The *Word Only* experiment uses just word collocations, whereas *Combined* uses both word and hypernym collocations. *Accuracy* is average for percent correct over ten trials in cross validation. STDEV is the standard deviation over the trials. The difference in the combination versus word-only experiments is statistically significant at $p < .01$ via a paired t-test.

shows that the *temporal* interpretation becomes more likely. Therefore, these class-based lexical associations reflect the intuitive use of the prepositions.

4.3.2.2 Results

The classification results for these prepositions in the Penn Treebank show that this approach is very effective. Table 4.10 shows the results when all of the prepositions are classified together. Unlike the general case for WSD, the sense inventory is the same for all the words here; therefore, a single classifier can be produced rather than individual classifiers. This has the advantage of allowing more training data to be used in the derivation of the clues indicative of each semantic role. Good accuracy is achieved when just using standard word collocations. Table 4.10 also shows that significant improvements are achieved using a combination of word and hypernym collocations. For the single-classifier case, the accuracy is 86.1%, using Weka's J4.8 classifier (Witten and Frank, 1999), which is an implementation of Quinlan's (1993) C4.5 decision tree learner. For comparison, Table 4.11 shows the results for individual classifiers created for each preposition (using Naive Bayes). In this case, the word-only collocations perform slightly better than the combined collocations: 78.5% versus 77.8% accuracy.

Preposition	Freq	Entropy	Baseline	Word Only	Combined
through	332	1.668	0.438	0.598	0.634
as	224	1.647	0.399	0.820	0.879
by	1043	1.551	0.501	0.867	0.860
between	83	1.506	0.483	0.733	0.751
of	30	1.325	0.567	0.800	0.814
out	76	1.247	0.711	0.788	0.764
for	1406	1.223	0.655	0.805	0.796
on	1927	1.184	0.699	0.856	0.855
throughout	61	0.998	0.525	0.603	0.584
across	78	0.706	0.808	0.858	0.748
from	1521	0.517	0.917	0.912	0.882
Total	6781	1.233	0.609	0.785	0.778

Table 4.11: ***Per-preposition disambiguation results over Treebank roles.*** A separate classifier is used for each preposition. *Freq* gives the frequency for the prepositions. The *Word Only* and *Combined* columns show averages for percent correct over ten trials. *Total* averages the values of the individual experiments (except for *Freq*). See Table 4.10 for information on the other columns.

4.3.3 FrameNet

A similar preposition word-sense disambiguation experiment is carried out over the FrameNet semantic role annotations involving prepositional phrases. Consider the sample annotation shown earlier:

⟨S TPOS="56879338"⟩
⟨T TYPE="sense2"⟩⟨/T⟩
It had a sharp, pointed face and
⟨C FE="BodP" PT="NP" GF="Ext"⟩ a feathery tail that ⟨/C⟩
⟨C TARGET="y"⟩arched⟨/C⟩
⟨C FE="**Path**" PT="*PP*" GF="Comp"⟩ over its back ⟨/C⟩
. ⟨/S⟩

The prepositional phrase annotation is isolated and treated as the sense of the preposition. This yields the following sense annotation:

It had a sharp, pointed face and a feathery tail that arched over$_{\text{Path}}$ its back.

The annotation frequencies for the most frequent prepositions are shown later in Table 4.16, again ordered by entropy. This illustrates that the role distributions are more complicated, yielding higher entropy values on average. In all, there are over 100 prepositions with annotations, 65 with ten or more instances each. (Several of the low-frequency cases are actually adverbs (e.g. 'anywhere'), but are treated as prepositions during the annotation extraction.)

4.3.3.1 Illustration with 'at'

Relation	P(R)	Example
addressee	.315	growled *at* the attendant
other	.092	chuckled heartily *at* this admission
phenomenon	.086	gazed *at* him with disgust
goal	.079	stationed a policeman *at* the gate
content	.051	angry *at* her stubbornness

Table 4.12: ***Prior probabilities of roles for 'at' in FrameNet.*** Only the top 5 of 40 applicable roles are shown; otherwise similar to Table 4.8.

Category	Relation	P(R|C)
ENTITY#1	addressee	0.28
ENTITY#1	goal	0.11
ENTITY#1	phenomenon	0.10
ENTITY#1	other	0.09
ENTITY#1	content	0.03
ABSTRACTION#6	addressee	0.22
ABSTRACTION#6	other	0.14
ABSTRACTION#6	goal	0.12
ABSTRACTION#6	phenomenon	0.08
ABSTRACTION#6	content	0.05

Table 4.13: ***Sample conditional probabilities of roles for 'at' in FrameNet.*** See Table 4.9 for the legend.

It is illustrative to compare the prior probabilities of the roles (i.e., P(R)) for FrameNet to those seen earlier for 'at' in Treebank. See Table 4.12 for the most frequent roles out of the 40 cases that were assigned to it. This highlights a difference between the two sets of annotations. The common *temporal* role

Experiment	Accuracy	STDEV
Word Only	48.9	0.94
Hypernym	48.0	1.32
Combined	49.4	0.59

# Instances:	27295
# Classes:	129
Entropy:	5.128
Baseline:	14.9

Table 4.14: ***Overall results for preposition disambiguation with FrameNet.*** All roles are considered. See Table 4.10 for the legend.

from Treebank is not directly represented in FrameNet, and it is not subsumed by another specific role. There is a *location* role in FrameNet, but it applied in less than 0.3% of all the role annotations. This reflects the bias of FrameNet towards roles that are an integral part of the frame under consideration: location and time apply to all frames, so these cases are generally not annotated.

4.3.3.2 Results

Table 4.14 shows the results of classification when all of the prepositions are classified together. Due to the very large number of roles, the overall results are not that high. However, the combined collocation approach still shows slight improvement (49.4% versus 49.0%). The FrameNet inventory contains many low-frequency relations that complicate this type of classification. By filtering out relations that occur in less than 1% of the role occurrences for prepositional phrases, significant improvement results, as shown in Table 4.15. Even with filtering, the classification is challenging (e.g., 25 classes with entropy 4.055).

Table 4.16 shows the results when using individual classifiers. This shows that the combined collocations produce better results: 70.3% versus 68.5% for word collocations alone. Unlike the case with Treebank, the single-classifier performance is below that of the individual classifiers. This is due to the fine-grained nature of the role inventory. When all the roles are considered together, prepositions are sometimes being incorrectly classified using roles that have not been assigned to them in the training data. This occurs when contextual clues are stronger for a commonly used role than for the appropriate one. Given Treebank's small role inventory, this problem does not occur in the corresponding experiments.

Experiment	Accuracy	STDEV
Word Only	59.5	1.20
Hypernym	58.4	1.32
Combined	60.5	1.14

# Instances:	22125
# Classes:	25
Entropy:	4.055
Baseline:	18.4

Table 4.15: ***Preposition disambiguation omitting rare FrameNet roles.*** Excludes roles with less than 1% relative frequency. Table 4.10 gives the legend.

Prep	Freq	Entropy	Baseline	Word Only	Combined
between	286	3.258	0.490	0.325	0.537
against	210	2.998	0.481	0.310	0.586
under	125	2.977	0.385	0.448	0.440
as	593	2.827	0.521	0.388	0.598
over	620	2.802	0.505	0.408	0.526
behind	144	2.400	0.520	0.340	0.473
back	540	1.814	0.544	0.465	0.567
around	489	1.813	0.596	0.607	0.560
round	273	1.770	0.464	0.513	0.533
into	844	1.747	0.722	0.759	0.754
about	1359	1.720	0.682	0.706	0.778
through	673	1.571	0.755	0.780	0.779
up	488	1.462	0.736	0.736	0.713
towards	308	1.324	0.758	0.786	0.740
away	346	1.231	0.786	0.803	0.824
like	219	1.136	0.777	0.694	0.803
down	592	1.131	0.764	0.764	0.746
across	544	1.128	0.824	0.820	0.827
off	435	0.763	0.892	0.904	0.899
along	469	0.538	0.912	0.932	0.915
onto	107	0.393	0.926	0.944	0.939
past	166	0.357	0.925	0.940	0.938
Total	10432	1.684	0.657	0.685	0.703

Table 4.16: ***Per-preposition disambiguation results over FrameNet roles.*** See Table 4.11 for the legend.

4.3.4 Factotum

Note that Factotum does not indicate the way the relationships are expressed in English. Similarly, WordNet does not indicate this, but it does include definition glosses. For example,

Factotum:
⟨drying, *is-function-of*, drier⟩

WordNet:

dry_{ALTER}	remove the moisture from and make dry
$dryer_{\text{APPLIANCE}}$	an appliance that removes moisture

These definition glosses might be useful in certain cases for inferring the *relation markers* (i.e., generalized case markers). As is, Factotum cannot be used to provide training data for learning how the relations are expressed in English. This contrasts with corpus-based annotations, such as Treebank II (Marcus et al., 1994) and FrameNet (Fillmore et al., 2001), where the relationships are marked in context.

4.3.4.1 Inferring Semantic Role Markers

To overcome the lack of context in Factotum, the relation markers are inferred through corpus checks, in particular through proximity searches involving the source and target terms. For example, using AltaVista's Boolean search,[6] this can be done via "source NEAR target." Unfortunately, this technique would require detailed post-processing of the web search results, possibly including parsing, in order to extract the patterns. As an expedient, common prepositions[7] are included in a series of proximity searches to find the preposition occurring the most with the terms. For instance, given the relationship ⟨drying, *is-function-of*, drier⟩, the following searches would be performed.

drying NEAR drier NEAR in
drying NEAR drier NEAR to
...
drying NEAR drier NEAR "around"

To account for prepositions that occur frequently (e.g., 'of'), mutual information (MI) statistics (Manning and Schütze, 1999) are used in place of the raw frequency when rating the potential markers. These are calculated as follows:

[6] AltaVista's Boolean search is available at www.altavista.com/sites/search/adv.

[7] The common prepositions are determined from the prepositional phrases assigned functional annotations in Penn Treebank II (Marcus et al., 1994).

$$MI_{prep} = \log_2 \frac{P(X,Y)}{P(X) \times P(Y)} \approx \log_2 \frac{f(\text{source NEAR target NEAR prep})}{f(\text{source NEAR target}) \times f(\text{prep})}$$

Such checks are done for the 25 most common prepositions to find the preposition yielding the highest mutual information score. Using this metric, the top three markers for the ⟨drying, *is-function-of*, drier⟩ relationship are 'during,' 'after,' and 'with.'

4.3.4.2 Method for Classifying Functional Relations

Given the functional relationships in Factotum along with the inferred relation markers, machine learning algorithms can be used to infer what relation most likely applies to terms occurring together with a particular marker. Note that the main purpose of including the relation markers is to provide clues for the particular type of relation. Because the source term and target terms might occur in other relationships, associations based on them alone might not be as accurate. In addition, the inclusion of these clue words (e.g., the prepositions) makes the task closer to what would be done in inferring the relations from free text. The task thus approximates preposition disambiguation, using the Factotum relations as senses.

Figure 4.4 gives the feature settings used in the experiments. This is a streamlined version of the feature set from Figure 4.3, which is used in the Treebank and FrameNet experiments, to account for the lack of sentential context. Figure 4.5 contains sample feature specifications from the experiments discussed in the next section. The top part shows the original relationships from Factotum; the first example indicates that *connaturalize* causes *similarity*. Also included is the most likely relation marker inferred for each instance. This shows that 'n/a' is used whenever a preposition for a particular relationship cannot be inferred. This happens in the first example because 'connaturalize' is a rare term.

The remaining parts of Figure 4.5 illustrate the feature values that would be derived for the three different experiment configurations, based on the inclusion of word and/or hypernym collocations. In each case, the classification variable is given by *Relation*. For brevity, the feature specification only includes collocation features for the most frequent relations. Sample collocations are also shown for the relations (e.g., 'vulgarity' for *is-caused-by*). In the word collocation case, the occurrence of 'similarity' is used to determine that the *is-caused-by* feature (WC_1) should be positive (i.e., '1') for the first two instances. Note that there is no corresponding hypernym collocation due to conditional-probability filtering. In addition, although 'new' is not included as a word collocation, one of its hypernyms, namely *Adj:early#2*, is used to determine that the *has-consequence* feature (HC_3) should be positive in the last instance.

Features:

POS_{source}:	part-of-speech of the source term
POS_{target}:	part-of-speech of the target term
Prep:	preposition serving as relation marker ('n/a' if not inferable)
$WordColl_r$:	1 iff context contains any word collocation for relation *r*
$HypernymColl_r$:	1 iff context contains any hypernym collocation for relation *r*

Collocation selection:

Frequency:	$f(\text{word}) > 1$
Relative conditional probability:	$(P(C\|\text{coll}) - P(C))/P(C) \geq .20$
Organization:	per-class-binary grouping

Model selection:
Decision tree using Weka's J4.8 classifier (Witten and Frank, 1999)

Figure 4.4: ***Features used in Factotum role classification experiments.*** Simplified version of Figure 4.3: context is simply the source and target terms.

4.3.4.3 Results

To make the task more similar to the Treebank and FrameNet cases covered above, only the functional relations in Factotum are used. These are determined by removing the hierarchical relations (e.g., *has-subtype* and *has-part*) along with the attribute relations (e.g., *is-property-of*). In addition, in cases where there are inverse functions (e.g., *causes* and *is-caused-by*), the most frequently occurring relation of each inverse pair is used. This is done because the relation marker inference approach does not account for argument order. The boldface relations in the listing shown earlier in Table 4.5 are those used in the experiment. Only single-word source and target terms are considered to simplify the WordNet hypernym lookup. The resulting dataset has 5,959 training instances. The dataset also includes the inferred relation markers (e.g., one preposition per training instance), thus introducing some noise. Figure 4.5 includes a few examples from this dataset. This shows that the original relationship ⟨similarity, *is-caused-by*, rhyme⟩ from Factotum is augmented with the 'by' marker prior to classification. Again, these markers were inferred via web searches involving the terms from the original relationship.

Table 4.17 shows the results of the classification. The combined use of both collocation types achieves the best overall accuracy at 71.2%, which is good considering that the baseline of always choosing the most common relation (*is-caused-by*) is 24.2%. This combination generalizes well by using hypernym collocations, while retaining specificity via word collocations. The

Relationships from Factotum with inferred markers:

Relationship	Marker
⟨similarity, *is-caused-by*, connaturalize⟩	n/a
⟨similarity, *is-caused-by*, rhyme⟩	by
⟨approximate, *has-consequence*, imprecise⟩	because
⟨new, *has-consequence*, patented⟩	with

Word collocations only:

Relation	POS_s	POS_t	Prep	WC_1	WC_2	WC_3	WC_4	WC_5	WC_6	WC_7
is-caused-by	NN	VB	n/a	1	0	0	0	0	0	0
is-caused-by	NN	NN	by	1	0	0	0	0	0	0
has-consequence	NN	JJ	because	0	0	0	0	0	0	0
has-consequence	JJ	VBN	with	0	0	0	0	0	0	0

Sample collocations:

is-caused-by	{bitterness, evildoing, monochrome, *similarity*, vulgarity}
has-consequence	{abrogate, frequently, insufficiency, nonplus, ornament}

Hypernym collocations only:

Relation	POS_s	POS_t	Prep	HC_1	HC_2	HC_3	HC_4	HC_5	HC_6	HC_7
is-caused-by	NN	VB	n/a	0	0	0	0	0	0	0
is-caused-by	NN	NN	by	0	0	0	0	0	0	0
has-consequence	NN	JJ	because	0	0	0	0	0	0	0
has-consequence	JJ	VBN	with	0	0	1	0	0	0	0

Sample collocations:

is-caused-by	{N:hostility#3, N:inelegance#1, N:humorist#1}
has-consequence	{V:abolish#1, *Adj:early#2*, N:inability#1, V:write#2}

Combined collocations:

The combination of the above specifications:
that is, ⟨Relation, POS_s, POS_t, Prep, WC_1, ... WC_7, HC_1, ... HC_7⟩.

Legend:

POS_s & POS_t are the parts of speech for the source and target terms; and WC_r & HC_r are the word and hypernym collocations as follows:
1. *is-caused-by* 2. *is-function-of* 3. *has-consequence* 4. *has-result*
5. *is-caused-by*$_{\text{mental}}$ 6. *is-performed-by* 7. *uses*

Figure 4.5: ***Sample feature specifications for Factotum experiments.*** Each relationship from Factotum is augmented with one relational marker inferred via web searches. Collocation features are omitted for low-frequency relations.

Experiment	Accuracy	STDEV
Word	68.4	1.28
Hypernym	53.9	1.66
Combined	*71.2*	1.78

# Instances:	5959
# Classes:	21
Entropy:	3.504
Baseline:	24.2

Table 4.17: ***Functional relation classification over Factotum.*** This uses the relational source and target terms with inferred prepositions. The accuracy figures are averages based on 10-fold cross validation. The gain in accuracy for the combined experiment versus the word experiment is statistically significant at $p < .01$ (via a paired t-test).

classification task is difficult, as suggested by the number of classes, entropy, and baseline values all being comparable to the filtered FrameNet experiment (see Table 4.15).

4.3.5 Combining the Different Semantic Role Inventories

For the application to differentia disambiguation, the classifiers learned over Treebank, FrameNet and Factotum need to be combined. This can be done readily in a cascaded fashion with the classifier for the most specific relation inventory (i.e., FrameNet) being used first and then the other classifiers being applied in turn whenever the classification is inconclusive. This has the advantage that new resources can be integrated into the combined relation classifier with minimal effort. However, the resulting role inventory will likely be heterogeneous and might be prone to inconsistent classifications. In addition, the role inventory could change whenever new annotation resources are incorporated, making the overall differentia disambiguation system somewhat unpredictable.

Alternatively, the annotations can be converted into a common inventory, and a separate relation classifier induced over the resulting data. This has the advantage that the target relation-type inventory remains stable whenever new sources of relation annotations are introduced. In addition, the classifier will likely be more accurate as there are more examples per relation type on average. The drawback however is that annotations from new resources must first be mapped into the common inventory before incorporation.

The latter approach is employed here. The common inventory incorporates some of the general relation types defined by Gildea and Jurafsky (2002) for their experiments in classifying semantic relations in FrameNet using a reduced relation inventory. They defined 18 relations (including a special-case *Null* role for expletives), as shown in Table 4.18. Most of these roles are con-

Agent	Cause	Degree	Experiencer	Force	Goal
Instrument	Location	Manner	Null	Path	Patient
Percept	Proposition	Result	Source	State	Topic

Table 4.18: ***Abstract roles defined by Gildea and Jurafsky based on FrameNet.*** Taken from (Gildea and Jurafsky, 2002).

tained in the common relation inventory developed as part of this work to support the differentia disambiguation. 26 total relations are defined, including a few roles based on the Treebank, Cyc and Conceptual Graphs inventories. Table 4.19 shows this role inventory along with a description of each case. In addition to traditional thematic relations, this includes a few specialization-type relations. Specialization relations are prevalent in dictionary differentia, as noted in the previous chapter (see Section 3.1.2). For example, *Characteristic* corresponds to the general relation from Conceptual Graphs for properties of entities; and, *Category* generalizes the corresponding FrameNet role, which indicates category types, to subsume other FrameNet roles related to categorization (e.g., *Topic*).

To apply the common inventory to the FrameNet data, annotations using the 141 FrameNet relations (see Table 4.2) need to be mapped into those using the 26 common relations shown in Table 4.19.[8] Results for the classification of the FrameNet data mapped into the common inventory are shown in Table 4.20. As can be seen, the performance improves by 6 percentage points compared to the full classification over FrameNet (see Table 4.14). Although the low-frequency role filtering yields the highest performance, as shown in Table 4.15, this comes at the expense of having 5,000 training instances discarded. Corpus annotations are a costly resource, so such waste is undesirable.

This illustrates that the reduced, common-role inventory has an additional advantage of improving performance in the classification, compared to a cascaded approach. This occurs because several of the miscellaneous roles in FrameNet cover subtle distinctions that are not relevant for differentia disambiguation. The common inventory therefore strikes a balance between the overly general roles in Treebank, which are easy to classify, and the overly specialized roles in FrameNet, which are quite difficult to classify. Nonetheless, a certain degree of classification difficulty is inevitable in order for the inventory to provide adequate coverage of the different distinctions present in dictionary

[8] See www.cs.nmsu.edu/~tomohara/differentia-extraction/relation-mapping.html for the complete mapping. This covers the mapping of most of the other role inventories discussed in this chapter, such as from Conceptual Graphs, into the common inventory.

Relation	Description
Accompaniment	entity that participates with another entity
Agent	entity that acts on another entity
Amount	quantity used as a measure of some characteristic
Area	region in which the action takes place
Category	general type or class of which the item is an instance
Cause	entity that produces an effect
Characteristic	general properties of entities
Direction	either spatial source or goal (same as in Treebank)
Distance	spatial extent of motion
Duration	period of time that the situation applies within
Experiencer	entity undergoing some physical experience
Goal	location that the theme ends up in
Ground	background or context for situation or predication
Instrument	entity or resource facilitating event occurrence
IntervalOfTime	reference time interval for situation
Location	reference spatial location for situation
Manner	property of the underlying process
Means	action taken to affect something
Medium	setting in which the theme is conveyed
Path	trajectory which is neither a source nor a goal
PointInTime	reference time point for situation
Product	entity present at end of event (same as Cyc *products*)
Recipient	recipient of the resources
Resource	entity utilized during event (same as Cyc *inputs*)
Source	initial position of the theme
Theme	entity somehow affected by the event

Table 4.19: ***Inventory of semantic relations for differentia disambiguation.*** This inventory of common roles is primarily based on FrameNet (Fillmore et al., 2001) and Conceptual Graphs (Sowa, 1999); it also includes roles based on the Treebank and Cyc inventories.

differentia. Note that, by using the annotations from Treebank and FrameNet, the end result is a general-purpose classifier, not one tied into dictionary text. Thus, it is useful for other tasks besides differentia disambiguation.

Experiment	Accuracy	STDEV
Word Only	54.5	0.94
Hypernym	53.0	0.75
Combined	55.5	0.54

# Instances:	27295
# Classes:	31
Entropy:	4.006
Baseline:	15.0

Table 4.20: ***Results for preposition disambiguation with common roles.*** The FrameNet annotations are mapped into the common inventory from Table 4.19. See Table 4.10 for the legend.

4.4 Differentia Disambiguation Algorithm

To summarize the approach taken to differentia disambiguation, Figure 4.6 presents the high-level algorithm for the process. Differentia disambiguation is done in two main steps. First, the source and target terms are disambiguated. Next, the relation-indicating terms are disambiguated into semantic relations by applying the common-inventory relation classifier. The disambiguation of prepositional phrases has been illustrated in depth in this chapter. Support for other types of relation indicators is sketched out later in Chapter 6, building upon the relation marker inference technique used for Factotum.

Note that the relation disambiguation system just combines the semantic role data from Treebank and FrameNet. The integration of the data from Factotum is not addressed due to time constraints. The next chapter shows how the disambiguated relations facilitate lexical augmentation and word-sense disambiguation.

Input Definition text and list of extracted lexical relationships:
⟨source-word, relation-function-word, target-word⟩

Output List of conceptual relationships:
⟨source-concept, *relation-type*, target-concept⟩

Example Disambiguating relationships extracted from definition of 'kennel'
Input: "A kennel is an outbuilding that serves as a shelter for a dog."
⟨4. noun:outbuilding, 5. pronoun:that, 6. verb:serves⟩
⟨6. verb:serves, 10. prep:for, 12. noun:dog⟩
⟨6. verb:serves, 7. prep:as, 9. noun:shelter⟩
Output:
⟨noun:outbuilding#1, *Agent*, verb:serves#1⟩
⟨verb:serves#1, *Reason*, noun:dog#1⟩
⟨verb:serves#1, *Manner*, noun:shelter#1⟩

Steps For each relationship:

1. Disambiguate the source and target words.
   For WordNet, this just incorporates the word-sense annotations provided by Extended WordNet. Application to other dictionaries require use of one of the WSD algorithms outlined in Section 2.4.3.
2. Disambiguate the relation function word.
   - Convert definition text into untagged annotation format:
     pre-context ⟨wf sense="?"⟩*function-word*⟨/wf⟩ *post-context*
     Example:
     A kennel is an outbuilding that serves ⟨wf sense="?"⟩as⟨/wf⟩ a shelter for a dog.
   - Run common-inventory relation classifier to determine the semantic role serving as the sense for the function word.
3. Consolidate the results of relation disambiguation with that of term disambiguation.
   This is a bookkeeping step necessary to coordinate the two different disambiguation systems.

Figure 4.6: ***Differentia disambiguation algorithm.*** Step 1 is addressed in Section 4.1; and step 2 is covered in the previous section. The final step is not discussed here but is documented in the program source available at www.cs.nmsu.edu/~tomohara/differentia-extraction.

CHAPTER 5
APPLICATION AND EVALUATION

To illustrate the usefulness of the differentiating relationships extracted and disambiguated using the methods from the previous chapters, two distinct application areas are discussed here, including detailed evaluations. The first area involves the use of this information to augment existing lexicons (i.e., lexicon augmentation). The differentiating information is directly evaluated by having humans assess the quality of random samples from the extracted relationship listings. The second area is word-sense disambiguation. The differentiating information is indirectly evaluated by comparing the performance of systems utilizing the differentia versus those based on typical approaches.

This chapter is organized as follows. Section 5.1 discusses the qualitative evaluation of the system output, including an inter-coder reliability analysis of the human judges. Section 5.2 presents results from two distinct word-sense application systems that utilize the differentiating relations to improve performance.

5.1 Lexicon Augmentation

From a lexical semantics point of view, the main purpose of this thesis is to augment existing semantic lexicons for natural language processing. Therefore, the first evaluation determines the quality of the information that would be added to the lexicons, in particular with respect to relation disambiguation as that is the focus of the research.

5.1.1 Overview of Extracted Relations

All the definitions from WordNet 1.7.1 were run through the differentia-extraction process. This involved a total of 111,223 synsets as shown earlier in Table 3.2. 10% of these had preprocessing or parse-related errors leading to no relations being extracted.

Table 5.1 shows the frequency of the relations that occur in the output from the differentia extraction process. The most common relation used is *Theme*, which accounts for four times as many of the cases as it does among the annotations. In the annotations, it is most often being tagged as the sense for 'of,' which also occurs significantly with roles *Source*, *Category*, *Ground*, *Agent*, *Characteristic*, and *Experiencer*. Some of these represent subtle distinctions, so it is likely that the difference in the text genre is causing the classifier to use the default case more often as a default. Note that *Theme* is a very

Abbreviation	Relation	Frequency
THME	Theme	0.316
GOAL	Goal	0.116
GROUND	Ground	0.080
CAT	Category	0.069
AGNT	Agent	0.069
CAUSE	Cause	0.061
MANR	Manner	0.058
RCPT	Recipient	0.053
MED	Medium	0.039
CHRC	Characteristic	0.022
RESOURC	Resource	0.021
MEANS	Means	0.021
SOURCE	Source	0.019
PATH	Path	0.017
EXPR	Experiencer	0.017
ACCM	Accompaniment	0.011
AREA	Area	0.010
DIR	Direction	0.001

Table 5.1: ***Frequency of extracted relations after disambiguation.*** WordNet definitions are analyzed with relations disambiguated with respect to the common relation inventory (Table 4.19), yielding about 19,000 total relationships.

generic relation that subsumes most of the other relations. Therefore, this type of overgeneration does not pose a problem with respect to lexicon augmentation.

Table 5.1 also shows that the specialization relations (e.g., *Category* and *Characteristic*) are more predominant in the extracted relations than in the data for the annotations (see Table 4.19). This is similar to the situation with the WordNet definitions annotations, where the specialization relations occur more often than in general text, reflecting the differentiating nature of definitions (discussed earlier in Section 3.1.2).

Figure 5.1 shows a random sample from the output of the system. This omits relationships due to modification and other types not considered during the disambiguation process. Therefore, no extracted relations are listed for the last case (*verb:lunge#1*).

noun:weaning#1: A weaning is the act of substituting other food for the mother 's milk in the diet of a child or young mammal.
    of#THME noun:child#1 (0.001953125)
    of#THME noun:mammal (0.001953125)
    n/a verb:weaning (0.0009765625)
    verb:is noun:act#2 (4.296875e-05)
        of#THME verb:substituting#3 (0.00390625)
            unknown:for#SOURCE noun:milk (0.00390625)

noun:area#1: An area is a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography).
    n/a verb:serving (0.0009765625)
    by#AGNT noun:people (0.0029296875)
    by#AGNT noun:culture (0.0029296875)
    by#AGNT noun:geography (0.0029296875)

noun:Bavaria#1: A Bavaria is a state in southwestern Germany famous for its beer; site of automobile factory.
    n/a noun:beer#1 (0.0009765625)
    n/a noun:state#2 (0.0001396484375)
        unknown:for#THME unknown:famous#1 (0.0009765625)

noun:slowdown#1: A slowdown is the act of slowing down or falling behind.
    verb:is noun:act#2 (0.000125)
        of#GROUND verb:falling#3 (0.001953125)
        of#GROUND verb:slowing (0.001953125)

verb:lunge#1: To lunge is to make a thrusting forward movement.
    n/a

Figure 5.1: ***Sample lexical relations extracted by system.*** Definitions from WordNet 1.7.1 used as input to process described in Figures 3.8 and 4.4. Syntactic relations such as modification are omitted for sake of brevity. Relation strengths derived via cue validity analysis are shown in parentheses.

5.1.2 Qualitative Evaluation

Six human judges were recruited to evaluate random samples of the relations that were extracted. Four are graduate students in computer science

with exposure to computational linguistics, and two are computer programmers with backgrounds in business and humanities, respectively. Each was given about 70 relationships to evaluate. To allow for inter-coder reliability analysis, each evaluator rated 35 samples that were also evaluated by the others, 25 as part of a training phase and the rest after training. In addition, they also evaluated 20 samples that were manually corrected beforehand. This provides a baseline against which the uncorrected results can be measured.

Because this thesis only addresses relations indicated by prepositional phrases, the evaluation is restricted to these cases. The evaluation also does not directly account for aspects related to prepositional attachment, although incorrect attachment decisions by the parser do negatively affect the evaluation. The judges only rated the assignment of relations to the prepositional phrases on a scale from 1 to 5, with 5 being an exact match. They were presented with the list of common relation types shown in the last chapter (Table 4.19), so that correctness is relative to this relation inventory.

For example, consider the 'kennel' example from the last chapter:

an outbuilding that serves as a shelter for a dog
⟨verb:serves#1, *Reason*, noun:dog#1⟩
⟨verb:serves#1, *Manner*, noun:shelter#1⟩

In this case, the *Reason* assignment might be rated as 3 since *Recipient* is more appropriate, and the *Manner* might be rated as 4 since *Goal* is a better relation to account for purpose. Figure 5.2 shows the instructions given to the judges. These are informal in nature, so as not to burden the judges in performing their task. This is appropriate given the volunteer nature of the evaluation; however, more detailed instructions are usually preferable to achieve better uniformity.

5.1.2.1 Inter-coder Reliability Analysis

To assess the reliability of the evaluations, the *kappa statistic* was calculated (Carletta, 1996):

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

This determines the extent to which the coders agree (P_a) less that which is due to chance agreement (P_e). Kappa can thus be negative if the actual agreement is less than that due to chance. Figure 5.3 shows how the intermediate agreement values are calculated for the case of two coders. With three or more

Instructions for evaluation of extracted relations
Please evaluate the following conceptual relation listing extracted from dictionary definitions. The evaluation is restricted to relations indicated by prepositions. So in each case, indicate the appropriateness of the relation on a scale of 1 (poor) to 5 (good). The listing contains a form for each of the cases to be annotated, such as the following:

⟨noun:mouth, *of#GOAL*, noun:river⟩ (bad) 1 __ 2 __ 3 __ 4 __ 5 __ (good)

The evaluation should be based on the selection of one of the 26 relations shown in the next section to serve as the meaning of the preposition. Please review the descriptions carefully before proceeding. In some cases, none of the relations might be a close match, so evaluate the extent to which the listed relation approximates the ideal one. In such cases where no relation is suitable, add a brief comment explaining what was expected, as done in the first example below.

The relation description is followed by a sample of five annotated definitions to give you an idea of the task. At the end is the actual sample of relationships to be evaluated.

Relation descriptions
See Table 4.19.

Sample annotations
sense: verb:repel#2
sentence: To repel is to be repellent to ; cause aversion in .
conceptual relations:

⟨verb:repel, *to#CAUSE*, verb:be⟩ (bad) 1 x 2 _ 3 _ 4 _ 5 _ (good)
⟨verb:repel, *has-object-2-6*, noun:repellent⟩
⟨noun:repellent, *to#CAUSE*, verb:cause⟩ (bad) 1 x 2 _ 3 _ 4 _ 5 _ (good)
⟨verb:cause, *has-object-9-10*, noun:aversion⟩

comments:
'to' is infinitive marker

sense: noun:reason#2
sentence: Reason is an explanation of the cause of some phenomenon .
conceptual relations:

⟨noun:explanation, *of#GROUND*, noun:cause⟩ (bad) 1 _ 2 _ 3 _ 4 x 5 _ (good)
⟨noun:cause, *of#GROUND*, noun:phenomenon⟩ (bad) 1 _ 2 _ 3 _ 4 x 5 _ (good)

comments:

...

Figure 5.2: ***Instructions for evaluation of extracted relations.*** Excerpt from the annotation instructions, omitting the actual relations to be judged. Four other sample annotations are included, along with the common-relation inventory table (Table 4.19).

Let G = #categories, n_{ij} the disagreements for categories i and j, and w_{ij} the weight for disagreement involving categories i and j

N	number of codings	$\sum_{i=1}^{G}\sum_{j=1}^{G} n_{ij}$
$p(r_i)$	row marginal probability for category i	$\frac{1}{N}\sum_{j=1}^{N} n_{ij}$
$p(c_j)$	column marginal probability for category j	$\frac{1}{N}\sum_{i=1}^{N} n_{ij}$
P_a	actual agreement	$\frac{1}{N}\sum_{i=1}^{G}\sum_{j=1}^{G} w_{ij} n_{ij}$
P_e	expected agreement	$\sum_{i=1}^{G}\sum_{j=1}^{G} w_{ij} p(r_i) p(c_j)$

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

Figure 5.3: ***Calculation of the weighted kappa statistic.*** Based on (Altman, 1997).

coders, the pairwise kappa scores are averaged. The figure shows the generalization of kappa that accounts for partial agreement in the case of ordinal data. A weight from 0 to 1 is defined for each pair of categories, indicating the amount of partial credit to be assigned when the coders chose the respective categories. Normally the weights are based on the absolute value of the difference in the ordinals or on the squared differences. Standard kappa can be viewed as defining a weight function that is 1 only when the category values are the same and otherwise 0.

Carletta (1996) indicates that a κ value of 0.8 or greater suggests a high level of reliability among raters, with values between 0.67 and 0.8 suggesting only moderate agreement. Altman (1997) provides an alternative interpretation, with .41 to .60 being considered moderate and .61 to .80 as good. Table 5.2 shows the results of the inter-coder reliability analysis over the judge ratings given to the manually corrected subset of the samples evaluated. The weighted kappa statistic uses the squared distance measure as implemented via the LKAPPA function in the R statistical environment (R Team, 2004). The overall kappa score is somewhat low even with weighting (.291), partly due to the difficulty of selecting specific scores (e.g., 4 versus 5). To alleviate this problem, the scores are converted into 3 values (*bad*, *ok* and *good*) by combining values 1 and 2 as well as 4 and 5. In this case, the weighted kappa measure increases to .400. Although the weighting scheme might need to be revised to better

Training phase

Groups	Cases	κ	κ_w
5	25	.079	.269
3	25	.231	.283

Final phase

Groups	Cases	κ	κ_w
5	10	.078	.291
3	10	.274	.400

Table 5.2: ***Inter-coder reliability analysis for evaluation.*** *Groups* gives the number of ranges for the assessment scores: with *3* categories, scores 1&2 and 4&5 are treated the same. *Cases* is the number of distinct assessments per coder. κ gives the kappa statistic, and κ_w the weighted kappa statistic.

Corrected

Metric	Training	Final
Cases	10	10
Scores	60	60
Mean	3.433	3.650
STDEV	1.466	1.424

Uncorrected

Metric	Training	Final
Cases	15	35.8
Scores	90	243
Mean	3.011	3.263
STDEV	1.510	1.465

Table 5.3: ***Mean assessment score for all extracted relationships.*** *Corrected* shows assessments over manually corrected output, whereas *Uncorrected* evaluates the system output as is. *Cases* is the (average) number of distinct relationships judged, and *Scores* is the number of individual assessments (i.e., total cases). *Mean* gives the mean of the assessment ratings (from 1 to 5), and STDEV is the corresponding standard deviation.

model the intuitive qualitative differences in the assessed scores, this suggests moderate agreement.

5.1.2.2 Results

The overall evaluation is based on averaging the assessment scores over the relationships. Table 5.3 shows the results from this evaluation, over the manually corrected and uncorrected subsets of the relationships. For the corrected output, the mean assessment value was 3.650, whereas for the uncorrected system output, the mean assessment value was 3.263. Therefore, although the absolute score is not high, the system's output is generally acceptable, especially given that the score for the uncorrected set of relationships is close to that of the manually corrected set.

5.2 Word Sense Disambiguation

Two different approaches are used for word-sense disambiguation (WSD). The first is based on a typical supervised approach for WSD using tagged training data to induce a statistical classifier for each word to be disambiguated. The second is a hybrid supervised/unsupervised approach that incorporates knowledge from WordNet to provide a model of the dependencies among word senses. It can be used to propagate empirical support derived from annotated corpora to related senses for which there might not be training data.

5.2.1 Supervised Classification

The typical approach using statistical classification for word-sense disambiguation was illustrated in Section 2.4.3.1. The same general approach is used here, but additional features based on differentia are included. In particular, a new type of collocation feature is included that checks whether words related to those occurring in the context of the target words are indicators of a particular sense. This adds a level of indirection to the standard word collocation selection scheme discussed in Section 4.3.1.1, allowing for words not occurring in the training data context to be included as collocations.

5.2.1.1 Feature Overview

Figure 5.4 shows the feature settings that are used in this application. Five of the feature variables are based on part of speech (*POS*$\pm$*i* for *i* from *-2* to *+2*). The *POS+0* feature is labeled *Morph* in Figure 2.3, since it uses the full set of part-of-speech tags rather than those corresponding to the traditional grammatical categories (e.g., noun and verb). There are also four adjacency-based collocational features (*Word*$\pm$*i* for i from -2 to $+2$), which were found to be beneficial in other work (Pedersen and Bruce, 1998; Ng and Lee, 1996).

The collocation variable $WordColl_s$ for each sense *s* is binary, corresponding to the presence (or absence) of any word in a set specifically chosen for *s*. A word *w* is considered as a collocation for sense *s* if the relative percent gain in the conditional probability over the prior probability is 20% or higher:

$$\frac{(P(s|w) - P(s))}{P(s)} \geq .20.$$

However, if the word only occurs once in the training data, it is ignored. This is a variation of the *per-class, binary organization* and the *conditional probability test* used by Wiebe et al. (1998a)

Features:

Morph:	morphology of the target word (i.e., part of speech)
POS$-$i:	part-of-speech of *i*th word to left
POS$+$i:	part-of-speech of *i*th word to right
Word$-$i:	*i*th word to the left
Word$+$i:	*i*th word to the right
WordColl_s:	occurrence of word collocation for sense *s* in context
RelatedColl_s:	occurrence of differentia related-word collocation for sense *s*

Collocation selection:

Word context:	anywhere in the sentence
Word collocation frequency:	$\text{f}(\text{word}) > 1$
Related-Word collocation frequency:	$\text{f}(\text{related-word}) > 4$
Conditional probability:	$\text{p}(\text{c}\|\text{coll}) >= .50$
Relative conditional probability (RCP):	$(\text{p}(\text{c}\|\text{coll}) - \text{p}(\text{c}))/\text{p}(\text{c}) >= .20$
Related-Word RCP:	$(\text{p}(\text{c}\|\text{coll}) - \text{p}(\text{c}))/\text{p}(\text{c}) >= .80$
Feature organization:	per-class-binary

Model selection:

Decision tree via Weka's J4.8 classifier (Witten and Frank, 1999)

Figure 5.4: ***Features for word-sense disambiguation with differentia.*** The *RelatedColl*$_s$ features represent the main change from those used in Figure 2.3.

Note that the use of the relative-percent-gain ratio is different from most other approaches to deriving sense-specific collocations, where usually just an absolute conditional probability threshold is used (e.g., $\text{P}(\text{s}|\text{w}) > .5$). The purpose is to account for cases when the prior probability is high to begin with. For example, if a sense occurs 70% of the time in the training data, then the conditional probability for words entirely independent of the sense would also be .70. Requiring a relative percent gain of .20 restricts potential collocations for which the conditional probability is .84 or higher.

5.2.1.2 Differentia-based Collocational Features

The relations extracted via differentia analysis are used to determine the semantic relatedness of words. Other sources for relatedness could be considered, including corpora and the WordNet hierarchy. Here the purpose is just to show that information derived from differentia analysis is useful for supervised word-sense disambiguation. O'Hara et al. (2004) illustrate other types of class-based collocations for WSD, as part of research done for Senseval III.

The context words are not disambiguated, so the relations for separate senses of the same word are conflated. When determining the potential col-

locations, the words strongly related to each context word in the training data are considered when tabulating the frequencies $f(c, coll)$ used in estimating the conditional probability table $P(c|coll)$. Instead of using a unit weight for each occurrence, the relation weight is used. In addition, a given related-collocation word might occur with more than one co-occurring word for the same sense, so the contributions are added. Afterwards, the conditional probability of the class given the relatedness collocation is estimated by dividing the weighted frequency by the sum of all such weighted co-occurrence frequencies for the class:

$$P(c|coll) \simeq \frac{wf(c, coll)}{\sum_{c_i} wf(c_i, coll)}$$

Here *wf(c, coll)* stands for the weighted co-occurrence frequency of the related-word collocation *coll* and class *c*.

A similar conditional probability test as before is used. However, the related-word collocations are less reliable given the level of indirection involved in their extraction. Therefore, tighter constraints are used in order to filter out extraneous potential collocations. In particular, the relative percent gain in the conditional probability over the prior probability must be 80% or higher. In addition, the words they are related to must occur more than four times in the training data. Recall that the related-word collocations themselves do not necessarily occur in the training data. Since they might be related to several different co-occurring words, the total number of distinct training instances need not be five. In fact, there might just be one training instance in case there are five different context words related to the same potential collocation.

5.2.1.3 System Results

Tables 5.4, 5.5, and 5.6 show the results of classifying the word-sense annotations from the Senseval II data for nouns, verbs, and adjectives, respectively. The entries are ordered by entropy, which measures the uniformity of the sense distribution (Manning and Schütze, 1999) and hence the general difficulty expected during classification. In each case, accuracy results are given for systems with and without the relatedness collocation features (*RelatedColl*$_s$), which are derived from the relations extracted from the WordNet definitions for the target words. The performance results for both systems are fairly close with the relatedness collocations leading to slight improvements. Overall, the differentia-based system achieves 63.8% accuracy versus 63.4% accuracy for the typical system, with a baseline accuracy of 57.7%.

Noun	Senses	Freq	Entropy	Baseline	-Diff	+Diff
bar	11	283	2.340	0.516	0.588	0.560
post	9	146	2.331	0.438	0.562	0.594
nature	5	90	2.003	0.489	0.569	0.534
channel	9	86	1.963	0.628	0.707	0.681
sense	5	107	1.905	0.393	0.533	0.517
stress	6	79	1.852	0.557	0.449	0.547
material	5	139	1.821	0.439	0.451	0.487
hearth	4	61	1.738	0.443	0.652	0.671
authority	8	162	1.698	0.623	0.783	0.764
art	4	169	1.676	0.503	0.731	0.727
mouth	8	117	1.648	0.504	0.452	0.497
restraint	6	91	1.614	0.659	0.707	0.723
facility	4	113	1.584	0.504	0.577	0.532
circuit	6	167	1.584	0.611	0.859	0.891
day	6	288	1.555	0.677	0.654	0.644
feeling	4	102	1.501	0.539	0.332	0.438
fatigue	6	84	1.476	0.583	0.846	0.858
bum	5	89	1.318	0.719	0.685	0.672
spade	4	62	1.192	0.677	0.755	0.712
church	3	111	1.095	0.568	0.712	0.695
grip	6	102	1.037	0.814	0.883	0.862
child	4	125	1.012	0.680	0.680	0.667
lady	4	90	0.906	0.822	0.742	0.773
detention	2	57	0.804	0.754	0.991	0.960
nation	3	57	0.797	0.825	0.771	0.724
dyke	2	50	0.722	0.800	0.752	0.787
chair	4	138	0.694	0.877	0.881	0.880
yew	2	54	0.605	0.852	0.842	0.862
holiday	2	61	0.208	0.967	0.967	0.967
Total				0.637	0.694	0.697

Table 5.4: ***Supervised WSD results over Senseval II noun training data.*** *Senses* is number of word senses; *Freq* gives the number of training instances, and *Entropy* measures the non-uniformity of the sense distributions. Accuracy results are given averaged over ten-fold cross validation: *Baseline* selects the most-frequent sense; *-Diff* uses all the features from Figure 5.4, except for the relatedness collocations derived from differentia ($RelatedColl_s$); and, *+Diff* includes the relatedness collocations as well.

Verb	Senses	Freq	Entropy	Baseline	-Diff	+Diff
draw	21	62	3.928	0.177	0.214	0.143
find	14	122	3.530	0.172	0.288	0.287
play	19	119	3.403	0.210	0.280	0.366
strike	14	86	3.366	0.198	0.268	0.235
carry	19	102	3.290	0.304	0.332	0.285
turn	15	76	3.272	0.263	0.508	0.548
see	17	128	3.173	0.320	0.402	0.411
develop	15	133	3.120	0.301	0.322	0.347
call	13	107	3.013	0.308	0.356	0.386
serve	11	99	2.953	0.263	0.429	0.386
leave	11	127	2.909	0.299	0.396	0.419
keep	15	112	2.887	0.429	0.338	0.323
work	12	96	2.676	0.344	0.327	0.329
train	9	125	2.564	0.272	0.435	0.485
drive	9	76	2.464	0.355	0.457	0.486
pull	10	69	2.440	0.391	0.283	0.317
match	8	86	2.375	0.360	0.278	0.343
drift	7	58	2.294	0.328	0.318	0.215
wash	6	16	2.233	0.375	0.500	0.633
treat	6	88	2.158	0.318	0.423	0.439
dress	10	87	2.118	0.517	0.645	0.618
live	6	116	1.864	0.569	0.600	0.615
begin	8	557	1.768	0.591	0.765	0.760
use	6	146	1.631	0.678	0.650	0.663
replace	4	86	1.569	0.512	0.525	0.473
face	7	186	1.056	0.833	0.814	0.818
wander	4	100	0.856	0.830	0.819	0.802
collaborate	3	57	0.719	0.860	0.922	0.907
Total				0.406	0.460	0.466

Table 5.5: ***Supervised WSD results over Senseval II verb training data.*** See Table 5.4 for the legend.

5.2.2 Probabilistic Spreading Activation

Spreading activation has been a popular technique in artificial intelligence for propagating support throughout a semantic network. A variety of approaches have been developed for doing this (Ide and Véronis, 1998), such as link counting techniques, marker passing, and approaches accounting for

Adjective	Senses	Freq	Entropy	Baseline	-Diff	+Diff
cool	8	94	1.629	0.649	0.737	0.779
fine	8	135	1.357	0.748	0.837	0.824
natural	8	200	1.322	0.765	0.828	0.854
simple	5	130	1.245	0.746	0.803	0.790
blind	5	102	1.239	0.725	0.795	0.786
fit	3	56	1.206	0.661	0.923	0.840
oblique	3	57	1.106	0.526	0.707	0.650
green	6	184	0.930	0.804	0.935	0.926
free	6	152	0.916	0.842	0.862	0.855
colorless	2	68	0.787	0.765	0.791	0.789
vital	4	74	0.728	0.865	0.930	0.932
graceful	2	56	0.592	0.857	0.760	0.797
local	3	75	0.539	0.907	0.926	0.950
solemn	2	52	0.457	0.904	0.860	0.887
faithful	2	47	0.149	0.979	0.975	0.990
Total				0.783	0.845	0.846

Table 5.6: ***Supervised WSD results over Senseval II adjective training data.*** See Table 5.4 for the legend.

fan-in and fan-out of nodes. Here Bayesian networks are used to implement a probabilistic version of spreading activation. They are used because they integrate well with empirical classifiers, such as the one just discussed above.

5.2.2.1 Bayesian Network Representation

The properties conveyed by dictionary differentia involve relations of varying strengths. Consider the WordNet definitions of ‘lock’ and ‘key’:

key: metal device shaped in such a way that when it is inserted into a lock the lock’s mechanism can be rotated
lock: a fastener fitted to a door or drawer to keep it firmly closed

The definition for ‘key’ indicates a strong relationship to $lock_{\text{FASTENER}}$, but the definition for ‘lock’ only indicates a moderate relationship to $door_{\text{BARRIER}}$. Therefore, a probabilistic representation is useful for representing the differentia. Moreover, given the asymmetries in the relationships, a *Bayesian network* (Pearl, 1988) is appropriate. Basically, Bayesian networks are acyclic, directed graphs in which nodes are associated with probability tables indicating the probabilistic

relation of their values to those of their parent nodes. See Appendix B for a brief primer on Bayesian networks.

Two important issues concern the use of Bayesian networks: 1) the interpretation of the links; and, 2) the derivation of the conditional probability tables (CPT's). Causality is the dominant type of link interpretation for Bayesian networks, because it is prevalent in medical domains (e.g., cold causes runny nose), which were the first major application area. However, interpreting differentia in terms of causality is possible but awkward (e.g., beagle causes small-size). Instead, *salience* is used to quantify how relevant a concept is for another, when considered as an attribute. That is, the salience value measures the degree to which an attribute is characteristic of the object. This notion is based on the psychological usage of salience for determining which properties are relevant for comparisons (Medin et al., 1993). Note that using salience rather than causality accords with the broader notion of causality discussed by Lauritzen and Spiegelhalter (1988, p. 160):

> 'Causality' has a broad interpretation as any natural ordering in which knowledge of a parent influences opinion concerning a child—this influence could be logical, physical, temporal or simply conceptual in that it may be most appropriate to think of the probability of children given parents.

A problem related to link interpretation is the directionality required in order to support belief propagation. As discussed by Wiebe et al. (1998b), it might be necessary to invert the logical direction, because evidence propagation generally occurs among the nodes for the children (not the parents).

The main problem with CPT derivation deals with how salience should be quantified. Although predefined salience values for each relation type can be determined based on intuitive judgment, a more empirical approach is desirable. One approach would be an extension of the idea of using the information retrieval term-weighting technique based on term frequency and inverse document frequency (*TF*IDF*), as proposed by Richardson (1997). To apply this to dictionary text, documents are defined as the cluster of definitions that refer to a particular headword. However, this does not model salience well because it does not take into consideration categories similar to the one being defined.

Here the relation weights are based on *cue validities*, which were discussed earlier in Chapter 3. The cue validity of a feature *F* for concept *C* is calculated as:

$$P(C|F) = \frac{P(F|C)}{\sum_i P(F|C_i)}$$

where C_i is a concept that contrasts with C. To determine the contrasting concepts for a given concept, its most-informative ancestor is first estimated based on frequency considerations using SemCor (Miller et al., 1994), as shown in Section 3.3.2. Other concepts subsumed by this most-informative ancestor are then considered as contrasting.

Another problem with CPT derivation concerns how to handle converging links (i.e., causal interactions among parent nodes in the Bayesian network). Several different models of causal independence were considered (Heckerman and Breese, 1994). The basic idea is to treat multiple causes as independent, so that interactions need not be quantified. To do this a model is chosen for determining how the individual cause strengths are combined. For instance, using the *Noisy-OR* model, the weights are combined in a manner analogous to the Boolean *OR* function (Wiebe et al., 1998b). To see how this is defined, first consider that the inclusive-or connective can be viewed as outputting a true value only when not all of the inputs are false:

$$\mathsf{output} = \neg((\neg \mathsf{v}_1) \wedge \cdots \wedge (\neg \mathsf{v}_n))$$

where each v_i is a logical-valued input variable. The extension to the case where probabilities are associated with each input is relatively straightforward:

$$\mathsf{child} = \neg((\neg \mathsf{v}_1) \wedge \cdots \wedge (\neg \mathsf{v}_n))$$
$$\mathsf{P}(\mathsf{child}|\mathsf{V}_1 = \mathsf{v}_1, ..., \mathsf{V}_n = \mathsf{v}_n) = 1.0 - \prod(1.0 - \mathsf{P}(\mathsf{child}|\mathsf{v}_i)), \qquad \forall_{\mathsf{V}_i} \mathsf{v}_i = \mathsf{T}.$$

The child's value is still an inclusive-or function of the parents. The probability of a positive value is 1 less the joint probability that the parents are negative, for those that are activated. Note that the use of these causal independence models is a simplifying assumption. Future work will investigate whether the dependencies can be induced from the data.

To illustrate the network structure that is derived from lexical relations, a small network was manually constructed based on the definitions for a few types of hounds, as shown in Figure 5.5. The figure also shows the relationships extracted from the definitions. The resulting semantic network is shown in Figure 5.6.

The Bayesian network representation for the relationships involving the hounds would have the same structure as that in Figure 5.5. To complete the Bayesian network specification, the CPT tables have to be defined. A variation of the Noisy-OR model is used for this purpose: each case is approximately the sum of the relation strengths for the incoming links derived via the cue validity measures. More precisely, given a node N with parents P_i, each row

hound: a hunting dog typically having large drooping ears
basset: a small hound with short legs
beagle: a small hound with a smooth coat
wolfhound: a large hound with a rough coat
greyhound: a large slender hound used as a racing dog
whippet: a small greyhound found in England
Italian greyhound: a very small greyhound

```
hound:
    is-a      dog (1.00)
    has-part      ears (1.00)
        size      large (0.33)
        attr      drooping (1.00)
    purpose       hunting (1.00)

basset:
    is-a      hound (0.25)
        size      small (0.25)
    has-part      legs (1.00)
        size      short (1.00)

beagle:
    is-a      hound (0.25)
        size      small (0.25)
    has-part      coat (0.50)
        attr      smooth (1.00)

wolfhound:
    is-a      hound (0.25)
        size      large (0.33)
    has-part      coat (0.50)
        attr      rough (1.00)

greyhound:
    is-a      hound (0.25)
        size      large (0.33)
        girth     slender (1.00)
    purpose       dog (1.00)
        attr      racing (1.00)

whippet:
    is-a      greyhound (0.50)
        size      small (0.25)
    location      England (1.00)

Italian greyhound:
    is-a      greyhound (0.50)
        size      small (0.25)
            attr      very (1.00)
```

Figure 5.5: ***Lexical relations for sample hound definitions.*** The definitions are simplified versions of the corresponding WordNet definitions. The relations were manually extracted and weighted using the cue validity process.

of the CPT table for *N* has a value based on the sum of the corresponding relation strengths for those P_i's that are positive (i.e., $P(N|...P_i = True...)$), with normalization to ensure proper probabilities. More details on this derivation process are given in the next section.

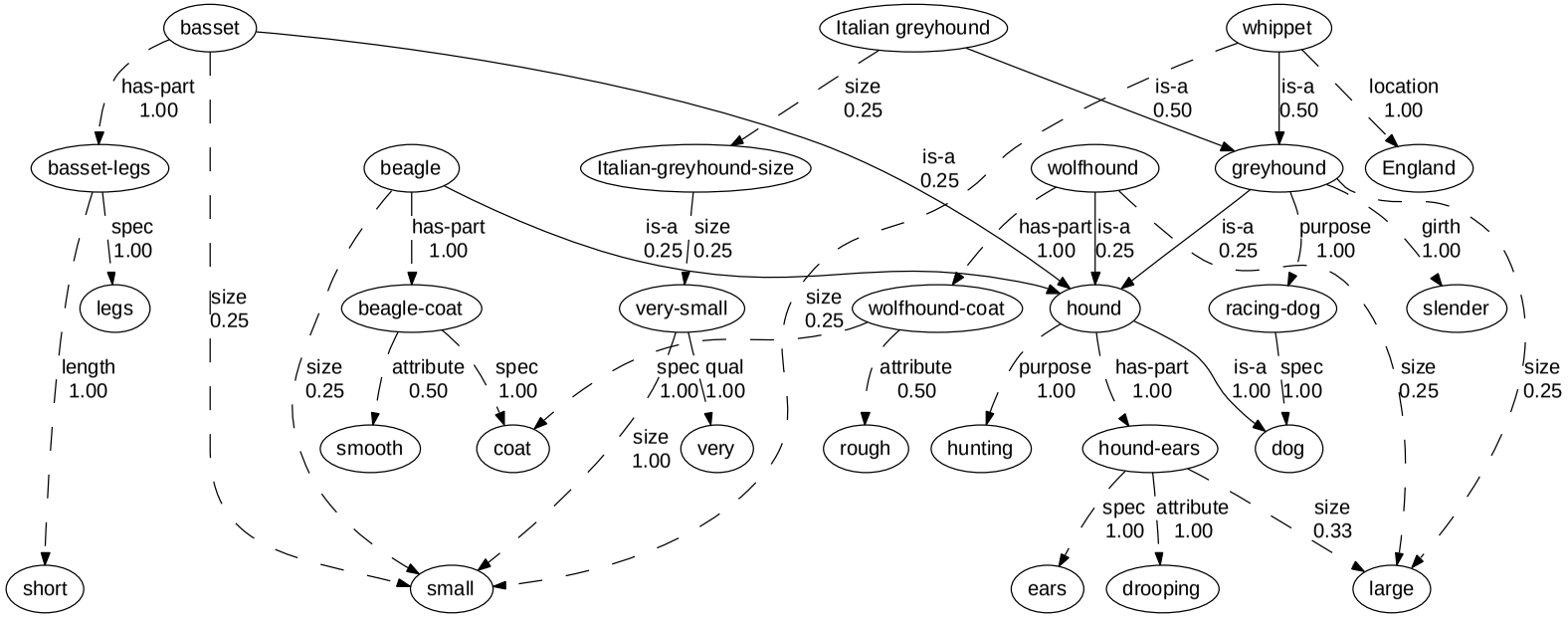


Figure 5.6: ***Semantic network of sample hound lexical relations.*** Bayesian network structure for relations in Figure 5.5. Solid links indicate *is-a* relations, and dashed links indicate differentiating relations.

As can be seen from the figure, embedded attributes are treated as separate nodes. Therefore, [*beagle* $\rightarrow_{\text{HAS-PART}}$ *coat* $\rightarrow_{\text{ATTRIBUTE}}$ *smooth*] is modeled as [*beagle* $\rightarrow_{\text{HAS-PART}}$ *beagle-coat*] and [*beagle-coat* $\rightarrow_{\text{ATTRIBUTE}}$ *smooth*]. This use of object-specific attribute nodes is inspired by the work by Koller and Pfeffer (1998) on probabilistic frame systems. Although it would be possible to use their system, they interpret the probabilities as the distribution of the possible values, not their salience. Embedded-attributes nodes are important when dealing with large networks, since otherwise the size of the underlying cliques might make direct evaluation intractable (Lauritzen and Spiegelhalter, 1988).

Although other probabilistic representations are possible, Bayesian networks offer advantages in terms of tractability and software availability. For example, by using undirected nodes as in *Markov Networks* (Pearl, 1988), problems with circularity are avoided; however, there are few implementations of Markov networks available, and this ignores the important information regarding directionality. It is also possible to use a more general representation to distinguish uncertainty from disbelief, such as in the Dempster-Shafer theory (Shafer, 1976; Shafer, 1987). This can be viewed as a switch from a point-based probability specification into an interval-based specification, allowing for three possibilities: belief, disbelief, and uncertainty. Naturally, the added flexibility complicates inferencing, making the representation less tractable than Bayesian networks. It is still an open issue whether general belief functions are necessary rather than just standard probabilities. See (Lindley et al., 1987) for a debate on the suitability of both for artificial intelligence, in particular expert systems. In addition, Almond (1995) explored the use of graphical belief functions for his dissertation, but he is unsure that the added expressivity is worth the extra computational costs.

Extensions to standard Bayesian networks are also possible. As mentioned above, Koller and Pfeffer (1998) have developed probabilistic frame systems, which are used to provide more structure to Bayesian network models. This also allows for better integration with existing knowledge bases.

5.2.2.2 System Overview

The application to word-sense disambiguation builds upon the framework laid out in (Wiebe et al., 1998b), which augments a traditional statistical classifier with probabilistic spreading activation. In particular, they use belief propagation in Bayesian networks to model the activation of similar word senses; the network is initialized with the local contextual support determined by the statistical classifier (similar to the supervised WSD system described above in Section 5.2.1). Their model can be viewed as propagating support

only along lines of *strong semantic similarity*, namely among nodes having a common ancestor in an *is-a* hierarchy. This is extended here to semantic relatedness (or *weak semantic similarity*) by relaxing the restriction on the paths along which the activation can occur. Specifically, the set of paths is expanded to include those incorporating differentia-based relations.

For this application, ambiguity in the dictionary differentia generally leads to degraded performance, so only those properties involving specific senses are considered. That is, if the Extended WordNet data did not have sense annotations for a particular word in a definition, relations involving it are not included. This occurs for less than 5% of the contents words in the WordNet definitions.

For each sentence with target words to be disambiguated, a separate Bayesian network is constructed to represent the interconnections among the various senses that are possible. As an example, consider the task of disambiguating 'community' and 'town' in the following sentence:

> The community leaders expressed concern about the town's spiritual decline.

In WordNet, the sense distinctions for these words follow:

> community:
> 1. a group of people living in a particular local area
> 2. a group of people having ethnic or cultural or religious characteristics in common
> 3. common ownership
> 4. a group of nations having common interests
> 5. the body of people in a learned occupation
> 6. agreement as to goals
> 7. a district where people live; occupied primarily by private residences
> 8. (ecology) a group of interdependent organisms inhabiting the same region and interacting with each other
>
> town:
> 1. an urban area with a fixed boundary that is smaller than a city
> 2. an administrative division of a county
> 3. the people living in a municipality smaller than a city

When defining a CPT incorporating the hypernym relations from WordNet, the conditional probability $\mathrm{P}(\mathrm{hyponym} \mid \mathrm{hypernym})$ is inversely proportional to the number of children that the hypernym synset has; however, the cue validity weights are used as is. For instance, *municipality#1* has two children in WordNet, so the following is the CPT for *town#1*.

P(town#1 | municipality#1)

municipality#1	P(town#1)
F	ϵ
T	0.500

This illustrates that logical zeros are encoded using an epsilon rather than 0.0. This is a requirement for Bayesian network inferencing (Lauritzen and Spiegelhalter, 1988). It also leaves open the remote possibility for the false case being applicable. In the case with multiple parents, the Noisy-OR inspired model is used. The next CPT shows the probabilities that are derived for *municipality#1*, given its hypernyms *urban_area#1* and *administration_dist#1*, which have 7 and 16 hyponyms, respectively. Thus, the probabilities of the parents in isolation would be as follows:

P(municipality#1 | urban_area#1) = 0.143
P(municipality#1 | administration_dist#1) = 0.0625

The CPT for *municipality#1* combines these basically by summing the positive probabilities. So the above $P(c|p_{ij})$ terms can be seen in the entries having just one T value.

P(municipality#1 | urban_area#1, administration_dist#1)

urban_area#1	administration_district#1	P(municipality#1)
F	F	ϵ
F	T	0.143
T	F	0.062
T	T	0.205

Lastly, in the case of synsets without hypernyms (i.e., “starters” in WordNet), a uniform distribution is assigned to the node.

P(location#1)
0.5

Figure 5.7 shows a graph from a Bayesian network based on the lexical relations pertaining to these words. This includes embedded attribute nodes, such as *town-smaller-city* (used for the main relation inferred by the definition for *town#1*). In the graph, solid links indicate the strong semantic similarity relations implied by the WordNet *is-a* hierarchy. In contrast, dashed links show

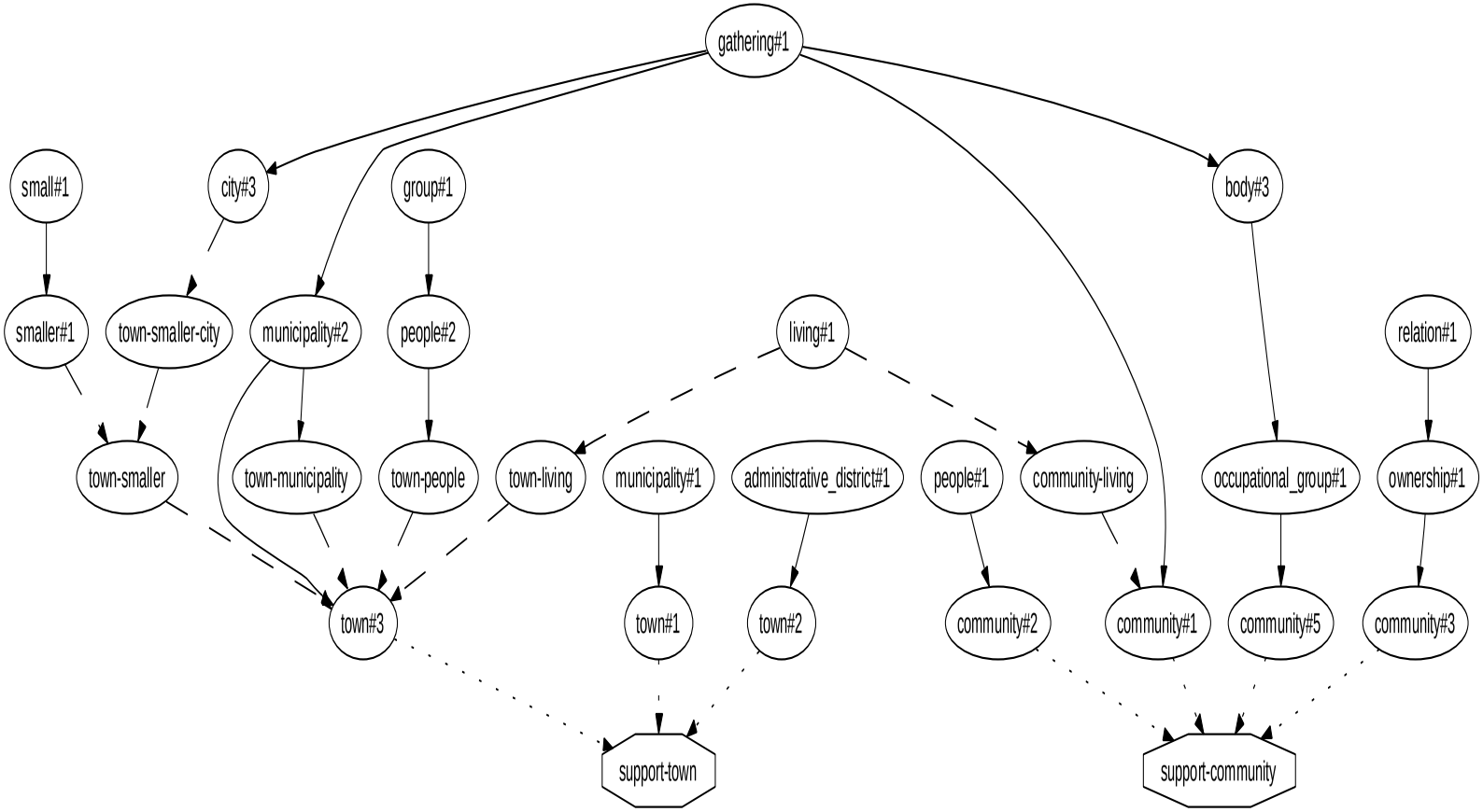


Figure 5.7: ***Relations for 'community' and 'town' in WordNet.*** Solid links indicate *is-a* relations, dashed links differentia, and dotted links context. Virtual evidence nodes are octagonal. Three senses of community omitted since untagged in DSO data.

the semantic-relatedness relations derived from dictionary differentia, and dotted links are used for evidence derived from context, as described below. All nodes with numeric suffixes represent senses (WordNet synsets), and the two octagonal nodes at the bottom encode the empirical support for each sense of the given words. These nodes are implemented as *virtual evidence nodes*, with the empirical distribution being encoded directly in the CPT's.

Virtual evidence nodes are binary-valued and do not have effect until clamped to a positive value. They influence the network indirectly through their incoming links. For example, assuming that the empirical distribution for 'town' is (.033, .263, .704), the following CPT would be created for its empirical support node (*support-town*):

P(support-town | town#1, town#2, town#3)

town#1	town#2	town#3	support-town
F	F	F	ϵ
F	F	T	.705
F	T	F	.264
F	T	T	ϵ
T	F	F	.034
T	F	T	ϵ
T	T	F	ϵ
T	T	T	ϵ

If the virtual evidence node *support-town* is not enabled, then the effect is as if the node were omitted from the network: the nodes for the senses of 'town' would be equally likely. However, when *support-town* is activated, it will cause the node for *town#3* to have a much higher likelihood of holding than the others.

5.2.2.3 System Results

To evaluate the Bayesian network word-sense disambiguation system, sentences having multiple, distinct sense annotations in the *Wall Street Journal* portion of the DSO corpus (Ng and Lee, 1996) were used. This is necessary since the system specifically addresses propagation of support among interdependent senses rather than just word-sense disambiguation in isolation, as with standard supervised WSD. Note that the Senseval II data used in the previous experiment (see Section 5.2.1.3) is not suitable for this purpose because the sense annotations are spread out through a much larger corpus. The top 100 sentences from the DSO corpus were chosen for the evaluation. Of these, six

sentences had eight distinct sense annotations, and there were just three sentences with fewer than four distinct annotations; the average was 6.2 different sense annotations per sentence.

Table 5.7 shows the results for the system over this data compared to a system based on (Wiebe et al., 1998b). In addition, a simple baseline of always selecting sense 1 for each word is included; this is a reasonable default choice because WordNet senses are ordered by frequency. As can be seen, using a Bayesian network with differentia-based relations generally leads to improvement over a system incorporating just the WordNet *is-a* relations. Overall, the differentia-based approach achieves a gain of nearly two percentage points (61.8% vs. 59.9%), a statistically significant difference. The baseline system that selects the most-frequent sense has accuracy of 55.2%. (The DSO data used for this experiment is entirely different from the Senseval II data used by the supervised classifier, so these results are not comparable to those in Table 5.4).

5.3 Summary

The distinguishing information in dictionary definitions can be exploited to improve word-sense disambiguation, as shown with two different approaches. Using the Senseval II data, a supervised classifier improves upon a typical approach to WSD through the use of relatedness collocations derived from differentia (63.8% accuracy versus 63.4%). A separate WSD system was developed that implements spreading activation over a Bayesian network based on WordNet augmented with differentiating relations. When evaluated over the DSO data, a significant improvement was achieved (61.8% versus 59.9%).

This chapter also discussed a qualitative evaluation by several human judges (with moderate agreement), showing that the relations extracted are generally acceptable. The next chapter discusses other possible application areas for this research (e.g., text segmentation). It also highlights the important differences of the techniques used here versus closely related work.

SentID	-Diff	+Diff	SentID	-Diff	+Diff	SentID	-Diff	+Diff
dj01-221	57.14	57.14	dj24-1329	66.67	66.67	dj39-1272	100.00	100.00
dj01-230	72.73	72.73	dj24-1805	57.14	71.43	dj39-1427	33.33	50.00
dj01-499	87.50	87.50	dj24-212	18.18	9.09	dj40-1033	83.33	83.33
dj01-559	100.00	100.00	dj25-195	42.86	42.86	dj40-133	42.86	57.14
dj02-100	50.00	50.00	dj25-2412	100.00	100.00	dj41-1856	57.14	28.57
dj03-1275	50.00	50.00	dj25-2466	42.86	42.86	dj41-2302	83.33	83.33
dj04-1900	87.50	87.50	dj26-1131	33.33	50.00	dj42-1378	90.00	90.00
dj05-1106	37.50	50.00	dj26-713	85.71	85.71	dj42-1380	50.00	33.33
dj05-1218	42.86	57.14	dj26-962	42.86	42.86	dj42-1382	55.56	66.67
dj05-1424	57.14	42.86	dj27-060	33.33	33.33	dj42-1638	71.43	57.14
dj05-244	62.50	75.00	dj27-722	83.33	83.33	dj42-496	50.00	50.00
dj06-358	87.50	87.50	dj28-1161	66.67	66.67	dj43-1747	66.67	50.00
dj07-272	50.00	50.00	dj28-1323	85.71	71.43	dj44-680	66.67	66.67
dj07-530	62.50	50.00	dj28-1780	33.33	33.33	dj45-809	100.00	100.00
dj11-624	87.50	87.50	dj28-325	66.67	83.33	dj46-173	100.00	100.00
dj11-627	71.43	71.43	dj29-1166	66.67	66.67	dj48-1129	66.67	83.33
dj11-771	69.23	76.92	dj29-699	33.33	33.33	dj48-1134	66.67	66.67
dj12-1069	42.86	57.14	dj30-128	83.33	83.33	dj48-1153	33.33	33.33
dj14-1675	100.00	100.00	dj30-1445	50.00	50.00	dj48-2082	37.50	62.50
dj14-1984	57.14	57.14	dj30-1980	100.00	100.00	dj49-1563	28.57	28.57
dj14-486	28.57	28.57	dj30-2154	25.00	50.00	dj49-578	83.33	83.33
dj14-573	57.14	57.14	dj32-249	40.00	40.00	dj51-1871	100.00	100.00
dj15-1422	57.14	42.86	dj33-1099	16.67	33.33	dj52-1462	50.00	50.00
dj15-1580	44.44	44.44	dj33-673	66.67	83.33	dj52-1657	100.00	100.00
dj15-1583	33.33	66.67	dj34-1671	83.33	83.33	dj53-1569	28.57	28.57
dj15-1622	28.57	57.14	dj34-1677	50.00	50.00	dj55-772	71.43	71.43
dj16-604	71.43	85.71	dj34-1737	66.67	50.00	dj56-625	50.00	50.00
dj18-619	14.29	14.29	dj34-561	33.33	33.33	dj57-1626	28.57	42.86
dj19-160	0.00	14.29	dj35-1092	83.33	100.00	dj57-1636	77.78	77.78
dj19-487	33.33	55.56	dj36-1781	33.33	50.00	dj57-1907	42.86	42.86
dj21-369	83.33	66.67	dj36-2313	50.00	50.00	dj59-046	16.67	16.67
dj22-1408	77.78	77.78	dj37-1389	66.67	33.33	dj60-566	71.43	71.43
dj22-2341	33.33	33.33	dj37-1538	83.33	83.33	Baseline	55.23	55.23
dj23-742	85.71	85.71	dj38-500	57.14	57.14	*Total*	*59.92*	*61.84*

Table 5.7: ***Bayesian network WSD classifier results.*** *SentID* gives the sentence ID from the DSO corpus. *-Diff* gives the accuracy when using just WordNet *hypernym* relations (*is-a*). *+Diff* gives the accuracy when using differentiating relations in addition to the *is-a* relations. *Baseline* always selects sense 1. *Total* gives the mean accuracy. The improvement of *+Diff* over *-Diff* is statistically significant at $p < .05$ via a paired t-test.

CHAPTER 6
DISCUSSION AND FUTURE WORK

This thesis work has touched on a variety of areas in computational linguistics. Section 6.1 highlights the important differences in the approaches taken compared to previous work. (See Chapter 2 for a discussion of other related work, as well as more details on the approaches mentioned below.) Section 6.2 presents ideas for how the work can be extended.

6.1 Comparison to Related Work

The background chapter presented a wide range of work that can be used for lexical acquisition. Here the main differences with closely related research are discussed, with an emphasis on the relation extraction and disambiguation processes.

6.1.1 Differentia Extraction

Most of the work addressing differentia extraction has relied upon manually constructed pattern matching rules (Vanderwende, 1995; Barrière, 1997; Rus, 2001), as discussed in Section 2.3.3. Here the emphasis is switched from transformation patterns for extracting relations to statistical classification for relation disambiguation, given tagged corpora with examples. Specifically, different classifiers are induced for each preposition using the feature organization shown in Figure 4.3. Different classifiers are also produced when targeting different role inventories; see Tables 4.11 and 4.16.

In Extended WordNet (Rus, 2001) relation disambiguation is not yet addressed: for instance, prepositions are converted directly into predicates in the underlying logical form representation (e.g., *by(e, x)*). In addition, the approach is closely tied into the grammar used by the parser, as a transformation rule is developed for each syntax rule. In a similar vein, Barnbrook's (2002) definition analysis system is tied into the specifics of a particular dictionary, namely *Collins Cobuild Student Dictionary*. Separating the surface-level relation extraction from the relation disambiguation helps to minimize such dependencies on the parser and on the definition format.

6.1.2 Relation Disambiguation

The work here addresses relation disambiguation specifically with respect to those indicated by prepositional phrases (i.e., preposition word-sense

disambiguation). Until recently, there has been little work on general-purpose preposition disambiguation. Litkowski (2002) and Srihari et al. (2001) present approaches using manually derived rules. Both approaches account only for a handful of prepositions, in contrast to the several dozen attempted here.

There have been a few machine-learning approaches that are more similar to the approach used in Chapter 4. Gildea and Jurafsky (2002) perform relation disambiguation using the FrameNet annotations as training data. However, they condition the classification on the predicating word (e.g., the verb corresponding to the frame under which the annotations are grouped). Therefore, the range of roles for a particular classification instance is more limited than in the experiments discussed here. Blaheta and Charniak (2000) use the Treebank annotations for relation disambiguation, addressing all adjuncts, not just prepositions. Therefore, they include the *nominal* and *adverbial* roles, which are syntactic and more predictable than the roles occurring with prepositional phrases.

6.1.3 Relation Weighting

Richardson (1997) illustrates how TF*IDF techniques can be used for relationship weighting, where documents are the set of definitions for the same dictionary entry word (see Section 2.4.2). The use of cue validity weights would produce comparable results if the set of contrasting concepts (see Figure 3.6) corresponded to those in the same entry word 'document.' This is because both measures are proportional to the co-occurrence frequency of an item and its reference class. However, the weighting will be different because the reference classes used here are determined based on semantic grounds (e.g., common most-informative ancestor) rather than morphological ones (e.g., same entry word).

6.1.4 Class-based Collocations

Scott and Matwin (1998) also use WordNet hypernyms for classification. Their approach is different in that they include a numeric density feature for any synset that subsumes words appearing in the document, potentially yielding hundreds of features. The hypernym collocations used here just involve a binary feature for each of the relations being classified, using indicative synsets based on the conditional probability test. In addition, adjective hypernyms are included rather than just nouns and verbs (see Section 4.3.1.1). Both approaches consider all senses of a word, distributing the alternative readings throughout the set of features. Gildea and Jurafsky (2002) instead just select the first sense for their hypernym features. They report marginal improvements

using the features, whereas the hypernym collocations resulted in significant improvement for the preposition disambiguation.

6.2 Areas for Future Work

The differentia extraction process was shown to provide useful information for word-sense disambiguation. This section sketches out a few other applications and mentions other areas for future work.

6.2.1 Extensions to Differentia Extraction Process

An obvious area for future work would be the application of the differentia extraction process to other types of dictionaries, both general-purpose dictionaries and specialized ones such as those used in medicine. Certain dictionaries, such as the *Longman Dictionary of Contemporary English* (LDOCE) and other learner's dictionaries, should be readily adaptable. Full-sized dictionaries would pose minor complications due to the wider range of formatting conventions, but this should mainly just affect the preprocessing stage.

The surface-level relations extracted by the system are dependent upon the parser used to analyze the definitions. Other parsers could be investigated, both dependency parsers as well as traditional phrase-structure parsers. For example, lexicalized grammars might be suitable, as they capture the dependencies among words via probabilities rather than parse rules (Jurafsky and Martin, 2000). Other direct extensions could be based on using different resources for the relation refinement. For example, the WordNet annotations discussed in Section 3.1.2 could be used as input into the relation disambiguation process discussed in Chapter 4, in addition to training over annotations from general text (e.g., in Treebank). Also helpful would be *The Preposition Project* (Litkowski, 2005), which is in the process of compiling detailed sense information for prepositions based on lexicographer analysis of definitions from machine readable dictionaries as well as annotations from FrameNet.

Several issues involved in the extraction process presented here have already been mentioned, such as structural ambiguity resolution and the assumption of uniformity in the definitions. One issue that has not been addressed is whether the information that is being extracted is indeed differential. This is not a focus of the current research, since it is assumed that dictionary definitions do not contain much extraneous information. Traditional dictionaries have always had considerable size constraints (Landau, 2001), so this is generally a safe assumption. Nonetheless, the use of *cue validities* could be used for this purpose. Section 3.3.2 showed how these are used for weighting the relations.

One way to evaluate how differential a particular relationship is would be to compare its cue-validity weight versus those for other relationships. Because this would have a bias towards incidental co-occurrences (e.g., 'greyhound' with 'motorcoach'), a separate measure could be used to quantify the usefulness of a relationship by seeing how frequently similar ones occur in the entire taxonomy as well as in a corpus. For instance, 'small' occurs 44 times in the WordNet definitions under the dog_{CANINE} branch, but only two times under the $hound_{\text{DOG}}$ sub-branch. In contrast, 'large' occurs 36 times under dog_{CANINE} but 10 times under $hound_{\text{DOG}}$. Thus, *small* is more differential than *large* in the context of hounds. Moreover, 'small' occurs frequently enough to be considered an important attribute for dogs, unlike 'bred by Pharaohs,' which only occurs once (for *Ibizan hound*).

Lastly, to facilitate structural disambiguation, class-based lexical associations could be investigated, for example, with the classes defined via WordNet. This line of research could use an extension of the approach taken by Hindle and Rooth (1993), discussed in Section 2.3.1.2. Instead of using word associations, the associations would be based on WordNet hypernym synsets. These associations would be different from those for the class-based collocations (see Section 4.3.1.1), because the latter involve word-sense distinctions rather than parse constituent placement decisions (e.g., prepositional phrase attachment). However, similar techniques could be used in the derivation of the associations.

6.2.2 Inferring Additional Semantic Role Markers

In Section 4.3.4.1, relational markers were inferred for the relationships in the Factotum semantic network. Recall that Factotum encodes the implicit relations among words in the Roget's Thesaurus, but does not indicate how the relations are manifested in English. The approach for inferring relational markers from the Factotum data checked for common prepositions occurring in proximity of the relational source and target terms. A similar approach can be taken for other types of relations, although it might be necessary to analyze the resulting text from the corpus checks to allow for a wider range of relation markers.

The Cyc Knowledge Base (KB) provides a large set of relations that could be used for this purpose, in particular providing a rich source of attribute relations. In the Cyc KB, properties for members of a category are specified via the *relationAllInstance* predicate. A few examples follow:

(relationAllInstance numberOfEdges Nonagon 9)
(relationAllInstance objectHasColor Slug BeigeColor)
(relationAllInstance hardnessOfObject StoneStuff Hard)

Note that the first *relationAllInstance* assertion is a shorthand notation (i.e., macro) for the following rule:

```
(implies
    (isa ?OBJ Nonagon)
    (numberOfEdges ?OBJ 9))
```

This is needed because Cyc distinguishes class-level concepts (i.e., *Collection*) from instance-level ones (i.e., *Individual*).

There are over 10,000 such class-level macro assertions in the KB, many of which deal with attribute specifications. For the first example, the relationship would be ⟨Nonagon, *numberOfEdges*, 9⟩. A proximity search of "nonagon NEAR 9" would produce hits similar to the following:

> AltaVista found 107 results About
> ...
> Math Forum - Ask Dr. Math Archives: College Geometry - Triangles/Polygons
> ... Equilateral triangle ABC has, near its center, point P, which is ... rectangle. *Nonagon* or Enneagon? [02/06/2003] Is 'enneagon' really the correct name for a *9-sided* polygon ...
> mathforum.com/library/drmath/sets/college_triangles.html
> More pages from mathforum.com
> ...
> Mr. Collins - 7th Grade Math - Math Vocabulary Sheet
> ... Trend Line—The line that can be drawn near the points on a scattergram ... Octagon–A polygon with 8 sides. 134. *Nonagon*–A polygon *with 9 sides*. 135
> ...
> members.aol.com/teacher677/mathvocabsheet.html
> More pages from members.aol.com ..

After analyzing the text of such hits, the patterns "9 sided" and "with 9 sides" would likely emerge. Analyzing the results over the entire set of Cyc's *numberOfEdges* assertions could produce the following patterns:

> with ⟨target⟩ sides
> has ⟨target⟩ sides
> ⟨target⟩-sided

This technique can also be extended to finding relation markers in foreign languages, such as Spanish, given a bilingual dictionary. Specifically, the

proximity searches can be done over the translation equivalents of the source and target terms. Ambiguous translations pose a complication, but in most of these cases, similar relation markers should result unless the alternatives for a term's translation have significantly different meanings. As an illustration, when the process is applied to the translated relationship for the example in Section 4.3.4.1, namely ⟨secar, *is-function-of*, secarador⟩, the top three markers are 'con,' 'de,' and 'para.'

6.2.3 Application to Text Segmentation

In addition to word-sense disambiguation, the conceptual differentia-extraction work can be applied to text segmentation. Since related words tend to occur together (Morris and Hirst, 1991), the co-occurrence frequency of related words can serve as an indication of text cohesion and thus can be used to estimate segment boundaries. Hearst's (1994) *TextTiling* program is representative of the general approach currently taken towards text segmentation. Hearst relies solely on word frequency, so the main issue is how to incorporate the data on word relatedness into this framework.

In TextTiling, the similarity computation is as follows (Hearst, 1993):

$$\mathrm{sim}(b_1, b_2) = \frac{\sum_t W(t, b_1) W(t, b_2)}{\sqrt{\sum_t W(t, b_1)^2 \sum_t W(t, b_2)^2}}$$

where $W(t, b)$ is weight of term t in block b, which is given by its frequency. In effect, the blocks are described by vectors with frequencies for each word:

$$V_i = \langle f_i(w_1) f_i(w_2) ... f_i(w_N) \rangle$$

where $f_i(w)$ is the frequency of word w in block i. Similarity is given by the cosine of the angle between the vectors.

A direct extension of this approach to one using conceptual differentia would be to assign semantic-relatedness classes to words based on occurrence in differentia-based relations. In effect, this augments the vectors computed by the original algorithm by a component consisting of the frequency counts for class labels:

$$V_i = \langle f_i(w_1) ... f_i(w_N), f_i(c_1) ... f_i(c_M) \rangle$$

6.2.4 Mapping Senses from other Dictionaries into WordNet

Given that different dictionaries emphasize different aspects of word meaning, it is desirable to combine the information acquired. This would likely

require human assistance, as determining which aspects of the different meanings apply to the same sense might involve subtle decision making. This thesis work could be integrated into an interactive system in which the relational analyses of the meanings for the same word in different dictionaries are presented to the user. The user then can combine the appropriate lexical relations that have been extracted into a single entry for incorporation into the lexicon. A similar system could be used to map senses from other dictionaries into WordNet.

Additional processes not discussed here would be required to facilitate the sense mapping. For example, due to the use of different sense inventories in different dictionaries, it is desirable for the computer to help with the determination of which sense definitions involve the same underlying concept. Although simple word-overlap schemes could help in this respect (Nastase and Szpakowicz, 2001), it would be beneficial to integrate work on aligning ontologies (O'Hara et al., 1998; Hovy, 1998) to account for WordNet's hierarchy.

6.2.5 Analyzing Lexical Gaps

Bilingual dictionaries are an important resource for machine translation. Usually, the entries consist merely of target language words with the same meaning as the source language word (i.e., *translation equivalents*). For instance,[1]

quintería	f. farm, grange.
perdido, -da	adj. lost. 2 mislaid. ...

However, when there is no word or commonly used phrase in the target language, the situation represents a *lexical gap*. In such cases, the entries in bilingual dictionaries give brief definitions more akin to monolingual dictionaries, as seen in the following example.

alhóndiga	public granary or grain market
traspapelarse	ref. to be mislaid among other papers

When definitions are used in place of translation equivalents, the differentia extraction system could be applied to the definition text to determine the conceptual relations and attributes that apply to the underlying concept being defined.

[1]These examples are taken from the Spanish-English Dictionary provided with the *NTC Languages of the World* CD-ROM from the National Textbook Company.

This would be beneficial in an interactive lexicon acquisition system that helps users create lexicon entries by either copying entries from an existing lexicon or creating one from scratch. For example, one way to bootstrap a foreign language lexicon is to apply transformations to an English lexicon based on a bilingual dictionary. In terms of a Mikrokosmos lexicon entry (Onyshkevych and Nirenburg, 1992), this would just modify the SYN-STRUC part of the frame structure while preserving the SEM-STRUC (see Section 2.1.3.3). The user would verify the suggested foreign language lexicon entry, making corrections if necessary. However, in the case of lexical gaps (e.g., no English lexicon entry to transform), the differentia extraction system could be used as a fallback mechanism to infer relations for the SEM-STRUC.

6.3 Summary

This chapter briefly highlighted the differences in this research versus closely related work (e.g., use of semantic role annotations for relation disambiguation instead of rules). It also mentioned some of the more promising areas for future work, such as text segmentation using word relatedness class derived from differentia. The next chapter includes a summary of the entire thesis, as well as speculations inspired by the work.

CHAPTER 7
CONCLUSION

This thesis has advocated an empirical methodology for extracting differentiating relations (i.e., *differentia extraction*), demonstrating a viable approach to exploiting information in text-based resources without involving the expense of manual knowledge extraction rules. This research has also touched upon a variety of other areas as illustrated in the previous chapter (e.g., class-based collocations for word-sense disambiguation).

This concluding chapter first summarizes the thesis to review the important points that were discussed (Section 7.1). In addition, I offer several observations resulting from the research. These encompass the main contributions of the research (Section 7.2) and include other insights based on the work (Section 7.3).

7.1 Summary of Thesis

This thesis improves upon previous work on extracting information from dictionary definitions by the use of data-driven relation disambiguation. This research exploits the Treebank and FrameNet semantic roles annotations mapped into a reduced inventory suitable for representing distinctions present in definitions. All the definitions from WordNet 1.7.1 were analyzed using this process. A random sample of the results was evaluated by six human judges, indicating that the quality is generally acceptable (e.g., compared to manually corrected output). In addition, the extracted information was shown to improve two separate approaches to word-sense disambiguation. Detailed summaries of the chapters follow.

7.1.1 Importance of Differentiating Relationships

Chapter 1 provided motivation for the research, in particular the need for including differentiating relations in semantic lexicons. For instance, differentiation is an integral part of categorization as revealed by research in cognitive psychology. Additional support for this hypothesis is based on the prevalence of differentiating relations in manually constructed lexicons versus those predominantly acquired using automated means. The introductory chapter also illustrated that dictionary definitions are still the best resource for extracting these relations, because corpus analysis over free text is not likely to be sufficiently directed at acquiring differentiating relations.

7.1.2 Approaches for Lexical Acquisition

Chapter 2 presented background material on lexical semantics and illustrated the common techniques used for acquiring semantic knowledge. Several different representations based on semantic networks were presented, in particular the early influential work by Schank (1973) on conceptual dependencies and by Wilks (1975b) on preference semantics. The ontological semantics approach is currently the state-of-the-art for lexicons that provide detailed semantics (Nirenburg and Raskin, 2004). Representing fine-grain distinctions would require more emphasis on stylistics and other pragmatic considerations (Edmonds and Hirst, 2002).

Manual acquisition is preferred when quality of lexicon entries is critical (Onyshkevych and Nirenburg, 1995; Burns and Davis, 1999). A variety of automatic approaches to lexical acquisition was discussed. Corpus approaches to lexical acquisition often involve the use of lexical associations, such as in words clustered according to similarity (Lin, 1998) or in preferences for verbal arguments (Resnik, 1995). Translation lexicons illustrate cross-lingual lexical associations (Melamed, 2000). The initial definition analysis attempts concentrated on extracting the main semantic category for the word being defined (*genus* extraction), as seen in the influential work by Amsler (1980). Later work addressed extracting the differentiating relationships from definitions (i.e., *differentia*). Such analysis can acquire precise relationships; however, it has relied primarily upon manually derived extraction rules (Barrière, 1997; Vanderwende, 1996; Rus, 2002).

7.1.3 Extraction of Differentiating Relations

Chapter 3 illustrated the steps in automatically extracting surface-level relations from dictionary definitions. Statistics on WordNet (Miller, 1990) version 1.7.1 were first presented, showing that it is equivalent in scope to a learner's dictionary such as the *Longman Dictionary of Contemporary English* (LDOCE), a popular dictionary for computational linguistics research (Procter, 1978). The WordNet definitions are not as uniform as those in commercial dictionaries (e.g. LDOCE), but they still tend to follow the classical *genus/differentia* format (Landau, 2001). The first extraction step involves preprocessing the definitions, such as in forming a complete sentence with the word being defined. The definition is then parsed using the Link Grammar (Sleator and Temperley, 1993), a dependency parser that produces a list of highly specialized grammatical relations among the words in the form of tuples.

The specialized dependency relations resulting from the definition parse are converted into higher-level grammatical relations, using a simple mapping

into relations like *modifier-of*. In addition, pairs of tuples involving the same function word in the target and source term positions are combined into a single tuple with the function word serving as the relation. The last extraction step involves the weighting of the grammatical relationships using the notion of cue validities (Smith and Medin, 1981). The relation type and target term are treated together as a feature of the source term, and the weighting ensures that features more specific to a given source term are weighted higher.

7.1.4 Disambiguation into Conceptual Relations

Chapter 4 presented the crucial disambiguation process that transforms the syntactically oriented relationships into conceptual ones. For the relation source and target terms, this amounts to word-sense disambiguation (WSD). Since WordNet definitions are being targeted here, the WSD annotations provided in Extended WordNet (Novischi, 2002) are incorporated. Information on two separate types of semantic role resources is provided. The emphasis is on corpus-based resources providing annotations of naturally occurring text as done with Treebank (Marcus et al., 1994) and FrameNet (Fillmore et al., 2001). In addition, semantic role inventories from knowledge bases are illustrated, in particular for Cyc (Lehmann, 1996) and Factotum (Cassidy, 2000).

The disambiguation concentrates on relations indicated by prepositional phrases, and is framed as word-sense disambiguation for the preposition in question. A new type of feature for word-sense disambiguation is introduced, using WordNet hypernyms as collocations rather than just words as typically done. For relationships derived from knowledge bases, the prepositions and other relational markers need to be inferred from corpora. A method for doing this is demonstrated using Factotum. In addition, to account for granularity differences in the semantic role inventories, the relations are mapped into a common inventory that was developed based on the inventories discussed in the chapter. This allows for improved classification in cases where inventories provide overly specialized relations (e.g., FrameNet).

7.1.5 Lexicon Augmentation and WSD Applications

Chapter 5 discusses two aspects on how the work can be applied, namely lexicon augmentation and word-sense disambiguation; in each case, detailed evaluations are presented (one direct and the other indirect). The output provided by the analysis can be used to augment existing lexicons, in particular WordNet. To directly evaluate the quality of the information that would be added, a random sample was selected and evaluated by four human judges familiar with computational linguistics and two others with computer science

backgrounds. To provide a baseline for the accuracy, part of the output was manually corrected prior to the evaluation. Inter-coder reliability analysis was performed using the Kappa statistic (Carletta, 1996). The evaluation illustrated that the quality of the uncorrected relationships is acceptable, based on comparisons of scores assigned to that of the manually corrected relations.

An indirect evaluation of the extracted information is illustrated with respect to word sense disambiguation. For a supervised WSD approach, a new-type of collocation feature is introduced that uses the differentia to expand the set of potential collocations. Standard collocations are derived using co-occurrence counts for the context words and the tagged word senses. For the differentia-based collocations, the relatedness weight is used in place of unit weights assigned to each co-occurrence. When tested over the Senseval II data (Edmonds and Kilgarriff, 2002), these features consistently yield improvements compared to just using word collocations. For a probabilistic spreading activation approach, the differentia properties are used to enhance the connectivity in a Bayesian network representing the interdependencies among the word senses for the target words being disambiguated. This builds upon the work of Wiebe et al. (1998b), where the connectivity is based solely on the WordNet *hypernym* relations (i.e., *is-a* hierarchy). When tested over the DSO data (Ng and Lee, 1996), this leads to statistically significant improvements.

7.1.6 Looking Backward and Then Forward

Chapter 6 first compared the work in this thesis to previous research. Since the background chapter covered earlier differentia-extraction work, this chapter concentrated on relation disambiguation, such as the recent work over Treebank (Blaheta and Charniak, 2000) and FrameNet (Gildea and Jurafsky, 2002). Other topics discussed include relation weighting (Richardson, 1997) and class-based collocations (Scott and Matwin, 1998). Chapter 6 also sketched areas for future work. One future application could use differentia-derived relatedness classes to augment existing approaches for text segmentation. The relation disambiguation methodology could also be extended to handle modification-type relations. For instance, information can be inferred from Cyc using similar techniques as those developed for Factotum.

7.2 Significance of Research

7.2.1 Empirical Acquisition of Conceptual Distinctions

Contrary to what some of the criticisms of machine-readable dictionary (MRD) research might imply (Amsler, 1995; Ide and Véronis, 1993), this type

of analysis can still be quite fruitful. This thesis provides an empirical methodology for extracting information from MRD's that is not tied into the specifics of the defining language used in the dictionary being analyzed. For instance, there are no pattern-matching rules tailored to the way the *instrument* relation is commonly indicated in definition text. In contrast, occurrences of this relation type are inferred using lexical associations derived from preposition usage in general text. Of course, such flexibility comes at a cost, which in this case is the requirement for tagged corpora indicating how the relations are expressed in natural language. By utilizing existing annotations prepared for general text (e.g., Treebank and FrameNet), such costs can be minimized.

7.2.2 Exploiting Resources on Relation Usage

This thesis demonstrated effective means of exploiting resources providing information on relation usage. In the case of corpus-based resources, annotations covering prepositional phrase usage were treated as sense annotations for the corresponding prepositions. In addition, in the case of FrameNet, the fine-grained relations were converted into a common relation inventory (see Table 4.19). Knowledge bases generally do not provide information on how specific relationships are indicated in natural language. Therefore, a way to infer relational markers was developed and illustrated using Factotum. Such techniques can be used to extract information from Cyc.

7.2.3 Bayesian Networks for Differentia Representation

Although popular in artificial intelligence, Bayesian networks are not commonly used in computational linguistics. The probabilistic spreading activation approach of Wiebe et al. (1998b) illustrates an effective use of Bayesian networks in modeling explicit relations in WordNet. This is extended here to include the implicit relations from the WordNet definitions. The links are defined based on the cue validity weights determined for the implicit relations: thus, causality is interpreted in terms of salience. The inclusion of relations based on differentia can lead to large networks. To avoid problems with overly large cliques in the underlying representation used for direct evaluation (Lauritzen and Spiegelhalter, 1988), embedded attributes are treated as separate nodes. Using Bayesian networks modeling differentia leads to significant improvements in a system for word-sense disambiguation.

7.2.4 Class-based Collocations for Sense Disambiguation

Supervised systems for word-sense disambiguation (WSD) often rely upon word-based keywords or collocations to provide clues on the most likely sense for a word given the context. In the second Senseval competition, these features figured predominantly among the feature sets for the leading systems (Mihalcea, 2002; Yarowsky et al., 2001). A limitation of such features is that the words selected must occur in the test data in order for the features to apply. To alleviate this problem, class-based approaches replace word-level features with category-level ones (Ide and Véronis, 1998). When applied to collocational features, this approach effectively uses class labels rather than wordforms in deriving the collocational features.

Two separate types of class-based collocation features for WSD were developed as part of the work for this thesis. The hypernym collocations were designed initially for preposition disambiguation, but they have been found useful for WSD in general (O'Hara et al., 2004). To derive the collocations, the input text is transformed by replacing each wordform with tokens representing each of the hypernyms in WordNet. This introduces noise due to ambiguity, but the sense-specific conditional probability tests used for collocation selection will compensate. Differentia-based collocations are derived in a similar process, except that co-occurrences are weighted based on relatedness rather than assuming unit weights.

7.3 Speculations

In concluding this thesis, some speculations are offered. These include aspects of the research not yet formally evaluated. It also covers the future direction of natural language processing (NLP), in particular, with respect to computational semantics.

7.3.1 Adaptability of Thesis Work

In Section 2.3.3, some criticisms of the work in machine-readable dictionaries (MRD) analysis were mentioned, casting doubts on the optimistic viewpoint that dictionaries can provide adequate information to form the basis for knowledge bases. Nevertheless, work in extracting information from MRD's can be beneficial to the extent that the techniques apply to knowledge acquisition from other resources such as technical dictionaries or encyclopedias. For example, Microsoft's MindNet (Richardson et al., 1998) was originally developed just using definition analysis, but it was enhanced via analysis of encyclopedia articles with minor change to the underlying extraction processes.

By isolating the disambiguation from the extraction step, this approach can be readily adapted to related tasks, such as in extracting semantic information from encyclopedia texts. Moreover, this approach can be more directly extended to handling foreign languages. The main requirements would be the availability of a parser for the target language as well as a tagged corpus for relation usage. Fortunately, these resources might already have been developed for other purposes. For instance, there are currently three projects on the creation of FrameNet-style lexicons for languages other than English (specifically German, Spanish, and Japanese).[1]

7.3.2 Computational Semantics in General

In many respects, the area of computational semantics is less ambitious than it was thirty years ago, when Schank and Wilks were developing rule-based systems for natural language understanding (Schank, 1973; Wilks, 1975a). Deep understanding is currently not attempted except for specialized domains. In addition, much effort is now being directed at aspects of NLP that were once taken for granted, such as parsing and word-sense disambiguation (e.g., using statistical approaches for better coverage). Such a pattern now seems inevitable for progress in natural language processing as with artificial intelligence in general: knowledge-based or heuristic approaches define new frontiers and then empirical approaches are used later on to provide improvements.

This might give the impression that knowledge-based approaches should be avoided in general. On the contrary, they often are necessary for clarifying techniques effective for certain problem areas before empirical approaches can be attempted to provide better coverage. The work in the last decade on named-entity recognition illustrates this pattern (Cowie and Lehnert, 1996; Srihari et al., 2001); and, the recent advances in question answering follow similar patterns (Moldovan and Rus, 2001; Ravichandran and Hovy, 2002).

When deep semantics becomes back in vogue, the ideas presented in this thesis can be used for important subtasks that will be required. For instance, as noun definitions are often indefinite descriptions, it is likely that analysis of the latter will be amenable to similar approaches. This research would also be helpful in the analysis of definite descriptions, although additional mechanisms would be required to account for anaphora and co-reference resolution. Additional knowledge-based work is likely to be required before such corpus-

[1] See www.icsi.berkeley.edu/ framenet/FNabroad.html.

based approaches are feasible in general (Vieira and Poesio, 2000; McShane and Nirenburg, 2002).

In short, although the techniques described here have only been applied to dictionary definitions, the research actually is a step towards deep understanding of general English text.

APPENDICES

APPENDIX A
PRIMER ON MACHINE LEARNING

This appendix presents a primer for the two machine learning techniques used in this thesis. Both are supervised approaches relying upon training data to provide examples along with the correct classification for each. There are several useful texts introducing machine learning techniques. Witten and Frank (1999) provide a practical introduction along with a discussion of their Java implementation. Mitchell (1997) provides a more-theoretical introduction but still is somewhat accessible. Several texts concentrate on the application of these techniques to natural language processing (Charniak, 1993; Manning and Schütze, 1999).

Bayesian Classification

The first technique covered uses probabilities for the various combinations of feature (attributes) and classification values (classes) in a purely statistical decision procedure. *Bayesian classification* derives its name from the use of Bayes Rule from probability theory, which provides a way to express the conditional probability $P(x|y)$ in terms of the inverse conditional probability $P(y|x)$. Bayes Rule follows:

$$P(x|y) = \frac{P(x \wedge y)}{P(y)} = \frac{P(y|x)P(x)}{P(y)}$$

In particular, the probability of a particular class given the features $P(c_i|f_1 ... f_n)$ is expressed in terms of the probability of the features given the class value $P(f_1 ... f_n|c_i)$. Specifically,

$$P(c_i|f_1 ... f_n) = \frac{P(c_i \wedge f_1 ... f_n)}{P(f_1 ... f_n)} = \frac{P(f_1 ... f_n|c_i)P(c_i)}{P(f_1 ... f_n)}.$$

Figure A.1 shows the basic steps in using Bayesian classification for machine learning. Classifiers using this technique are often referred to as *Naive Bayes*. Note that the simplification via the conditional independence assumption in step 4 is omitted in general Bayesian classification. The normalization constant (z) is actually determined after the probabilities are determined as the inverse of the total sum. Advanced approaches also use more sophisticated

Input: Instance I described in terms of *features* F_i.

Goal: Determine the class $c_j \in C$ that best describes the input.

Method:

1. Collect large sample of known classifications:
   $\langle \{f_1, ..., f_n\}, c_i \rangle$
2. Estimate probability of each class value
   $P(C = c_j) \simeq \frac{f(c_j)}{\sum_i f(c_i)}$
3. Estimate probability of each feature given each class value:
   $P(F_i = f_i | C = c_j) \simeq \frac{f(f_i, c_j)}{f(c_j)}$
4. Choose class value that maximizes posterior probability:

$$
\begin{aligned}
& P(C_j = c_j | F_1 = f_1, ..., F_n = f_n) \\
& \quad = \frac{P(f_1, ..., f_n | c_j) P(c_j)}{P(f_1, ..., f_n)} && \text{Bayes Rule} \\
& \quad \simeq \frac{\prod_{i=1}^{n} P(f_i | c_j) P(c_j)}{P(f_1, ..., f_n)} && \text{Conditional Independence Assumption} \\
& \quad = \prod_{i=1}^{n} P(f_i | c_j) P(c_j) z && \text{Normalization Constant}
\end{aligned}
$$

Figure A.1: ***Simple Bayesian classification ("Naive Bayes").*** Based on (Weiss and Kulikowski, 1991; Mitchell, 1997).

	F_1	F_2	F_3	C
Type	Flavor	Fat	Carbos	OK?
Anchovies	spicy	high	low	No
Bananas	bland	none	medium	Yes
Burritos	hot	moderate	high	Yes
French fries	mild	high	high	No
Hamburgers	mild	moderate	low	Yes
Hotdogs	bland	high	low	No
Jalapeños	hot	none	low	No
Liver	bland	moderate	low	No
Meatloaf	mild	moderate	low	No
Pizza	spicy	high	high	Yes
Sushi	spicy	low	low	Yes
Tacos	spicy	moderate	medium	No
Zucchini	mild	low	low	Yes
Enchiladas	hot	moderate	high	???

Table A.1: ***Data for favorite-foods example.*** The top part shows the training data, and the bottom indicates the test data.

probability approximation techniques than used in step 2 to account for problems with spare data.

As a simple example, consider the task of classifying food preferences based only on the characteristics of flavor, fat content, and complex carbohydrate content. The feature descriptions follow, and sample data is shown in Table A.1.

F_1=Flavor: $f_1 \in \{\text{bland}, \text{mild}, \text{spicy}, \text{hot}\}$

F_2=Fat: $f_2 \in \{\text{none}, \text{low}, \text{moderate}, \text{high}\}$

F_3=Carbos: $f_3 \in \{\text{low}, \text{medium}, \text{high}\}$

To determine the acceptability of a new type of food not shown in the table, such as enchiladas, the following steps are done, assuming the input instance to be classified is $\{$Flavor=hot, Fat=moderate, Carbos=high$\}$.

1. Obtain training data on food preferences: $\langle\{F_1, ..., F_n\}, C_i\rangle$

   See Table A.1.

2. Estimate probability of each class value

$$P(C = c_j) \simeq \frac{f(c_j)}{\sum_i f(c_i)}$$

$$P(C = \text{no}) \simeq \frac{7}{13}$$

$$P(C = \text{yes}) \simeq \frac{6}{13}$$

3. Estimate $P(F_i = f_i|c_j)$ as $\frac{f(f_i, c_j)}{f(c_j)}$

$$P(\text{Flavor} = \text{hot}|C = \text{Yes}) \simeq \frac{f(\text{hot}, \text{Yes})}{f(\text{Yes})} = \frac{1}{6}$$

$$P(\text{Flavor} = \text{hot}|C = \text{No}) \simeq \frac{f(\text{hot}, \text{No})}{f(\text{No})} = \frac{1}{7}$$

...

$$P(\text{Carbos} = \text{high}|C = \text{Yes}) \simeq \frac{f(\text{high}, \text{Yes})}{f(\text{Yes})} = \frac{2}{6}$$

$$P(\text{Carbos} = \text{high}|C = \text{No}) \simeq \frac{f(\text{high}, \text{No})}{f(\text{No})} = \frac{1}{7}$$

4. Find c_j maximizing $P(C = c_j|F_1 = f_1, ..., F_n = f_n)$

$$\begin{aligned} &P(\text{no}|\text{hot}, \text{moderate}, \text{high}) \\ &\quad \simeq \frac{P(\text{hot}|\text{no})P(\text{moderate}|\text{no})P(\text{high}|\text{no})P(\text{no})}{P(\text{hot}, \text{moderate}, \text{high})} \\ &\quad \simeq P(\text{hot}|\text{no})P(\text{moderate}|\text{no})P(\text{high}|\text{no})P(\text{no})z \\ &\quad = (1/7 \times 3/7 \times 1/7 \times 7/13)z = 0.0047z = 0.356 \end{aligned}$$

$$\begin{aligned} &P(\text{yes}|\text{hot}, \text{moderate}, \text{high}) \\ &\quad \simeq \frac{P(\text{hot}|\text{yes})P(\text{moderate}|\text{yes})P(\text{high}|\text{yes})P(\text{yes})}{P(\text{hot}, \text{moderate}, \text{high})} \\ &\quad \simeq P(\text{hot}|\text{yes})P(\text{moderate}|\text{yes})P(\text{high}|\text{yes})P(\text{yes})z \\ &\quad = (1/6 \times 2/6 \times 2/6 \times 6/13)z = 0.0085z = 0.644 \end{aligned}$$

Thus, enchiladas would be classified as acceptable. The probability of acceptance (i.e., $P(\text{yes}|...)$) is nearly twice that of rejection, even though the latter is more common overall (i.e., $P(\text{no}) > P(\text{yes})$).

Decision Trees

Decision trees involve a more heuristic decision procedure than Bayesian classification. The idea is to apply a series of attribute value tests to partition

the data into subsets that are more predictable than the original data. Then the majority class for a subset is chosen as the classification matching the attribute tests. Figure A.2 shows a decision tree for the favorite-foods example. It first checks the fat content of the food. Low fat foods are accepted immediately. Most of the remaining cases are then decided by just checking flavor. However, in two cases the carbohydrate content needs to be checked as well. For example, spicy foods high in fat are only accepted if also high in carbohydrates.

```
if (Fat = low) then Yes
if (Fat = none) then
    if (Flavor = mild) then Yes
    if (Flavor = spicy) then null
    if (Flavor = hot) then No
    if (Flavor = bland) then Yes
if (Fat = moderate) then
    if (Flavor = mild) then
        if (Carbos = low) then No
        if (Carbos = medium) then No
        if (Carbos = high) then null
    if (Flavor = spicy) then null
    if (Flavor = hot) then Yes
    if (Flavor = bland) then No
if (Fat = high) then
    if (Flavor = mild) then No
    if (Flavor = spicy) then
        if (Carbos = low) then No
        if (Carbos = medium) then null
        if (Carbos = high) then Yes
    if (Flavor = hot) then null
    if (Flavor = bland) then No
```

Figure A.2: ***Decision tree for favorite-foods example.*** Combinations of features not covered in the training data yield *null* classification.

Decision trees are induced in a process that recursively splits the training examples based on the feature that partitions the current set of examples to maximize *information gain* (Mitchell, 1997; Witten and Frank, 1999). This is commonly done by selecting the feature that minimizes the *entropy* of the distribution (i.e., yields least uniform distribution). Entropy is a measure of the uniformity of a distribution of values. Higher entropy values signify higher uniformity (or randomness). Entropy can be viewed as the weighted average of the

information content associated with each probability of a distribution (Manning and Schütze, 1999):

$$\text{Entropy} = \sum_{i} -p(x_i)\log_2(p(x_i))$$

With N classes, the entropy ranges from 0 to $\log_2(N)$; so, for a binary distinction, as in the favorite-foods example, the entropy is in the range from 0 to 1.

As an illustration, consider the steps in using entropy to determine the first attribute to split. For the entire dataset, the entropy of the class distribution with 6 Yes's and 7 No's is as follows:

$$\begin{aligned}\text{Entropy} &= \sum_{i} -p(x_i)\log_2(p(x_i)) \\ &= -P(\text{yes})\log_2(P(\text{yes})) - P(\text{no})\log_2(P(\text{no})) \\ &= -6/13 \times \log_2(P(6/13)) - 7/13 \times \log_2(P(7/13)) = .996\end{aligned}$$

If the *Flavor* attribute is chosen first to split the data, then the resulting partitions would be as follows, yielding a small decrease in entropy (i.e., to .951 on average).[1]

Flavor	Classification distribution	Entropy
bland	{ No, No, Yes }	0.918
hot	{ Yes, No }	1.000
mild	{ No, No, Yes, No, Yes}	0.971
spicy	{ Yes, No, Yes }	0.918

If attribute *Fat* is used instead, the decrease in entropy is better (i.e., to .835).

Fat	Classification distribution	Entropy
high	{ No, Yes, No, No }	0.811
low	{ Yes }	0.000
moderate	{ Yes, No, No, Yes, No }	0.971
none	{ No, Yes, Yes }	0.918

Lastly, if the *Carbos* attribute is used first, the split would be slightly higher than the first (i.e., to .952).

[1]The overall entropy for the partition split is based on a weighted average of the splits (i.e., $.951 = 3/13 * .918 + 2/13 * 1.0 + 5/13 * .917 + 3/13 * .918$).

Carbos	Classification distribution	Entropy
high	{ No, Yes, Yes }	0.918
low	{ No, No, No, No, No, Yes, Yes, Yes }	0.954
medium	{ No, Yes }	1.000

Therefore, using *Fat* to partition the data first yields the lowest average entropy and thus leads to the highest information gain. This process is then repeated for the remaining attributes in turn on each of the partitions. For the example in Figure A.2, this involves using *Flavor* and then *Carbos*.

Quinlan (1986; 1993) has developed a series of decision trees that perform very well for a variety of tasks. *ID3* is the simplest and just uses information gain for splitting the nodes in the tree, as described above. *C4.5* is an extension that uses statistical tests to address the problem with overfitting of data that can occur with decision trees. For example, when deciding whether to split a node, it checks whether the difference in the information content can be attributed to chance. If so, then the majority-test decision is applied instead of further attribute checks.

APPENDIX B
PRIMER ON BAYESIAN NETWORKS

Bayesian networks provide a convenient way to represent probabilistic relations in a graphical format. They are suitable for problems where the probabilistic dependencies are such that only a small number of variables have direct influence over a given variable. For example, when determining whether you can afford a particular purchase, you only need to consider whether you have enough money at your disposal, not the various ways by which to obtain more money.

As a simple example, consider the problem of whether you should order an espresso drink (e.g., Cafe Latte) or just plain coffee. Espresso drinks generally taste better and are much stronger; however, they usually cost twice as much as regular coffee. Figure B.1 depicts this situation, from the perspective of a student (e.g., low funds at end of semester).

An assumption underlying Bayesian networks is that nodes are only dependent upon those directly connected to it in the graph. For example, the *Buy Espresso* is not directly dependent upon *Just Paid*. Therefore, the conditional probability table (CPT) for *Buy Espresso* does not include *Just Paid*. Instead, it only accounts for the parent nodes *Real Tired* and *Little Money*. As with other statistical representations, the inference implicitly involves calculating the joint probability for all the variables represented by the nodes. Without the independence assumption, the calculation would involve many combinations of the variables. Via the chain rule, the joint probability can be determined as follows: (omitting the *GO* node for simplicity):

$$\begin{aligned}&P(BE, RT, LM, ES, JP) = \\ &\quad P(JP) \times P(ES|JP) \times P(LM|ES, JP) \times P(RT|ES, JP, LM) \times P(BE|ES, JP, LM, RT)\end{aligned}$$

However, by accounting for the independent assumptions represented in the graph, this formula can be simplified to the following:

$$\begin{aligned}&P(BE, RT, LM, ES, JP) = \\ &\quad P(JP) \times P(ES) \times P(LM|ES, JP) \times P(RT) \times P(BE|LM, RT)\end{aligned}$$

For the most part, evaluation of the network involves working top-down through the network propagating the prior probabilities for nodes without parents to those nodes directly connected to them and then recursively propagating the resulting posterior probabilities. For a particular node with parents, the

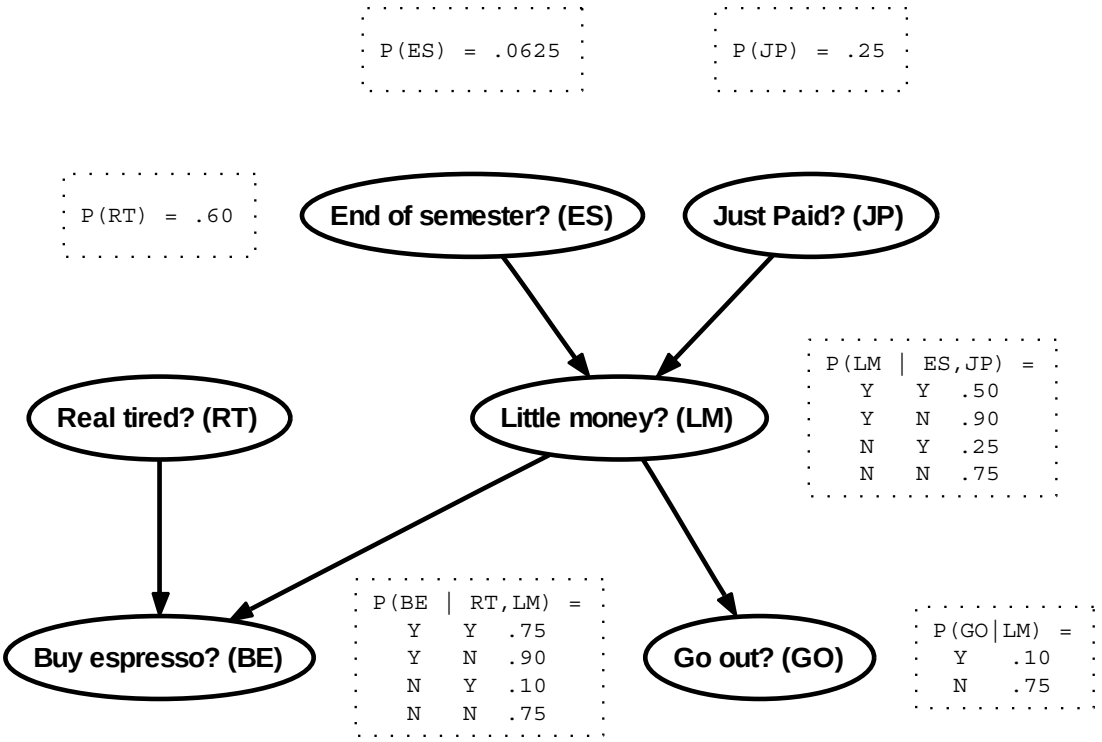


Figure B.1: ***Bayesian network for choosing espresso over regular coffee.*** Oval nodes represent random variables, and boxes indicate their probability distributions. Only the positive case probabilities are shown.

value is based on weighting each entry of its CPT by the probability that the particular combination of parent values occurs. For example, the value for the embedded node *LM* would be based on weighting the four possibilities for *ES* and *JP*, namely $(\text{False}, \text{False})$, $(\text{False}, \text{True})$, $(\text{True}, \text{False})$, and $(\text{True}, \text{True})$:

$$\begin{aligned}
\text{P}(\text{LM}) &= .75\text{P}(\overline{\text{ES}})\text{P}(\overline{\text{JP}}) + .25\text{P}(\overline{\text{ES}})\text{P}(\text{JP}) + .90\text{P}(\text{ES})\text{P}(\overline{\text{JP}}) + .50\text{P}(\text{ES})\text{P}(\text{JP}) \\
\text{P}(\text{LM}) &= .75(1 - .0625)(1 - .25) + .25(1 - .0625)(.25) \\
&\quad + .90(.0625)(1 - .25) + .50(.0625)(.25) \\
&= .636
\end{aligned}$$

Thus, by default, *Little Money* holds 64% of the time. A similar formula applies for the *BE* node, which has a default value of .617 (i.e., buying espresso holds 62% of the time).

If there were evidence that any of the ancestors nodes (e.g., *JP*) hold particular values, then the formulas would be the same as above except that the value of the given node would be fixed. For example, if *Just Paid* holds then $\text{P}(\overline{\text{JP}})$ is zero, so the cases involving it are effectively ignored, yielding a probability of 27% for *Little Money* and increasing *BE* to 75%. The calculations for $\text{P}(\text{LM})$ in this case follow:

$$\begin{aligned}
\text{P}(\text{LM}) &= .75\text{P}(\overline{\text{ES}})\text{P}(\overline{\text{JP}}) + .25\text{P}(\overline{\text{ES}})\text{P}(\text{JP}) + .90\text{P}(\text{ES})\text{P}(\overline{\text{JP}}) + .50\text{P}(\text{ES})\text{P}(\text{JP}) \\
\text{P}(\text{LM}) &= .75(1 - .0625)(0) + .25(1 - .0625)(1) + .90(.0625)(0) + .50(.0625)(1) \\
&= .266
\end{aligned}$$

A special case causes evaluation to be different from the simple top-down process. If any descendant node for an interior node is set, then the interior node is given a posterior distribution that would have produced the same value for the descendant node. For example, if the *Go Out* variable is known to hold, then *Little Money* is not likely to hold. Details on this special case can be found in (Charniak, 1992; Russell and Norvig, 1995), as well as information on an efficient algorithm for propagating values through the network.

GLOSSARY

annotations Markup added by humans to text corpora to make certain language phenomena explicit (e.g., relation types implied by prepositions).

attribute The *properties* of an entity expressed via qualities (not other entities).

collocation A word that tends to co-occur with another word. Collocations can be viewed as generalizing word associations.

concept Abstract representation for a class of entities.

differentia Differentiating relations as distinguished from taxonomic relations like *is-a*. Also, the part of a dictionary definition describing how a term differs from ones involving the same *genus* headword.

entity An instance of a *concept*, either an object, an attribute, or an event.

entry word The word being defined in a dictionary definition.

genus The type of a term. Also, the part a dictionary definition describing the general category for a term.

headword The main word in a phrase, determining overall syntactic properties.

property A feature of a concept, either an *attribute* or a *relation*.

relation The characterization of how two entities can be related (i.e., *relation type*). Also, a particular instance of this (i.e., *relationship*).

relation instance Same as *relationship*.

relation type A *concept* indicating how two entities can be related.

relationship A particular instance of a relation type, including the source and target terms (i.e., ⟨source, *relation-type*, target⟩).

sense One of the meanings a *word* can convey.

term A label used for a concept, typically given via a word or phrase.

word A language unit used to convey distinct meanings (i.e., lexeme).

wordform A spoken or written instance of a word (e.g., conveying inflection).

word-sense disambiguation The process of selecting a particular *sense* for an ambiguous *word*.

REFERENCES

Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. 1999. Statistical machine translation: Final report, JHU Workshop 1999. Technical report, Johns Hopkins University.

R. Almond. 1995. *Graphical Belief Modeling*. London, Chapman and Hall.

H. Alshawi. 1989. Analysing the dictionary definitions. In Boguraev and Briscoe (Boguraev and Briscoe, 1989), pages 153–169.

D. Altman. 1997. *Practical Statistics for Medical Research*. London, Chapman and Hall.

R. Amsler. 1980. *The Structure of the Merriam-Webster Pocket Dictionary*. Ph.D. thesis, University of Texas at Austin.

R. Amsler. 1995. Introduction. In Guo (Guo, 1995b), pages 1–13.

B. Atkins. 1995. The dynamic database. In Guo (Guo, 1995b), pages 131–143.

J. Ayto. 1983. On specifying meaning. In R. Hartmann, editor, *Lexicography: Principles and Practice*, pages 89–98. Academic Press, Inc., London.

K. Barker. 1998. *Semi-Automatic Recognition of Semantic Relationships in English Technical Texts*. Ph.D. thesis, Department of Computer Science, University of Ottawa.

G. Barnbrook. 2002. *Defining Language: A Local Grammar of Definition Sentences*. John Benjamins Publishing Company, Amsterdam.

C. Barrière. 1997. *From Machine Readable Dictionaries to a Lexical Knowledge Base of Conceptual Graphs*. Ph.D. thesis, Simon Fraser University.

R. Basili, M. Pazienza, and P. Verlardi. 1996a. A context driven conceptual clustering method. In B. Boguraev and J. Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*. MIT Press, Cambridge, MA.

R. Basili, M. Pazienza, and P. Verlardi. 1996b. An empirical symbolic approach to natural language processing. *Artificial Intelligence*, 85:59–99.

H. Béjoint. 1994. *Tradition and Innovation in Modern English Dictionaries*. Clarendon Press, Oxford.

A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G. Kim, M. Marcinkiewicz, and B. Schasberger. 1995. Bracketing guidelines for Treebank II style: Penn Treebank project. Technical report, University of Pennsylvania.

D. Blaheta and E. Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the American Association for Computational Linguistics (NAACL-2000)*, pages 234–40.

B. Boguraev and T. Briscoe, editors. 1989. *Computational Lexicography for Natural Language Processing*. Longman, London.

T. Briscoe, A. Copestake, and A. Lascarides. 1995. Blocking. In Saint-Dizier and Viegas (Saint-Dizier and Viegas, 1995), pages 273–302.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312.

R. Bruce and L. Guthrie. 1991. Building a noun taxonomy from a machine readable dictionary. Technical Report MCCS-91–207, Computing Research Laboratory, NMSU.

R. Bruce and J. Wiebe. 1999. Decomposable modeling in natural language processing. *Computational Linguistics*, 25(2):195–208.

B. Bruce. 1975. Case systems for natural language. *Artificial Intelligence*, 6:327–360.

K. Burns and A. Davis. 1999. Building and maintaining a semantically adequate lexicon using Cyc. In Viegas (Viegas, 1999), pages 121–143.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22 (2):249–254.

X. Carreras and L. Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Natural Language Learning (CoNLL-2004)*.

P. Cassidy. 2000. An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proceedings of the First International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2000)*.

E. Charniak. 1992. Bayesian networks without tears. *AI Magazine*, 12(4):50–63.

E. Charniak. 1993. *Statistical Language Learning*. MIT Press, Cambridge, MA.

G. Chierchia and S. McConnell-Ginet. 2000. *Meaning and Grammar*. MIT Press, Cambridge, MA, second edition.

T. Chklovski and R. Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*.

J. Cowie and W. Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

J. Cowie, J. Guthrie, and L. Guthrie. 1992. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 359–365.

D. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

I. Dagan, A. Itai, and U. Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*.

K. Dalgren and J. McDowell. 1986. Using commonsense knowledge to disambiguate prepositional phrase modifiers. In *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86)*, pages 589–593.

A. Van den Bosch and S. Buchholz. 2002. Shallow parsing on the basis of words only: A case study. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL'02)*, pages 433–440.

B. Dorr, N. Habash, and D. Traum. 1998. A thematic hierarchy for efficient generation from lexical-conceptual structure. Technical Report 022, UMIACS, University of Maryland.

B. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

D. Dowty. 1979. *Word Meaning and Montague Grammar*. D. Reidel Publishing, Holland.

P. Edmonds and S. Cotton, editors. 2001. *Proceedings of the* SENSEVAL *2 Workshop*. Association for Computational Linguistics.

P. Edmonds and G. Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

P. Edmonds and A. Kilgarriff. 2002. Special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4). Editors.

P. Edmonds. 1999. *Semantic Representations of Near-Synonyms for Automatic Lexical Choice*. Ph.D. thesis, University of Toronto.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

C. Fillmore, C. Wooters, and C. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*.

C. Fillmore. 1968. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York.

C. Fillmore. 1977. The case for case reopened. In *Syntax and Semantics 8*.

W. Frawley. 1992. *Linguistic Semantics*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

P. Fung and K. Church. 1994. K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*.

W. Gale, K. Church, and D. Yarowsky. 1993. A method for disambiguating words in a large corpus. *Computers and the Humanities*, 26:415–439.

I. Gati and A. Tversky. 1984. Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology*, 16:341–370.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

B. Gillon. 1999. The lexical semantics of English count and mass nouns. In Viegas (Viegas, 1999), pages 19–37.

C. Guo. 1995a. Constructing a MTD from LDOCE. In *Machine Tractable Dictionaries: Design and Construction* (Guo, 1995b), pages 145–225.

C. Guo, editor. 1995b. *Machine Tractable Dictionaries: Design and Construction*. Ablex Publishing Corporation, Norwood, NJ.

M. Halliday. 1956. The linguistic basis of a mechanical thesaurus, and its application to English preposition classification. *Mechanical Translation*, 3(2):81–88.

S. Harabagiu, G. Miller, and D. Moldovan. 1999. WordNet 2—A morphologically and semantically enhanced resource. In *Proceedings of the SIGLEX Workshop on Standardizing Lexical Resources*, pages 1–7.

M. Hearst. 1993. TextTiling: A quantitative approach to discourse segmentation. Technical Report 93/24, University of California at Berkeley.

M. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

D. Heckerman and J. Breese. 1994. Causal independence for probability assessment and inference using Bayesian networks. Technical Report MSR-TR-94-08, Microsoft Research, (Revised October, 1995).

I. Heim and A. Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell Publishers, Malden, MA.

D. Heylen. 1995. Lexical functions, generative lexicons and the world. In Saint-Dizier and Viegas (Saint-Dizier and Viegas, 1995), pages 125–140.

D. Hindle and M. Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

G. Hirst and D. St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malaproprisms. In Fellbaum (Fellbaum, 1998).

G. Hirst. 1986. Why dictionaries should list case structures. In *Proceedings of the Conference on Advances in Lexicography*, pages 147–162.

G. Hirst. 1988. Resolving lexical ambiguity computationally with spreading activation and polaroid words. In S. Small, G. Cottrell, and M. Tanenhaus, editors, *Lexical Ambiguity Resolution: Perspectives From Psycholinguistics, Neuropsychology, and Artificial Intelligence*, pages 73–107. Morgan Kaufmann, Los Altos, CA.

G. Hirst. 1995. Near-synonymy and the structure of lexical knowledge. In Klavans (Klavans, 1995), pages 51–56.

E. Hovy. 1998. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*.

N. Ide and J. Véronis. 1993. Extracting knowledge bases from machine readable dictionaries: Have we wasted our time? In *Proceedings of the International Conference on Building and Sharing of Very Large-Scale Knowledge Bases (KB&KS '93)*, pages 256–266.

N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1–40.

D. Inkpen and G. Hirst. 2001. Building a lexical knowledge-base of near-synonym differences. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*.

R. Jackendoff. 1983. *Semantics and Cognition*. MIT Press, Cambridge, MA.

R. Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, MA.

N. Japkowicz and J. Wiebe. 1991. Translating spatial prepositions using conceptual information. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 153–160.

K. Jensen and J. Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics*, 13(3–4):251–260.

D. Jurafsky and J. Martin. 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.

M. Kayaalp, T. Pedersen, and R. Bruce. 1997. A statistical decision making method: A case study on prepositional phrase attachment. In *Proceedings of the Workshop on Computational Language Learning (CoNLL '97)*.

A. Kilgarriff and M. Palmer, editors. 2000a. *Computers and the Humanities: Special Issue on* SENSEVAL, volume 34(1–2). Kluwer Academic Publishers, Dordrecht, the Netherlands.

A. Kilgarriff and M. Palmer. 2000b. Introduction to the special issue on SENSEVAL. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 15–48.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 15–48.

A. Kilgarriff. 1997. “I don’t believe in word senses”. *Computers and the Humanities*, 31(2):91–113.

A. Kilgarriff. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC ’98)*, pages 581–588.

J. Klavans, editor. 1995. *Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*. AAAI Spring Symposium Series.

D. Koller and A. Pfeffer. 1998. Probabilistic frame-based systems. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 580–587.

S. Landau. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, Cambridge, second edition.

H. Langone, B. Haskell, and G. Miller. 2004. Annotating WordNet. In *Proceedings of the Workshop on Frontiers in Corpus Annotation*.

S. Lauritzen and D. Spiegelhalter. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, B 50:157–224.

G. Leech. 1974. *Semantics*. Middlesex, Penguin Books.

F. Lehmann. 1996. Big posets of participatings and thematic roles. In P. Eklund, G. Ellis, and G. Mann, editors, *Conceptual Structures: Knowledge Representation as Interlingua*, pages 50–74, Berlin. Springer-Verlag.

W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. 1992. University of Massachusetts: Description of the CIRCUS system as used for MUC-4. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 282–288.

W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson, F. Feng, C. Dolan, and S. Goldman. 1993. UMass/Hughes: Description of the CIRCUS system used for MUC-5. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*.

D. Lenat and R. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, Reading, Massachusetts.

D. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11).

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the Fifth International Conference on Systems Documentation (SIGDOC '86)*, pages 24–26.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL-98)*, pages 768–774.

D. Lindley, G. Shafer, D. Spiegelhalter, et al. 1987. Special issue on probability in expert systems. *Statistical Science*, 1(2).

K. Litkowski. 1997. Automatic creation of lexical knowledge bases: New developments in computational lexicology. Technical Report 97–03, CL Research.

K. Litkowski. 2002. Digraph analysis of dictionary preposition definitions. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*.

K. Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12.

K. Litkowski. 2005. The Preposition Project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*.

R. Liu and V. Soo. 1993. An empirical study on thematic knowledge acquisition based on syntactic clues and heuristics. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93)*.

J. Lyons. 1977. *Semantics*. Cambridge University Press, Cambridge.

K. Mahesh and S. Nirenburg. 1995. A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, et al. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.

J. Markowitz, T. Ahlswede, and M. Evens. 1986. Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 112–119.

J. McCawley. 1986. What linguists might contribute to dictionary making if they could get their act together. In P. Bjarkman and V. Raskin, editors, *The Real-World Linguist: Linguistic Applications in the 1980s*, pages 3–18. Ablex, Norwood, NJ.

M. McShane and S. Nirenburg. 2002. Reference and ellipsis in ontological semantics. Technical Report MCCS-02-329, Computing Research Laboratory, NMSU.

D. Medin, R. Goldstone, and D. Gentner. 1993. Respects for similarity. *Psychological Review*, 100:252–278.

D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–49.

I. Mel'čuk and A. Polguere. 1987. A formal lexicon in the meaning-text theory (or how to do lexica with words). *Computational Linguistics*, 13(3–4):261–275.

R. Mihalcea and D. Moldovan. 2001. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools, 10(1–2)*, pages 5–21.

R. Mihalcea. 2002. Instance based learning with automatic feature selection applied to word sense disambiguation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4): Special Issue on WordNet.

G. Miller, R. Beckwith, C. Fellbaum, and D. Gross. 1993. Introduction to WordNet.

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*.

G. Miller. 1990. Introduction. *International Journal of Lexicography*, 3(4): Special Issue on WordNet.

G. Miller. 1996. *The Science of Words*. Scientific American Library, New York, second edition.

F. Mish, editor. 1996. *Merriam Webster's Collegiate Dictionary*. Merriam-Webster, Inc., Springfield, Massachusetts, 10th edition.

T. Mitchell. 1997. *Machine Learning*. New York, McGraw-Hill.

D. Moldovan and V. Rus. 2001. Logic form transformation of WordNet and its applicability to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

V. Nastase and S. Szpakowicz. 2001. Word sense disambiguation in Roget's thesaurus using WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*, pages 17–22.

V. Nastase and S. Szpakowicz. 2003. Augmenting WordNet's structure using LDOCE. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*.

H. Ng and H. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

S. Nirenburg and V. Raskin. 2004. *Ontological Semantics*. MIT Press, Cambridge, MA.

A. Novischi. 2002. Accurate semantic annotations via pattern matching. In *Proceedings of Florida Artificial Intelligence Research Society (FLAIRS-2002)*.

T. O'Hara, K. Mahesh, and S. Nirenburg. 1998. Lexical acquisition with WordNet and the Mikrokosmos ontology. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 94–101.

T. O'Hara, J. Wiebe, and R. Bruce. 2000. Selecting decomposable models for word-sense disambiguation: The GRLING-SDM system. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 159–164.

T. O'Hara, R. Bruce, J. Donner, and J. Wiebe. 2004. Class-based collocations for word-sense disambiguation. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 199–202.

B. Onyshkevych and S. Nirenburg. 1992. Lexicon, ontology, and text meaning. In J. Pustejovsky and S. Bergler, editors, *Lexical Semantics and Knowledge Representation*, pages 289–303. Berlin, Springer-Verlag.

B. Onyshkevych and S. Nirenburg. 1995. A lexicon for knowledge-based MT. *Machine Translation*, 10(2):5–57.

OpenCyc. 2002. OpenCyc release 0.6b. http://www.opencyc.org.

M. Stone Palmer. 1990. *Semantic Processing for Finite Domains*. Cambridge University Press, Cambridge.

D. Pearce. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

T. Pedersen and R. Bruce. 1998. Knowledge-lean word-sense disambiguation. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)*, pages 800–805.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

W. Phillips and E. Riloff. 2002. Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 125–132.

P. Procter, editor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, Harlow, Essex.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

M. Quillian. 1968. Semantic memory. In M. Minsky, editor, *Semantic Information and Processing*, pages 227–270. MIT Press, Cambridge, MA.

R. Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

J. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California.

R Team, 2004. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

V. Raskin and S. Nirenburg. 1995. Lexical semantics of adjectives: A microtheory of adjectival meaning. Technical Report MCCS-95-288, Computing Research Laboratory, NMSU.

D. Ravichandran and E. Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–47.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relations*. Ph.D. thesis, University of Pennsylvania.

P. Resnik. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora (WVLC-95)*.

S. Richardson, W. Dolan, and L. Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic*.

S. Richardson. 1997. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*. Ph.D. thesis, City University of New York.

E. Riloff and M. Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)*.

E. Rosch and C. Mervis. 1975. Family resemblance studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

E. Rosch. 1973. Natural categories. *Cognitive Psychology*, 4:328–350.

V. Rus. 2001. High precision logic form transformation. In *Proceedings of the International Conference with Tools in Artificial Intelligence*.

V. Rus. 2002. *Logic Forms for WordNet Glosses*. Ph.D. thesis, Computer Science Department, Southern Methodist University.

S. Russell and P. Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Upper Saddle River, NJ.

P. Saint-Dizier and E. Viegas, editors. 1995. *Computational Lexical Semantics*. Cambridge University Press, Cambridge.

R. Schank. 1973. Identification of conceptualizations underlying natural language. In R. Schank and K. Colby, editors, *Computer Models of Thought and Language*, pages 187–247. W. H. Freeman and Company, San Francisco.

R. Schvaneveldt, D. Dearholt, and F. Durso. 1988. Graph theoretic foundations of Pathfinder networks. *Computers and Mathematics with Applications*, 15:337–345.

S. Scott and S. Matwin. 1998. Text classification using WordNet hypernyms. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 38–44.

G. Shafer. 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

G. Shafer. 1987. Probability judgment in artificial intelligence and expert systems. *Statistical Science*, 2:3–44.

B. Slator and Y. Wilks. 1987. Towards semantic structures from dictionary entries. Technical Report MCCS-87-96, Computing Research Laboratory, NMSU.

B. Slator, S. Amirsoleymani, S. Andersen, K. Braaten, J. Davis, R. Ficek, H. Hakimzadeh, L. McCann, J. Rajkumar, S. Thangiah, and D. Thureen. 1990. Towards empirically derived semantic classes. In *Proceedings of the 5th Annual Rocky Mountain Conference on Artificial Intelligence (RMCAI-90)*, pages 257–262.

D. Sleator and D. Temperley. 1993. Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*.

S. Small and C. Rieger. 1982. Parsing and comprehending with word experts (a theory and its realization). In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Processing*, Hillsdale, NJ. Lawrence Erlbaum Associates.

E. Smith and D. Medin. 1981. *Categories and Concepts*. Harvard University Press, Cambridge, MA.

H. Somers. 1987. *Valency and Case in Computational Linguistics*. Edinburgh University Press, Edinburgh, Scotland.

J. Sowa. 1984. *Conceptual Structures in Mind and Machines*. Addison-Wesley, Reading, MA.

J. Sowa. 1999. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, Pacific Grove, CA.

R. Srihari, C. Niu, and W. Li. 2001. A hybrid approach for named entity and subtype tagging. In *Proceedings of the 6th Applied Natural Language Processing Conference*.

M. Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*.

J. Taylor. 1993. Prepositions: patterns of polysemization and strategies of disambiguation. In Zelinsky-Wibbelt (Zelinsky-Wibbelt, 1993).

I. Arturo Trujillo. 1995. *Lexicalist Machine Translation of Spatial Prepositions*. Ph.D. thesis, University of Cambridge.

A. Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

H. van Riemsdijk and E. Williams. 1986. *Introduction to the Theory of Grammar*. MIT Press, Cambridge, MA.

L. Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 782–88.

L. Vanderwende. 1995. Ambiguity in the acquisition of lexical information. In Klavans (Klavans, 1995), pages 174–179.

L. Vanderwende. 1996. *Understanding Noun Compounds using Semantic Information Extracted from On-Line Dictionaries*. Ph.D. thesis, Georgetown University.

J. Veenstra, A. van den Bosch, S. Buchholz, W. Daelemans, and J. Zavrel. 2000. Memory-based word sense disambiguation. In *Computers and the Humanities* (Kilgarriff and Palmer, 2000a), pages 171–177.

J. Veronis and N. Ide. 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*.

E. Viegas, B. Onyshkevych, V. Raskin, and S. Nirenburg. 1996. From submit to submitted via submission: On lexical rules in large-scale lexicon acquisition. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

E. Viegas, editor. 1999. *Breadth and Depth of Semantic Lexicons*. Kluwer, Dordrecht.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):525–579.

P. Vossen, P. Diez-Orzas, and W. Peters. 1997. The multilingual design of EuroWordNet. In *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

S. Weiss and C. Kulikowski. 1991. *Computer Systems that Learn*. Morgan Kaufmann Publishers, San Mateo, CA.

J. Wiebe, R. Bruce, and L. Duan. 1997. Probabilistic event categorization. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*.

J. Wiebe, K. McKeever, and R. Bruce. 1998a. Mapping collocational properties into machine learning features. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233.

J. Wiebe, T. O'Hara, and R. Bruce. 1998b. Constructing Bayesian networks from WordNet for word-sense disambiguation: Representational and processing issues. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 23–30.

Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, and B. Slator. 1989. A tractable machine dictionary as a resource for computational semantics. In Boguraev and Briscoe (Boguraev and Briscoe, 1989), pages 193–228.

Y. Wilks, B. Slator, and L. Guthrie. 1996. *Electric Words*. MIT Press, Cambridge, MA.

Y. Wilks. 1975a. An intelligent analyzer and understander of English. *CACM*, 18(5):264–274.

Y. Wilks. 1975b. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74.

Y. Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(2):197–223.

M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, and A. Vizedom. 2003. An interactive dialogue system for knowledge acquisition in Cyc. In *Proceedings of the Workshop on Mixed-Initiative Intelligent Systems*.

I. Witten and E. Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

D. Yarowsky, S. Cucerzan, R. Florian, C. Schafer, and R. Wicentowski. 2001. The Johns Hopkins SENSEVAL2 system descriptions. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 163–166.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460.

C. Zelinsky-Wibbelt, editor. 1993. *The Semantics of Prepositions: From Mental Processing to Natural Language Processing*. Mouton de Gruyter, Berlin.